The Pennsylvania State University
The Graduate School

CONTRIBUTIONS TO SEMIPARAMETRIC INFERENCE AND ITS APPLICATIONS

A Dissertation in
Statistics
by
Seong-ho Lee

© 2023 Seong-ho Lee

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2023

The dissertation of Seong-ho Lee was reviewed and approved by the following:

Yanyuan Ma
Professor of Statistics
Dissertation Advisor
Chair of Committee

Bharath Kumar Sriperumbudur
Associate Professor of Statistics

Necdet Serhat Aybat
Associate Professor of Industrial and Manufacturing Engineering

Bing Li
Verne M. Willaman Professor of Statistics
Chair of Graduate Studies

Abstract

This dissertation focuses on developing statistical methods for semiparametric inference and its applications. Semiparametric theory provides statistical tools that are flexible and robust to model misspecification. Utilizing the theory, this work proposes robust estimation approaches that are applicable to several scenarios with mild conditions, and establishes their asymptotic properties for inference. Chapter 1 provides a brief review of the literature related to this work. It first introduces the concept of semiparametric models and the efficiency bound. It further discusses two nonparametric techniques employed in the following chapters, kernel regression and B-spline approximation. The chapter then addresses the concept of dataset shift.

In Chapter 2, novel estimators of causal effects for categorical and continuous treatments are proposed by using an optimal covariate balancing strategy for inverse probability weighting. The resulting estimators are shown to be consistent for causal contrasts and asymptotically normal, when either the model explaining the treatment assignment is correctly specified, or the correct set of bases for the outcome models has been chosen and the assignment model is sufficiently rich. Asymptotic results are complemented with simulations illustrating the finite sample properties. A data analysis suggests a nonlinear effect of BMI on self-reported health decline among the elderly.

In Chapter 3, we consider a semiparametric generalized linear model and study estimation of both marginal mean effects and marginal quantile effects in this model. We propose an approximate maximum likelihood estimator and rigorously establish the consistency, the asymptotic normality, and the semiparametric efficiency of our method in both the marginal mean effect and the marginal quantile effect estimation. Simulation studies are conducted to illustrate the finite sample performance, and we apply the new tool to analyze non-labor income data and discover a new interesting predictor.

In Chapter 4, we propose a procedure to select the best training subsample for a classification model. Identifying patient's disease status from electronic health records (EHR) is a frequently encountered task in EHR related research. However, assessing patient's phenotype is costly and labor intensive, hence a proper selection of EHR as a training set is desired. We propose a procedure to tailor the training subsample for a classification model minimizing its mean squared error (MSE). We provide theoretical justification on its optimality in terms of MSE. The performance gain from our method is illustrated through simulation and a real data example, and is found often satisfactory under criteria beyond mean squared error.

In Chapter 5, we study label shift assumption and propose robust estimators for quantities of interest. In studies ranging from clinical medicine to policy research, the quantity of interest is often sought for a population from which only partial data is available, based on complete data from a related but different population. In this work, we consider this setting under the so-called label shift assumption. We propose an estimation procedure that only needs standard nonparametric techniques to approximate a conditional expectation, while by no means needs estimates for other model components. We develop the large sample theory for the proposed estimator, and examine its finite-sample performance through simulation studies, as well as an application to the MIMIC-III database.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my gratitude to those who have been with me throughout my Ph.D. journey for the past five years. Thanks to their encouragement and support, I was able to successfully complete my doctoral studies.

First and foremost, I would like to thank my advisor, Professor Yanyuan Ma. From the moment I passed the qualifying exam as a student to becoming a Ph.D. recipient and embarking on a new career as a researcher, her sharp academic insights and warm support have been invaluable. Whenever I needed to discuss our research or seek advice from her, she was always responsive and prioritized my requests despite her own schedule. The research findings in this dissertation would not have been achieved without her dedication and guidance. I'm truly proud of being one of her students.

Secondly, I would like to express my gratitude to my Ph.D. committee members: Profs. Bing Li, Bharath Kumar Sriperumbudur, and Necdet Serhat Aybat. Even before I asked them to be part of my committee, they had already taught me a great deal about statistics and mathematics through their wonderful lectures. They not only served willingly as my committee members, but also devoted their time and effort to provide helpful feedback. I sincerely appreciate their mentorship.

I extend my gratitude to the collaborators worked on the research projects that constitute this dissertation: Drs. Xavier de Luna, Elvezio Ronchetti, Jinbo Chen, Ying Wei, and Jiwei Zhao. It was an honor to have the opportunity to collaborate with them through my advisor. Their insights, expertise, and the critical aspects that I have learned from them will serve as key foundations for my future research.

Next, I would like to offer special thanks to Prof. Hyungsuk Tak for providing abundant advice during my job application process, and Prof. Jenny Shook for writing a teaching recommendation letter for my application. I am also grateful to Jaewoo Park, Kyongwon Kim, and Samidha Shetty for sharing their valuable experiences. I further express my gratitude to my friends at Penn State. I could not have accomplished this without all of you.

Lastly and most importantly, I appreciate my family, their love, and support: my father, Chunbok Lee, my mother, Youngseon Ko, and my brother, Sungbin Lee.

Chapter 1
Introduction

1.1 Semiparametric models and efficiency bound

A majority of statistical problems involve probability models, where we treat observations in a dataset as realizations of a vector of random variables $\mathbf{Z}$ drawn from a certain population of interest. For probability models, we typically use a parameter to characterize a distribution. Our objective is to estimate the parameter or a subset of its elements that best describes the distribution from which the observations are drawn. We can represent a class of probability densities as

$$\mathcal{P} \equiv \{f(\mathbf{z}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega\}.$$

Depending on the problem at hand, we may need to consider the class that is so large that the parameter $\boldsymbol{\theta}$ needs to be treated as infinite-dimensional. By allowing the parameter to be infinite-dimensional, we are imposing fewer restrictions on the probability model, which leads to solutions that are more comprehensive and robust than those obtained with finite-dimensional parametric models.

In some situations, we may split the parameter $\boldsymbol{\theta}$ into a p-dimensional parameter of interest $\boldsymbol{\beta}$ and an infinite-dimensional nuisance parameter $\boldsymbol{\eta}$, i.e., $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta})$. In other situations, the parameter of interest $\boldsymbol{\beta}$ may be expressed as a function of $\boldsymbol{\eta}$, denoted as $\boldsymbol{\beta}(\boldsymbol{\eta})$. These models are referred to as semiparametric models in the statistical literature, because they comprise both a parametric component $\boldsymbol{\beta}$ and a nonparametric component $\boldsymbol{\eta}$. In a semiparametric model, we assume that the observations $\mathbf{z}_i$ $(i = 1, \cdots, n)$ are independent and identically distributed random vectors generated from a population

with a density that belongs to the class

$$\mathcal{P} = [f\{\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\eta}(\cdot)\} : \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{\eta}(\cdot) \text{ is infinite-dimensional}].$$

A comprehensive literature on semiparametric models, their estimations, and asymptotic results can be found in Tsiatis (2006).

Since most of the reasonable estimators of $\boldsymbol{\beta}$ are asymptotically linear in either parametric or semiparametric models, our focus is on constructing asymptotically linear estimators. The asymptotically linear estimator of $\boldsymbol{\beta}_0$, $\widehat{\boldsymbol{\beta}}_n$, can be characterized as

$$\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = n^{-1} \sum_{i=1}^{n} \boldsymbol{\phi}(\mathbf{z}_i) + o_p(n^{-1/2}), \tag{1.1}$$

where $\boldsymbol{\phi}(\mathbf{z}_i)$ is a p-dimensional measurable random function such that $\mathrm{E}\{\boldsymbol{\phi}(\mathbf{Z})\} = \mathbf{0}$ and $\mathrm{E}\{\boldsymbol{\phi}^{\otimes 2}(\mathbf{Z})\}$ is finite and nonsingular. $\boldsymbol{\phi}(\mathbf{z}_i)$ is called an influence function of the ith observation on the estimator $\widehat{\boldsymbol{\beta}}_n$. In addition, we only consider regular estimators to avoid the problems related to super-efficient estimators, that is, the limiting distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ does not depend on the local data generating process.

Using the central limit theorem and Slutsky's theorem on (1.1), it is easy to show that the asymptotic variance of an regular and asymptotically linear (RAL) estimator is the variance of its influence function. Thus, to construct the efficient RAL estimator, we need to find its influence function, the efficient influence function. When faced with infinite-dimensional problems, it is common to first work with a finite-dimensional problem that well approximates the target problem, and then take the limit on the solution to the finite-dimensional problem. Similarly, we begin by considering a simpler finite-dimensional parametric model within the semiparametric model, and then extend the theory and methods to the latter.

We start by defining a parametric submodel $\mathcal{P}_{\boldsymbol{\beta},\boldsymbol{\gamma}}$, a class of densities characterized by the finite-dimensional parameter $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. A parametric submodel, or a submodel, $\mathcal{P}_{\boldsymbol{\beta},\boldsymbol{\gamma}}$ satisfies two conditions: (i) $\mathcal{P}_{\boldsymbol{\beta},\boldsymbol{\gamma}}$ is a subset of $\mathcal{P}$, and (ii) the true density $f_0(\mathbf{z}) \equiv f\{\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0(\cdot)\}$ belongs to $\mathcal{P}_{\boldsymbol{\beta},\boldsymbol{\gamma}}$. Here, the dimension of $\boldsymbol{\gamma}$, denoted as q, can vary depending on the choice of the submodel. We now define a nuisance tangent space of the submodel. Let $\mathcal{H}$ be a Hilbert space which consists of p-dimensional measurable functions that are mean-zero and finite-variance, and is equipped with the covariance

inner product. Then the submodel nuisance tangent space $\Lambda_{\boldsymbol{\gamma}} \in \mathcal{H}$ is defined as

$$\Lambda_{\boldsymbol{\gamma}} = \left\{ \mathbf{B}\mathbf{S}_{\boldsymbol{\gamma}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) : \mathbf{B} \in \mathbb{R}^{p \times q} \right\},$$

where $\mathbf{S}_{\boldsymbol{\gamma}}(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \partial \log f(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}$. Then the submodel efficient score $\mathbf{S}_{\text{eff}}^{\boldsymbol{\beta},\boldsymbol{\gamma}}$ is obtained by projecting the score for $\boldsymbol{\beta}$ onto $\mathcal{H}$, and subtracting its projection onto the submodel nuisance tangent space $\Lambda_{\boldsymbol{\gamma}}$. That is,

$$\mathbf{S}_{\text{eff}}^{\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \mathbf{S}_{\boldsymbol{\beta}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) - \Pi \left\{ \mathbf{S}_{\boldsymbol{\beta}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) \mid \Lambda_{\boldsymbol{\gamma}} \right\},$$

where $\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \partial \log f(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}) / \partial \boldsymbol{\beta}$. Then the submodel efficient influence function is

$$\boldsymbol{\phi}_{\text{eff}}^{\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{z}) = \left[\mathrm{E} \left\{ \mathbf{S}_{\text{eff}}^{\boldsymbol{\beta},\boldsymbol{\gamma}\otimes 2}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) \right\} \right]^{-1} \mathbf{S}_{\text{eff}}^{\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0).$$

It is known that the smallest asymptotic variance among the submodel RAL estimators is the variance of the submodel efficient influence function, $[\mathrm{E}\{\mathbf{S}_{\text{eff}}^{\boldsymbol{\beta},\boldsymbol{\gamma}\otimes 2}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\}]^{-1}$. A detailed discussion on the efficient influence function with the finite-dimensional nuisance parameter can be found in Chapter 3 of Tsiatis (2006).

We now describe the semiparametric efficient influence function. Let us first introduce a semiparametric nuisance tangent space. A semiparametric nuisance tangent space Λ is a mean-square closure of the submodel nuisance tangent space $\Lambda_{\boldsymbol{\gamma}}$, i.e.,

$$\begin{aligned} \Lambda \;=\; & \left\{ \mathbf{h}(\mathbf{z}) \in \mathcal{H} : \|\mathbf{h}(\cdot)\|^2 < \infty, \; \exists \text{ a sequence of submodels } \mathbf{B}_j \mathbf{S}_{\boldsymbol{\gamma} j}(\mathbf{z}) \right. \\ & \left. \text{such that } \; \|\mathbf{h}(\cdot) - \mathbf{B}_j \mathbf{S}_{\boldsymbol{\gamma} j}(\cdot)\|^2 \to 0 \text{ as } j \to \infty \right\}, \end{aligned}$$

where $\|\mathbf{h}(\cdot)\|^2 = \mathrm{E}\{\mathbf{h}^{\mathrm{T}}(\mathbf{Z})\mathbf{h}(\mathbf{Z})\}$. To ensure that the projection theorem can be applied, we assume that Λ is a linear and closed space. Then the semiparametric efficient score for $\boldsymbol{\beta}$ is

$$\mathbf{S}_{\text{eff}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = \mathbf{S}_{\boldsymbol{\beta}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) - \Pi \left\{ \mathbf{S}_{\boldsymbol{\beta}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) \mid \Lambda \right\}.$$

Then we can obtain the efficient influence function $\boldsymbol{\phi}_{\text{eff}}(\mathbf{z})$ by constructing a unique element of $\mathcal{H}$ that satisfies the conditions (i) $\mathrm{E}\{\boldsymbol{\phi}(\mathbf{Z})\mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\} = \mathrm{E}\{\boldsymbol{\phi}(\mathbf{Z})\mathbf{S}_{\text{eff}}^{\mathrm{T}}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\} = \mathbf{I}$ and (ii) $\Pi\{\boldsymbol{\phi}(\mathbf{z}) \mid \Lambda\} = \mathbf{0}$. Specifically,

$$\boldsymbol{\phi}_{\text{eff}}(\mathbf{z}) = \left[\mathrm{E}\{\mathbf{S}_{\text{eff}}^{\otimes 2}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\} \right]^{-1} \mathbf{S}_{\text{eff}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0).$$

It has been shown that the semiparametric efficiency bound is equal to the inverse of

the variance matrix of the semiparametric efficient score, $[\mathrm{E}\{\mathbf{S}_{\mathrm{eff}}^{\otimes 2}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\}]^{-1}$, and we can easily see that the efficient influence function $\boldsymbol{\phi}_{\mathrm{eff}}(\mathbf{z})$ achieves the semiparametric efficiency bound.

1.2 Kernel regression

In statistics, nonparametric methods are techniques for estimating population quantities without assuming any specific probability distribution for the data. Thus, nonparametric methods are useful when the data distribution is unknown or when the underlying relationship between variables is complex and not easily modeled by a parametric approach. One such nonparametric method is the estimation of conditional expectations. It involves estimating the conditional expectation function without making any assumptions about the functional form of the relationship between the variables.

One of the commonly used methods is kernel regression. The idea behind kernel regression is to estimate the expected value of a response Y given a predictor X by averaging over the observed values of Y for nearby values of X. A basic example is the Nadaraya-Watson estimator. This involves estimating the expectation at a given point as the weighted average of the observed data points with weights determined by a kernel function. Specifically, let $x_1, \cdots, x_n$ be a set of observed predictors and $y_1, \cdots, y_n$ be their corresponding responses. The Nadaraya-Watson estimator of $\mathrm{E}(Y \mid X = x)$ is given by

$$\widehat{m}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}$$

where $K_h(x) = K(x/h)/h$, $K(\cdot)$ is the kernel function, and h is the smoothing parameter known as the bandwidth.

The kernel function $K(\cdot)$ is typically chosen to be a non-negative, symmetric function that integrates to 1 over its support. A commonly used kernel function is the Gaussian kernel defined as $K(x) = \exp(-x^2/2)/\sqrt{2\pi}$. Other kernel functions such as the Epanechnikov, triangular, or quartic kernels can also be used depending on the problem at hand. The bandwidth parameter h determines the amount of smoothing in the estimate. A larger bandwidth will result in a smoother estimate, but with less sensitivity to local variations in the data. Conversely, a smaller bandwidth will result in a more variable estimate, but with more sensitivity to local features in the data.

The local polynomial kernel regression estimator extends the Nadaraya-Watson estimator by fitting a local polynomial function of degree p which best approximates the

response values in the neighborhood around the point of interest. The coefficients of the polynomial function can be estimated by solving the following weighted least square problem

$$\underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(x - x_i) \left\{ y_i - \beta_0 - \beta_1(x_i - x) - \cdots - \beta_p(x_i - x)^p \right\}^2 . \tag{1.2}$$

For notational brevity, let us define

$$\mathbf{y} = (y_1, \cdots, y_n)^{\mathrm{T}}, \ \boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^{\mathrm{T}}, \ \mathbf{K}_x = \operatorname{diag}\left\{ K_h(x - x_1), \cdots, K_h(x - x_n) \right\},$$

and

$$\mathbf{X}_x = \begin{bmatrix} 1 & (x_1 - x) & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - x) & \cdots & (x_n - x)^p \end{bmatrix}$$

Then we can rewrite (1.2) as $(\mathbf{y} - \mathbf{X}_x\boldsymbol{\beta})^{\mathrm{T}} \mathbf{K}_x (\mathbf{y} - \mathbf{X}_x\boldsymbol{\beta})$ which leads to the solution $\widehat{\boldsymbol{\beta}}_x = \left(\mathbf{X}_x^{\mathrm{T}}\mathbf{K}_x\mathbf{X}_x\right)^{-1} \mathbf{X}_x^{\mathrm{T}}\mathbf{K}_x\mathbf{y}$, and further, the estimator of $\mathrm{E}(Y \mid X = x)$ is

$$\widehat{\beta}_{0x} = \mathbf{e}_1^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_x = \mathbf{e}_1^{\mathrm{T}} \left(\mathbf{X}_x^{\mathrm{T}}\mathbf{K}_x\mathbf{X}_x\right)^{-1} \mathbf{X}_x^{\mathrm{T}}\mathbf{K}_x\mathbf{y}$$

where $\mathbf{e}_1 = (1, 0, \cdots, 0)^{\mathrm{T}}$. The estimator with $p = 0$ is equivalent to the Nadaraya-Watson estimator, and the estimator with $p = 1$ is often referred to as the local linear estimator, which will be denoted as $\widehat{m}_{\mathrm{LL}}(x)$.

We present the asymptotic properties of the basic kernel regression estimators provided in Theorem 3.1 of Fan & Gijbels (1996). Assume the predictor X is a random variable with density $f(x) > 0$ and let $m(x) = \mathrm{E}(Y \mid x)$ and $\sigma^2(x) = \mathrm{var}(Y \mid x)$. Also, the kernel function $K(\cdot)$ is symmetric and satisfies $\int K(u)du = 1$, $\int uK(u)du = 0$, and $\sigma_K^2 \equiv \int u^2K(u)du < \infty$. Further, suppose $n \to \infty$, $h \to 0$, and $nh \to \infty$. Then the Nadaraya-Watson estimator $\widehat{m}_{\mathrm{NW}}(\cdot)$ satisfies

$$\begin{aligned} \operatorname{bias}\left\{\widehat{m}_{\mathrm{NW}}(x)\right\} &= h^2\sigma_K^2 \left\{ \frac{m''(x)}{2} + \frac{m'(x)f'(x)}{f(x)} \right\} + o(h^2), \\ \operatorname{var}\left\{\widehat{m}_{\mathrm{NW}}(x)\right\} &= \frac{\|K(\cdot)\|_2^2}{nh} \frac{\sigma^2(x)}{f(x)} + o\{(nh)^{-1}\}, \end{aligned}$$

where $\|K(\cdot)\|_2^2 = \int K^2(u)du$. For the local linear estimator $\widehat{m}_{\mathrm{LL}}(\cdot)$,

$$\begin{aligned} \text{bias}\{\widehat{m}_{\mathrm{LL}}(x)\} &= h^2\sigma_K^2\frac{m''(x)}{2} + o(h^2), \\ \text{var}\{\widehat{m}_{\mathrm{LL}}(x)\} &= \frac{\|K(\cdot)\|_2^2}{nh}\frac{\sigma^2(x)}{f(x)} + o\{(nh)^{-1}\}. \end{aligned}$$

1.3 B-spline approximation

A B-spline curve is a piecewise polynomial curve defined by a set of control points and a set of basis functions. Let m be the number of control points, and let r be the degree of the basis functions. Then, the B-spline curve of degree r with m control points is defined as follows:

$$c_r(t) = \sum_{k=1}^{m} B_{r,k}(t)\gamma_k$$

where γ_k is the kth control point, and $B_{r,k}(t)$ is the kth basis function of degree r. The basis functions are defined recursively as follows:

$$B_{0,k}(t) = \begin{cases} 1 & \text{if } t_k \leqslant t < t_{k+1} \\ 0 & \text{otherwise,} \end{cases}$$

$$B_{r,k}(t) = \frac{t - t_k}{t_{k+r} - t_k}B_{r-1,k}(t) + \frac{t_{k+r+1} - t}{t_{k+r+1} - t_{k+1}}B_{r-1,k+1}(t),$$

where $t_1, \cdots, t_m$ are the knot values, which determine how the curve interpolates between control points. The knot values are typically chosen such that they are non-decreasing and satisfy certain properties, such as being asymptotically equidistant or having multiple knots at certain locations.

There have been several theoretical results established on the approximation properties of B-spline curves. One important result is the approximation order of B-spline curves. De Boor (1978) showed that B-spline curves of degree r can approximate a wide class of smooth functions up to order $(r-1)$. More specifically, for a function f that is q times continuously differentiable on a closed interval $[a, b]$, there exists a B-spline curve $c_r(t)$ of degree $r \geqslant (q+1)$ such that

$$\|c_r(\cdot) - f(\cdot)\|_\infty \leqslant C_q h^q \|f^{(q)}(\cdot)\|_\infty,$$

where $\|f(\cdot)\|_\infty \equiv \sup_{t\in[a,b]} |f(t)|$ is the supremum norm of a function f, C_q is a constant

which depends only on q, $h \equiv \max_{k=1,\cdots,m}(t_{k+1} - t_k)$ is the maximum knot spacing, and $f^{(q)}$ is the qth derivative of a function f. This result shows that B-spline curves can approximate smooth functions with high accuracy, and provides a theoretical foundation for their use in applications.

Utilizing the approximation properties of B-spline curves, its various applications have been studied in the statistics literature. One important example is nonparametric estimation of a regression curve. Suppose a response variable Y and a predictor variable X are associated as $y = m(x) + \epsilon$, where m is a unknown function and ϵ is error. In B-spline regression, the conditional expectation of Y given X is modeled as a linear combination of B-spline basis functions $\sum_{k=1}^{m} B_{r,k}(x_i)\gamma_k$. The coefficients of the B-spline curve $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_m)$ can be estimated by solving several optimization problems studied in the literature. One basic example is a least square method, i.e., the coefficients are estimated by

$$\widehat{\boldsymbol{\gamma}} = \text{argmin}_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \left\{ y_i - \sum_{k=1}^{m} B_{r,k}(x_i)\gamma_k \right\}^2 .$$

Once the coefficients are estimated, we can estimate $m(x)$ by $\widehat{m}(x) \equiv \sum_{k=1}^{m} B_{r,k}(x_i)\widehat{\gamma}_k$. Agarwal & Studden (1980) showed that, under certain assumptions on the smoothness of the true regression curve and the number of knots in the spline, the convergence rate of the estimator is $n^{-q/(2q+1)}$, where n is the sample size and q is the smoothness of the true curve. This rate of convergence is optimal (Stone 1982), in the sense that it cannot be improved upon by any estimator that depends only on the observed data.

Another important statistical application of B-spline curves is for density estimation. Consider the problem of estimating an unknown density function f based on sample data. Also, let F be its cumulative distribution function and $Q \equiv F^{-1}$ be its quantile function. Stone (1990) proposed the estimators of f, F, Q constructed based on the maximum likelihood estimation of a log-spline model. The log-spline model is defined as

$$f(\cdot, \boldsymbol{\gamma}) = \frac{\exp\{\sum_{k=1}^{m} B_{r,k}(\cdot)\gamma_k\}}{\int \exp\{\sum_{k=1}^{m} B_{r,k}(t)\gamma_k\} dt},$$

where $\boldsymbol{\gamma}$ satisfies $\sum_{k=1}^{m} \gamma_k = 0$. Then $\boldsymbol{\gamma}$ is estimated by

$$\begin{aligned} \widehat{\boldsymbol{\gamma}} &= \text{argmax}_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \log f(z_i, \boldsymbol{\gamma}) \\ &= \text{argmax}_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \left[\sum_{k=1}^{m} B_{r,k}(z_i)\gamma_k - \log \int \exp\left\{ \sum_{k=1}^{m} B_{r,k}(t)\gamma_k \right\} dt \right] \end{aligned}$$

under the constraint $\sum_{k=1}^{m} \gamma_k = 0$. Stone (1990) showed that the estimators achieve the optimal rate of convergence under mild conditions, and investigated the asymptotic behavior of the corresponding confidence bounds. As detailed in Kooperberg & Stone (1991), the log-spline density estimation has several noteworthy advantages. Firstly, this method is effective in producing accurate estimates even for densities that have some degree of irregularity. Additionally, the log-spline density estimator relies upon a parsimonious set of parameters, facilitating its applicability to diverse applications such as bootstrapping and robust regression. Lastly, the technique is naturally suited for quantile estimators and their confidence intervals.

1.4 Dataset shift

Dataset shift is a common phenomenon in real-world applications and scenarios, where the testing data experiences a change in distribution of a single feature, a combination of features, or the class boundaries (Moreno-Torres et al. 2012). Formally, dataset shift occurs when the joint distributions of training and test data differ, that is, when $f_{\mathrm{s}}(y, \mathbf{x}) \neq f_{\mathrm{t}}(y, \mathbf{x})$ where f_{s} and f_{t} denote the distribution functions of observations from training (source) and test (target) data respectively. This leads to a violation of the common assumption that the training and testing data follow the same distributions. Therefore, it is necessary to develop robust learning models that can handle such variations in data distributions.

To address dataset shift, it is essential to characterize the relationships between the covariates and the label, as this characteristic determines what kinds of shift can occur and affect a given problem. We can classify these relationships as two types, which will be referred to as $\mathbf{X} \to Y$ and $Y \to \mathbf{X}$ problems. In $\mathbf{X} \to Y$ problems, the label is causally determined by the values of the covariates, while in $Y \to \mathbf{X}$ problems, the label causally determines the values of the covariates.

We can further distinguish different types of dataset shifts that can occur. One type of shift is covariate shift, which appears only in $\mathbf{X} \to Y$ problems. It is defined as the case where $f_{\mathrm{s}}(y \mid \mathbf{x}) = f_{\mathrm{t}}(y \mid \mathbf{x})$ and $f_{\mathrm{s}}(\mathbf{x}) \neq f_{\mathrm{t}}(\mathbf{x})$. Another type of shift is label shift, which appears only in $Y \to \mathbf{X}$ problems. Label shift is defined as the case where $f_{\mathrm{s}}(\mathbf{x} \mid y) = f_{\mathrm{t}}(\mathbf{x} \mid y)$ and $f_{\mathrm{s}}(y) \neq f_{\mathrm{t}}(y)$. Additionally, concept shift is defined as $f_{\mathrm{s}}(y \mid \mathbf{x}) \neq f_{\mathrm{t}}(y \mid \mathbf{x})$ and $f_{\mathrm{s}}(\mathbf{x}) = f_{\mathrm{t}}(\mathbf{x})$ in $\mathbf{X} \to Y$ problems, and $f_{\mathrm{s}}(\mathbf{x} \mid y) \neq f_{\mathrm{t}}(\mathbf{x} \mid y)$ and $f_{\mathrm{s}}(y) = f_{\mathrm{t}}(y)$ in $Y \to \mathbf{X}$ problems. Concept shift presents the hardest challenge among the different types of dataset shift that has been tackled so far.

There are several possible causes for dataset shift, and two of the most important ones are sample selection bias and non-stationary environments. Sample selection bias occurs when the distribution of the training data does not represent the operating environment where the classifier is to be deployed. This is usually caused by a biased sampling method, leading to an over- or under-representation of certain types of samples in the training set. In contrast, non-stationary environments arise when the distribution changes over time or space. This can occur, for example, when the data is collected from different sources, or when the underlying statistical relationship between the covariates and label changes due to external factors.

Sample selection bias have four different forms: missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR), and missing at random-class (MARC). MCAR occurs when the probability of being selected into training data is independent of both the covariates and label, i.e., $P(S = 1 \mid y, \mathbf{x}) = P(S = 1)$ where a binary variable S is the indicator whether a sample belongs to the training set or not. In this case, there is no systematic bias introduced by the sampling mechanism, and hence no dataset shift. MAR occurs when the selection probability depends on the covariates but not on the label, so that $P(S = 1 \mid y, \mathbf{x}) = P(S = 1 \mid \mathbf{x})$. This kind of bias can potentially cause covariate shift, where the covariate distribution of the training set differs from that of the test set. MNAR occurs when the selection probability depends on both the covariates and label, so that $P(S = 1 \mid y, \mathbf{x})$ is not independent of $\mathbf{x}$ or y. This kind of bias can introduce one or more types of dataset shift. Finally, MARC occurs when the selection probability depends on the label but not on the covariates, so that $P(S = 1 \mid y, \mathbf{x}) = P(S = 1 \mid y)$. This kind of bias can potentially cause label shift, where the distribution of the label in the training set differs from that of the test set.

The second cause, non-stationary environments, is also common in real-world applications, where the distribution of the covariates and/or label can change over time or space. Non-stationarity can introduce different kinds of dataset shift depending on the type of problem. In $\mathbf{X} \to Y$ problems, a non-stationary environment can create changes in either $f(\mathbf{x})$ or $f(y \mid \mathbf{x})$, respectively leading to covariate shift or concept shift. On the other hand, in $Y \to \mathbf{X}$ problems, non-stationarity can generate label shift with a change in $f(y)$ or concept shift with a change in $f(\mathbf{x} \mid y)$.

1.5 Organization of dissertation

In Chapter 2 of the dissertation, covariate balancing for causal inference on categorical and continuous treatments is discussed. The chapter discusses balancing scores and preliminaries for categorical treatments in Section 2.2.1, and for continuous treatments in Section 2.3.1. The asymptotic properties for two types of treatments are presented in Sections 2.2.2 and 2.3.2 respectively. Simulation experiments are carried out in Section 2.4, and data application is discussed in Section 2.5 where we investigate the effect of body mass index on self-reported health decline. The technical proofs for this chapter can be found in Appendix A.

Chapter 3 of the dissertation presents a semiparametric approach to estimation of marginal and marginal quantile effects. The chapter provides a discussion of the methodology in Section 3.2, which addresses the efficiency bounds of marginal and marginal quantile effect estimation, and the estimation procedure. The theoretical properties of the proposed method are then discussed in Section 3.3, including the asymptotic normality and efficiency. Section 3.4 presents simulation experiments and Section 3.5 applies the method to non-labor income data in Switzerland. The technical proofs for this chapter can be found in Appendix B.

Chapter 4 of the dissertation proposes an optimal sampling approach for positive-only electronic health record data. The chapter addresses the main contributions of the work in Section 4.2, which consists of the methodology, theory, and classification given a certain variable. Section 4.3 presents simulation experiments, covering correct and wrong model specification, and robustness to violation of positive-only assumption. The proposed method is applied to a real-world dataset in Section 4.4 to identify patients with depression symptoms using their electronic health records. The technical proofs for this chapter can be found in Appendix C.

Chapter 5 of the dissertation discusses a method for doubly flexible estimation under label shift. The chapter begins with an introduction in Section 5.1 and a discussion of the model structure in Section 5.2. The proposed doubly flexible estimation approach is presented in Section 5.3, which includes a general approach and the proposed estimator, and an alternative singly flexible estimator is also discussed in Section 5.4. Section 5.5 presents simulation experiments, and Section 5.6 applies the method to the MIMIC-III database. The chapter concludes with a discussion in Section 5.7. The technical proofs for the chapter can be found in Appendix D, which also includes the proposed approach for a general quantity of interest.

Chapter 2
Covariate balancing for causal inference on categorical and continuous treatments

2.1 Introduction

Encouraged by the recent booming development of the causal inference literature, we devise and study a novel inference tool for categorical and continuous treatments by using covariate balancing strategies for inverse probability weighting (e.g., Imai & Ratkovic 2014, Wang & Zubizarreta 2020, Fan et al. 2022, Sant'Anna et al. 2022). Our study is built on foundational work on optimal covariate balancing by Fan et al. (2022), while we overcome additional methodological and theoretical challenges.

When estimating a causal effect on an outcome, weighting based on the propensity score (model for the probability of the treatment given observed pre-treatment covariates) is often used to construct optimal estimators by an augmentation using fitted models for the outcome given the covariates. These augmented inverse probability weighting estimators have robustness properties, and are locally efficient (e.g., Robins & Rotnitzky 1995, Scharfstein et al. 1999, Cantoni & de Luna 2020). Most of the literature on causal inference has focused on binary treatments, where the causal parameter of interest is a contrast between two treatments. Nevertheless, there is an increasing interest in the multi-valued treatments (e.g., Fong et al. 2018, Kennedy et al. 2017, Yang et al. 2016, Ao et al. 2021, Lee 2018) often encountered in applied work, in both the medical and social sciences. Causal effects of categorical treatment were formalized by Imbens (2000) and Robins et al. (2000), while Cattaneo (2010) deduced the semiparametric efficiency bound; see also Yang et al. (2016) for a review. Causal effects of continuous treatments were formalized by Robins et al. (2000), Laan & Robins (2003), Hirano & Imbens (2004) and Galvao & Wang (2015) and studied in the context of mediation in Huber et al.

(2020). In contrast to these, Kennedy et al. (2017) proposed a double robust estimation strategy, avoiding parametric specification of the dose-response curve.

We contribute to the somewhat less rich literature on robust estimation for categorical and continuous treatments by using an estimation strategy based on covariate balancing propensity score estimation for inverse probability weighting (e.g., Imai & Ratkovic 2014, Fong et al. 2018). Fan et al. (2022) recently obtained key results in the binary treatment case by specifying which covariate functions should be balanced for efficient inference: the propensity score model should be fitted by balancing a set of bases for the outcome models in the space spanned by the covariates. We provide corresponding results for the categorical and continuous treatment cases, thereby completing the story. In particular, the procedures we proposed balance the "most suitable" functions of the covariates when the propensity score is correctly specified, in the sense that they minimize the variability of the causal effect estimation. When the propensity score is misspecified but the outcome basis functions are correct, the procedure looks for an approximate balance by minimizing the squared bias of the resulting estimator. As other recent proposals for the binary treatment case (Wang & Zubizarreta 2020, Athey et al. 2018, Zubizarreta 2015, Wong & Chan 2018), the method presented here does not necessarily try to achieve exact balance where this is not feasible, although in practice exact balance can always be targeted by enriching the assignment model.

For both the categorical and continuous treatment case, the proposed estimators are shown to be robust, i.e. consistent and asymptotically normal for causal contrasts of interest, both when the model explaining treatment assignment is correctly specified, and when the correct set of bases for the outcome models has been chosen and the propensity score model is sufficiently rich. For the categorical treatment case, we show that the estimator proposed attains the semiparametric efficiency bound when both the treatment assignment model and the outcome basis are correctly specified. For the continuous case, the causal parameter of interest is a function. The latter is not parametrized and the estimators proposed are shown to have bias and variance of the classical nonparametric order under typical regularity conditions, hence with a usual bias-variance trade-off. Unlike other double robust procedures, the proposed method avoids estimating parameters involved in the outcome model, regardless of whether the outcome model is correct or not. In addition, the proposed procedure for estimating the propensity score can be combined with outcome model fitting, and this is observed to sometimes improve the finite sample performance of the classical augmented inverse probability weighting estimator.

The rest of the paper is organized as follows. Sections 2.2 and 2.3 deal with the categorical and the continuous treatment cases, respectively. In both sections, inverse probability weighting estimators are introduced, where a working model for the generalized propensity score is estimated by balancing basis functions for the outcome models. We establish the theoretical properties of the estimators. Simulation studies are conducted in Section 2.4 to illustrate the finite sample performance of our methods. In Section 2.5, we use data from a nine year follow-up study of the elderly to estimate the dose-response curve of BMI on self-reported health decline. Section 2.6 concludes the paper, while all proofs are relegated to the Appendix.

2.2 Categorical treatments

2.2.1 Balancing scores and preliminaries

Consider $K+1$ treatments, $A = 0, 1, \ldots, K$, and their respective potential outcomes $Y^0, \ldots, Y^K$. We observe a random sample $(A_i, Y_i, \mathbf{X}_i), i = 1, \ldots, n$, where we assume $Y_i = Y_i^k$ if $A_i = k$, and $\mathbf{X}_i \in \mathbb{R}^d$ is a vector of pre-treatment covariates. We also assume ignorability of the treatment assignment, i.e. $E(Y_i^k \mid \mathbf{X}_i, A_i) = E(Y_i^k \mid \mathbf{X}_i) \equiv m(k, \mathbf{X}_i)$ and $\text{pr}(A_i = k \mid \mathbf{X}_i = \mathbf{x}) \equiv \pi_0(k, \mathbf{x}) > \delta > 0$ for all $k \in \{0, 1, \ldots, K\}$ and all $\mathbf{x}$, where $\pi_0(k, \mathbf{x})$ is named generalized propensity score in the literature (Imbens 2000).

Let $\theta_k \equiv E(Y_i^k)$ for $k = 0, 1, \ldots, K$ be the average response to the different treatment levels. The parameters of interest are typically differences between these average responses, i.e. causal contrasts such as $\theta_k - \theta_0$, if $k = 0$ is a treatment level of reference. We consider a parametric working model $\pi(k, \mathbf{x}, \boldsymbol{\beta})$ for $\pi_0(k, \mathbf{x})$, with $\boldsymbol{\beta} \in \mathbb{R}^p$, and vectors of basis functions, $\mathbf{B}(k, \mathbf{X}) : \mathbb{R}^{d+1} \to \mathbb{R}^q$, aiming at spanning $m(k, \mathbf{x})$. For notational simplicity we assume q does not depend on k. Thus, correct specification will imply that there exists a value $\boldsymbol{\beta}_0$ with

$$\pi(k, \mathbf{x}, \boldsymbol{\beta}_0) = \pi_0(k, \mathbf{x}), \tag{2.1}$$

and there exists $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_K^{\mathrm{T}})^{\mathrm{T}}$ with

$$\boldsymbol{\alpha}_k^{\mathrm{T}} \mathbf{B}(k, \mathbf{x}) = m(k, \mathbf{x}), \tag{2.2}$$

for all k and all $\mathbf{x}$. Misspecification, i.e. situations where (2.1) or (2.2) does not hold for any value of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, will also be considered in the sequel. Note that one of the

advantages of the balancing approach studied here is that the parameter $\boldsymbol{\alpha}$ does not need to be known or estimated. We hence do not use a subscript 0 on $\boldsymbol{\alpha}$ and $m(\cdot)$ to distinguish the true parameter value and the correct model since this will be clear from the context.

To estimate θ_k under the above assumptions one needs to control for the covariates $\mathbf{X}_i$ by using one or both working models. In particular, $\pi(k, \mathbf{x}, \boldsymbol{\beta})$ is a balancing score in the sense that $\mathbf{X}_i \perp\!\!\!\perp A_i \mid \pi(k, \mathbf{x}, \boldsymbol{\beta}_0)$ under (2.1) (Rosenbaum & Rubin 1983). Thus, for the binary case ($K = 1$), Imai & Ratkovic (2014) proposed to solve

$$\sum_{i=1}^{n} \left\{ \frac{I(A_i = 1)}{\pi(1, \mathbf{X}_i, \boldsymbol{\beta})} - \frac{I(A_i = 0)}{\pi(0, \mathbf{X}_i, \boldsymbol{\beta})} \right\} \mathbf{b}(\mathbf{X}_i) = 0$$

to obtain $\widehat{\boldsymbol{\beta}}$, where $\mathbf{b}(\mathbf{X}_i)$ is a vector valued function of the covariates. Based on the resulting fitted propensity score $\pi(k, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})$, we focus on an inverse probability weighting estimator for θ_k, i.e.,

$$\widehat{\theta}_k = n^{-1} \sum_{i=1}^{n} \frac{I(A_i = k) Y_i}{\pi(k, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})}. \tag{2.3}$$

Two issues arise regarding the above procedure. The first is that if the propensity score model (2.1) is misspecified, then $\widehat{\theta}_k$ is generally biased. The second is the choice of $\mathbf{b}(\mathbf{X})$, which is largely left unsupervised. Fan et al. (2022) overcame these two issues in the binary case ($K = 1$), and proposed an optimal choice for $\mathbf{b}(\mathbf{X})$, in the sense that the resulting treatment effect estimator is consistent when (2.1) is correct, or when (2.2) is correct and (2.1) has sufficient flexibility, and is efficient if both are correct.

We aim to achieve the same kind of optimality and robustness in the categorical treatment case. Two different estimators are introduced with different properties, which we discuss heuristically below, before giving a formal treatment in the next section. The first is to estimate $\boldsymbol{\beta}$ by solving the following balancing condition

$$\sum_{i=1}^{n} \left[\left\{ \frac{I(A_i = k)}{\pi(k, \mathbf{X}_i, \boldsymbol{\beta})} - 1 \right\} \mathbf{B}(k, \mathbf{X}_i) - \left\{ \frac{I(A_i = 0)}{\pi(0, \mathbf{X}_i, \boldsymbol{\beta})} - 1 \right\} \mathbf{B}(0, \mathbf{X}_i) \right] = \mathbf{0} \tag{2.4}$$

at all $k = 1, \ldots, K$, i.e. a system of qK equations. GMM, as described below, can be used if $qK \geqslant p$. This balancing condition is motivated by pushing the bias of the contrast estimator $\widehat{\theta}_k - \widehat{\theta}_0$ towards zero. In fact, it will be shown that the asymptotic

bias of $\widehat{\theta}_k - \widehat{\theta}_0$ is equal to

$$E\left[\{I(A_i = k)/\pi(k, \mathbf{X}_i, \boldsymbol{\beta}) - 1\}\, m(k, \mathbf{X}_i) - \{I(A_i = 0)/\pi(0, \mathbf{X}_i, \boldsymbol{\beta}) - 1\}\, m(0, \mathbf{X}_i)\right].$$

An alternative to setting the bias of $\widehat{\theta}_k - \widehat{\theta}_0$ to zero for $k = 1, \ldots, K$, is to directly put the bias of $\widehat{\theta}_k$ to zero, for $k = 0, \ldots, K$, by separately balancing both terms in (2.4), i.e. solving the condition

$$\sum_{i=1}^{n} \left\{ \frac{I(A_i = k)}{\pi(k, \mathbf{X}_i, \boldsymbol{\beta})} - 1 \right\} \mathbf{B}(k, \mathbf{X}_i) = \mathbf{0} \tag{2.5}$$

at all $k = 0, \ldots, K$, i.e. a system of $q(K + 1)$ equations. We will use GMM allowing for $q(K + 1) \geqslant p$; see (2.6) below.

An interesting discovery here is that the two alternatives are not necessarily equivalent. The former choice allows for biased estimation of $\widehat{\theta}_k$; the only aim being to estimate the contrast $\theta_k - \theta_0$ without bias. We find that, if $\widehat{\theta}_k$ is indeed biased, then $\widehat{\theta}_k - \widehat{\theta}_0$ will not be efficient. This is because local efficiency holds when the fitted propensity score is correctly specified and its parameters are consistently estimated, which is not the case when (2.5) does not hold. Due to this consideration, below we focus on solving (2.5) and show that the resulting estimator of θ_k in (2.3) has, under certain conditions, a robust property and, when all working models are correctly specified, reaches the asymptotic semiparametric efficiency bound. This had not yet been established in the literature. The consideration of the categorical case further allows us to extend the method to the continuous treatment case, viewing this as infinitely many categories and incorporating smoothness.

2.2.2 Asymptotic properties

We now establish a robustness property and the asymptotic distribution results of the estimator in (2.3), where $\boldsymbol{\beta}$ is estimated through covariate balancing (2.5); see A.1 for proofs. To gain an intuitive understanding of the robustness property, we can verify that when the propensity score model is correctly specified, i.e. when (2.1) holds for all k and all $\mathbf{x}$, $\widehat{\boldsymbol{\beta}}$ is $\sqrt{n}-$consistent under the standard regularity conditions for GMM estimation (Newey & McFadden 1994), and $\pi(k, \mathbf{x}, \widehat{\boldsymbol{\beta}}) \to \pi(k, \mathbf{x}, \boldsymbol{\beta}_0) = \pi_0(k, \mathbf{x})$ in probability as n tends to infinity. The consistency of $\hat{\theta}_k$ is a consequence of

$$E\left[\{I(A_i = k)/\pi(k, \mathbf{X}_i, \boldsymbol{\beta}_0) - 1\}\, \mathbf{B}(k, \mathbf{X}_i)\right] = \mathbf{0}$$

in combination with the regularity conditions, irrespective of whether or not a correct basis for the outcome models is specified. This then leads to the convergence of

$$E(\widehat{\theta}_k) = E\left\{ n^{-1} \sum_{i=1}^{n} \frac{I(A_i = k) Y_i}{\pi(k, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})} \right\} \rightarrow E\left\{ \frac{I(A_i = k) Y_i^k}{\pi_0(k, \mathbf{X}_i)} \right\} = \theta_k,$$

as $n \rightarrow \infty$. On the other hand, when the outcome model basis is actually correctly specified, i.e. when (2.2) holds for all k and $\mathbf{x}$, then the propensity model (2.1) does not need be correct as long as (2.5) has a solution. In such case, $\widehat{\boldsymbol{\beta}}$ is consistent for some value $\boldsymbol{\beta}^*$, hence $\pi(k, \mathbf{x}, \widehat{\boldsymbol{\beta}})$ converges in probability to some function $\pi(k, \mathbf{x})$. We then have

$$\begin{aligned} E(\widehat{\theta}_k) &= E\left\{ n^{-1} \sum_{i=1}^{n} \frac{I(A_i = k) Y_i}{\pi(k, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})} \right\} \rightarrow E\left\{ \frac{I(A_i = k) Y_i^k}{\pi(k, \mathbf{X}_i)} \right\} \\ &= E\left[\left\{ \frac{\pi_0(k, \mathbf{X}_i)}{\pi(k, \mathbf{X}_i)} - 1 \right\} m(k, \mathbf{X}_i) + m(k, \mathbf{X}_i) \right] = \theta_k, \end{aligned}$$

as $n \rightarrow \infty$, where the last equality is the result of (2.2) and (2.5).

To be more formal, let

$$\mathbf{f}_{ki}(\boldsymbol{\beta}) \equiv \left\{ \frac{I(A_i = k)}{\pi(k, \mathbf{X}_i, \boldsymbol{\beta})} - 1 \right\} \mathbf{B}(k, \mathbf{X}_i),$$

$\mathbf{f}_i(\boldsymbol{\beta}) \equiv \{\mathbf{f}_{0i}(\boldsymbol{\beta})^{\mathrm{T}}, \ldots, \mathbf{f}_{Ki}(\boldsymbol{\beta})^{\mathrm{T}}\}^{\mathrm{T}}$, $\mathbf{V}(\boldsymbol{\beta}) \equiv E\{\mathbf{f}_i(\boldsymbol{\beta})\mathbf{f}_i(\boldsymbol{\beta})^{\mathrm{T}}\}$, $\widehat{\mathbf{V}}(\boldsymbol{\beta}) \equiv n^{-1} \sum_{i=1}^{n} \mathbf{f}_i(\boldsymbol{\beta})\mathbf{f}_i(\boldsymbol{\beta})^{\mathrm{T}}$, $\mathbf{A}(\boldsymbol{\beta}) \equiv E\left\{\partial \mathbf{f}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{\mathrm{T}}\right\}$ and $\widehat{\mathbf{A}}(\boldsymbol{\beta}) \equiv n^{-1} \sum_{i=1}^{n} \partial \mathbf{f}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{\mathrm{T}}$. Further, let $\boldsymbol{\theta} \equiv (\theta_0, \ldots, \theta_K)^{\mathrm{T}}$, $g_{ki}(\boldsymbol{\beta}) \equiv I(A_i = k) Y_i / \pi(k, \mathbf{X}_i, \boldsymbol{\beta}) - E\{m(k, \mathbf{X}_i)\}$, $\mathbf{g}_i(\boldsymbol{\beta}) = \{g_{1i}(\boldsymbol{\beta}), \ldots, g_{Ki}(\boldsymbol{\beta})\}^{\mathrm{T}}$ and $\mathbf{B}(\boldsymbol{\beta}) \equiv E\{\partial \mathbf{g}_i(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta}^{*\mathrm{T}}\}$. We solve for a solution of (2.5) by minimizing

$$\{\sum_{i=1}^{n} \mathbf{f}_i(\boldsymbol{\beta})\}^T \widehat{\mathbf{V}}(\boldsymbol{\beta})^{-1} \{\sum_{i=1}^{n} \mathbf{f}_i(\boldsymbol{\beta})\}. \tag{2.6}$$

We will use the following regularity conditions:

A0. $\boldsymbol{\beta}^*$ is the unique solution of $E\{\mathbf{f}_i(\boldsymbol{\beta})\} = 0$.

A1. The variance-covariance matrix $\mathbf{V}(\boldsymbol{\beta}^*)$ has bounded positive eigenvalues.

A2. $\mathbf{f}_i(\boldsymbol{\beta})$ is differentiable with respect to $\boldsymbol{\beta}$.

A3. The matrix $\mathbf{A}(\boldsymbol{\beta}^*)$ is bounded and has full column rank.

A4. $\mathbf{g}_i(\boldsymbol{\beta})$ is differentiable with respect to $\boldsymbol{\beta}$.

These are classical regularity conditions. Condition A0 requires the existence and uniqueness of a solution, where the uniqueness can be relaxed to local uniqueness. The existence requirement is automatic when the $\pi(k, \mathbf{x}, \boldsymbol{\beta})$ model is correct. In this case $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. It is also natural and standard when $(K+1)q$, the number of equations in $E\{\mathbf{f}_i(\boldsymbol{\beta})\}$, is not larger than p, the dimension of $\boldsymbol{\beta}$ which is achievable through enriching the $\pi(k, \mathbf{x}, \boldsymbol{\beta})$ model. Thus, regardless of whether or not $\pi(k, \mathbf{x}, \boldsymbol{\beta})$ is correctly specified, we can always justify Condition A0.

Theorem 2.2.1. *Assume that either (2.1) holds for all k and $\mathbf{x}$, or (2.2) holds for all k and $\mathbf{x}$. Then, under regularity conditions A0 to A4, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ has asymptotic normal distribution with mean zero and variance*

$$\begin{aligned}\boldsymbol{\Sigma} \;=\; & \mathbf{B}(\boldsymbol{\beta}^*)\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{B}(\boldsymbol{\beta}^*)^{\mathrm{T}} + \mathbf{C}(\boldsymbol{\beta}^*) \\ & -\mathbf{B}(\boldsymbol{\beta}^*)\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{D}(\boldsymbol{\beta}^*) \\ & -\mathbf{D}(\boldsymbol{\beta}^*)^{\mathrm{T}}[\mathbf{B}(\boldsymbol{\beta}^*)\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}]^{\mathrm{T}},\end{aligned}$$

where $\mathbf{C}(\boldsymbol{\beta}^*) \equiv E\{\mathbf{g}_i(\boldsymbol{\beta}^*)^{\otimes 2}\}$ and $\mathbf{D}(\boldsymbol{\beta}^*) \equiv E\{\mathbf{f}_i(\boldsymbol{\beta}^*)\mathbf{g}_i(\boldsymbol{\beta}^*)^{\mathrm{T}}\}$.

Theorem 2.2.1 highlights a robustness property. On the one hand, if the propensity score is correctly specified then we will have a consistent estimator of the treatment contrast even if the outcome basis is misspecified. On the other hand, we can also afford to misspecify the propensity score model, provided that the outcome basis functions are correctly specified. In the latter case, Condition A0 plays a pivotal role and it is crucial that it be satisfied. An example is to use the model $\pi(k, \mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}_{(k)}^{\mathrm{T}}\mathbf{B}(k, \mathbf{x})$, $k = 0, \ldots, K$, with $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(0)}^{\mathrm{T}}, \cdots, \boldsymbol{\beta}_{(K)}^{\mathrm{T}})^{\mathrm{T}}$ so that $\boldsymbol{\beta}$ has length $p = q(K+1)$. Then (2.5) is the derivative of the loss function

$$\sum_{i=1}^{n}[I(A_i = k)\log\{\boldsymbol{\beta}_{(k)}^{\mathrm{T}}\mathbf{B}(k, \mathbf{X}_i)\} - \boldsymbol{\beta}_{(k)}^{\mathrm{T}}\mathbf{B}(k, \mathbf{X}_i)], \tag{2.7}$$

for $k = 0, \ldots, K$, hence the minimizer is a root of (2.5). The utilization of the same basis of functions for both nuisance models is also used in Wang & Zubizarreta (2020). To further accommodate one's favorite propensity model, we can also make a linear combination of this model and any other relevant candidate model.

The asymptotic variance simplifies greatly when all models are correctly specified, and a local efficiency result is obtained.

Corollary 2.2.1. *Assume that (2.1) and (2.2) hold for all k and $\mathbf{x}$ and let $var(Y_i^k \mid \mathbf{X}_i) = v(k, \mathbf{X}_i)$. Then, under the regularity conditions of Theorem 2.2.1, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ has asymptotic normal distribution with mean zero and variance*

$$\boldsymbol{\Sigma} = \mathbf{C}(\boldsymbol{\beta}_0) - \mathbf{B}(\boldsymbol{\beta}_0)\{\mathbf{A}(\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}_0)^{-1}\mathbf{A}(\boldsymbol{\beta}_0)\}^{-1}\mathbf{B}(\boldsymbol{\beta}_0)^{\mathrm{T}},$$

where

$$\begin{aligned}
\mathbf{A}_k(\boldsymbol{\beta}_0) &= E\left\{-\frac{\mathbf{B}(k, \mathbf{X}_i)\pi'_{\boldsymbol{\beta}}(k, \mathbf{X}_i, \boldsymbol{\beta}_0)^{\mathrm{T}}}{\pi(k, \mathbf{X}_i, \boldsymbol{\beta}_0)}\right\}, \\
\mathbf{B}_k(\boldsymbol{\beta}_0) &= E\left\{-\frac{m(k, \mathbf{X}_i)\pi'_{\boldsymbol{\beta}}(k, \mathbf{X}_i, \boldsymbol{\beta}_0)^{\mathrm{T}}}{\pi(k, \mathbf{X}_i, \boldsymbol{\beta}_0)}\right\}, \\
\mathbf{V}_{kl}(\boldsymbol{\beta}_0) &= E\left[\left\{\frac{I(k=l)}{\pi(k, \mathbf{X}_i, \boldsymbol{\beta}_0)} - 1\right\}\mathbf{B}(k, \mathbf{X}_i)\mathbf{B}(l, \mathbf{X}_i)^{\mathrm{T}}\right]
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{C}_{kl}(\boldsymbol{\beta}_0) &= E\left\{I(k=l)\frac{m(k, \mathbf{X}_i)^2 + v(k, \mathbf{X}_i)}{\pi(k, \mathbf{X}_i, \boldsymbol{\beta}_0)} - m(k, \mathbf{X}_i)m(l, \mathbf{X}_i)\right\} \\
&\quad + E\left([m(k, \mathbf{X}_i) - E\{m(k, \mathbf{X}_i)\}][m(l, \mathbf{X}_i) - E\{m(l, \mathbf{X}_i)\}]\right).
\end{aligned}$$

Remark 2.2.1. *The variance $\boldsymbol{\Sigma}$ may be estimated without knowing or estimating $\boldsymbol{\alpha}$, by approximating the original definitions of the matrices involved, i.e. $\mathbf{B}(\boldsymbol{\beta}_0) \equiv E\{\partial \mathbf{g}_i(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}_0{}^{\mathrm{T}}\}$ and $\mathbf{C}(\boldsymbol{\beta}_0) \equiv E\{\mathbf{g}_i(\boldsymbol{\beta}_0)^{\otimes 2}\}$, instead of the expression involving $m(\cdot)$ and $v(\cdot)$ given in Corollary 2.2.1.*

Corollary 2.2.2. *Under the assumptions of Corollary 2.2.1, the variance of $\widehat{\boldsymbol{\theta}}$ attains the semiparametric efficiency bound $\boldsymbol{\Sigma}_{\text{eff}}$, where the (k, l) entry of $\boldsymbol{\Sigma}_{\text{eff}}$ is*

$$\boldsymbol{\Sigma}_{\text{eff,k,l}} = I(k=l)E\{v(k, \mathbf{X})/\pi(k, \mathbf{X})\} + E([m(k, \mathbf{X}) - E\{m(k, \mathbf{X})\}][m(l, \mathbf{X}) - E\{m(k, \mathbf{X})\}]).$$

Remark 2.2.2. *We point out an interesting generalization of the double robustness property in the special case where the set $\{0, 1, \ldots, K\}$ can be split into two non-overlapping subsets S_1, S_2, so that $\pi(k, \mathbf{x}, \boldsymbol{\beta}) = \pi(k, \mathbf{x}, \boldsymbol{\beta}_1)$ for all $k \in S_1$ and $\pi(k, \mathbf{x}, \boldsymbol{\beta}) = \pi(k, \mathbf{x}, \boldsymbol{\beta}_2)$ for all $k \in S_2$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are subvectors of $\boldsymbol{\beta}$ that do not share any components. In this case, based on (2.5), the properties established above are applicable in S_1 and S_2 separately. Hence, we can allow either one of the two models $\pi(k, \mathbf{x}, \boldsymbol{\beta}_1)$ and $\mathbf{B}(k, \mathbf{x})$ to be correct for $k \in S_1$, and either one of the two models $\pi(k, \mathbf{x}, \boldsymbol{\beta}_2)$ and $\mathbf{B}(k, \mathbf{x})$ to be*

correct for $k \in S_2$.

The choice of the basis functions $\mathbf{B}(k, \mathbf{x})$ is crucial but not intuitively obvious, since we do not know the underlying truth. With this in mind, we point out some relevant facts that can help both the implementation and the interpretation of the result. Note that $\mathbf{B}(0, \mathbf{x})$ spans the baseline conditional mean $m(0, \mathbf{x})$; and similarly for $k = 1, \ldots, K$. Thus, researchers can choose such bases to form $\mathbf{B}(0, \mathbf{x})$ and $\mathbf{B}(k, \mathbf{x}), k = 1, \ldots, K$, so that (2.5) gives the mean of the bases balanced across the treatments. For example, choosing $\mathbf{B}(k, \mathbf{x}) = \mathbf{x},\ k = 0, \ldots, K$, will result in balancing the mean of the covariates $\mathbf{X}_i$. In addition, if some functions are unlikely to belong to the bases of the baseline mean, but may still appear in the bases of the average treatment effects, say $\mathbf{B}^*(\mathbf{x})$ is such a function, then one can set $\mathbf{B}(k, \mathbf{x}), k = 1, \ldots, K$ as $[\mathbf{B}^{\mathrm{T}}(0, \mathbf{x}), \mathbf{B}^{*\mathrm{T}}(\mathbf{x})]^{\mathrm{T}}$. These components will then be balanced.

We end this section by proposing a possible extension: combining our method with outcome augmentation. The well-known augmented inverse probability weighting (AIPW) estimator is defined as

$$\widehat{\theta}_k \equiv n^{-1} \sum_{i=1}^{n} \left[\frac{I(A_i = k)}{\pi(k, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})} Y_i + \left\{ 1 - \frac{I(A_i = k)}{\pi(k, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})} \right\} \widehat{m}(k, \mathbf{X}_i) \right]. \tag{2.8}$$

It has been shown that when both the propensity score and the outcome mean models are correctly specified, the AIPW estimator achieves the same kind of optimality as our method; see Robins & Rotnitzky (1995) and Scharfstein et al. (1999). However, in finite sample situations, it is natural to expect that the AIPW estimator performs more efficiently in practice, since it has more parameters to fit. Hence, to further improve the performance, practitioners can consider implementing our method in combination with augmentation using (2.8), with $\widehat{\boldsymbol{\beta}}$ obtained by minimizing (2.6) and the fitted outcome models $\widehat{m}(k, \mathbf{X}_i)$. See Section 2.4 for its practical performance.

2.3 Continuous treatments

2.3.1 Balancing scores and preliminaries

We now consider a continually valued treatment A, say taking values a in $[0, 1]$. In this case, it is reasonable to assume that the potential outcome Y^a changes with a smoothly. We write Y^a as $Y(a)$ in a more conventional notation. Note that the observed outcome

for the ith observation, Y_i, is assumed to be $Y_i(a_i)$ when we observe $A_i = a_i$. We observe a random sample $(A_i, Y_i, \mathbf{X}_i), i = 1, \ldots, n$, where $\mathbf{X}_i \in \mathbb{R}^d$ is a vector of pre-treatment covariates observed for all units. Following the literature convention, we assume ignorability of the treatment assignment, in the sense that $E\{Y_i(a) \mid \mathbf{X}_i, A_i\} = E\{Y_i(a) \mid \mathbf{X}_i\}$, and the generalized propensity score is the conditional probability density function of the continuous treatment A_i given the covariates $\mathbf{X}_i$: $\pi_0(a, \mathbf{x}) \equiv f_{A|\mathbf{X}}(a, \mathbf{x}) > \delta > 0$ for all $a \in [0, 1]$ and all $\mathbf{x}$. We write the expected conditional potential outcome as $m(a, \mathbf{x}) \equiv E\{Y_i(a) \mid \mathbf{X}_i = \mathbf{x}\}$.

In such case, the parameter of interest is the treatment response function or the dose-response function, denoted as $\theta(a) = E\{Y_i(a)\}$ for $a \in [0, 1]$. The average causal effects between two treatment doses, say a and b, are obtained by taking their contrast $\theta(a) - \theta(b)$. We consider a parametric working model $\pi(a, \mathbf{x}, \boldsymbol{\beta})$ for the propensity score $\pi_0(a, \mathbf{x})$, where $\boldsymbol{\beta} \in \mathbb{R}^p$, and consider a set of basis functions $\mathbf{B}(a, \mathbf{x}) : \mathbb{R}^{d+1} \to \mathbb{R}^q$ aiming at spanning $m(a, \mathbf{x})$. Thus, correctly specified situations will be such that there exists $\boldsymbol{\beta}_0$ so that

$$\pi(a, \mathbf{x}, \boldsymbol{\beta}_0) = \pi_0(a, \mathbf{x}), \tag{2.9}$$

and there exists $\boldsymbol{\alpha}$ such that

$$\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{B}(a, \mathbf{x}) = m(a, \mathbf{x}), \tag{2.10}$$

for all $a \in [0, 1]$ and all $\mathbf{x}$. Misspecification, i.e. situations where one of (2.9) and (2.10) does not hold, will be allowed in the sequel.

The balancing consideration then leads us to the condition

$$\sum_{i=1}^{n} \left[\left\{ \frac{K_l(A_i - a)}{\pi(a, \mathbf{X}_i, \boldsymbol{\beta})} - 1 \right\} \mathbf{B}(a, \mathbf{X}_i) - \left\{ \frac{K_l(A_i - b)}{\pi(b, \mathbf{X}_i, \boldsymbol{\beta})} - 1 \right\} \mathbf{B}(b, \mathbf{X}_i) \right] = \mathbf{0}$$

for two arbitrary a, b values in $[0, 1]$. Following the same considerations as in Section 2.2, we strengthen the above requirement and consider the balancing equations

$$\sum_{i=1}^{n} \left\{ \frac{K_l(A_i - a)}{\pi(a, \mathbf{X}_i, \boldsymbol{\beta})} - 1 \right\} \mathbf{B}(a, \mathbf{X}_i) = \mathbf{0} \tag{2.11}$$

at all $a \in [0, 1]$. Here, $K_l(\cdot) = l^{-1} K(\cdot / l)$, where $K(\cdot)$ is a kernel function and l is a bandwidth. Kernel estimation was also used in Flores et al. (2012), but in a different

way. Practically, we propose to solve (2.11) at a set of chosen a values, typically those observed for A_i, and minimize

$$\sum_{j=1}^{n}\left\|\sum_{i=1}^{n}\left[\left\{\frac{K_l(A_i-A_j)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta})}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right\|_2^2\{\sum_{i=1}^{n}K_l(A_i-A_j)\} \tag{2.12}$$

with respect to $\boldsymbol{\beta}$ to get $\widehat{\boldsymbol{\beta}}$. Once we obtain $\widehat{\boldsymbol{\beta}}$, we estimate the causal parameter $\theta(a)$ with an inverse probability weighting estimator

$$\widehat{\theta}(a)=n^{-1}\sum_{i=1}^{n}\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}, \tag{2.13}$$

for any a within the range of observed values for A_i. Here, h is a bandwidth.

Remark 2.3.1. *The nonparametric estimator (2.13) can be viewed as an approximation of*

$$\frac{n^{-1}\sum_{i=1}^{n}Y_iK_h(A_i-a)/\pi(A_i,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}{n^{-1}\sum_{i=1}^{n}K_h(A_i-a)/\pi(A_i,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})},$$

which is the solution to

$$\min_c\sum_{i=1}^{n}\frac{(Y_i-c)^2K_h(A_i-a)}{\pi(A_i,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}.$$

Thus, we can understand (2.13) as a weighted local constant estimator of $\theta(a)$. Similar to the generalization from local constant to local polynomial estimators in nonparametrics, we can also generalize (2.13) to more sophisticated versions. For example, through obtaining $\widehat{c}_0$ from

$$\min_{c_0,c_1}\sum_{i=1}^{n}\frac{\{Y_i-c_0-c_1(A_i-a)\}^2K_h(A_i-a)}{\pi(A_i,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})},$$

we can obtain the weighted local linear estimator of $\theta(a)$.

2.3.2 Asymptotic properties

We now study the limiting properties of the estimator (2.13) using (2.12); see A.2 for proofs. Denote by $\boldsymbol{\beta}^*$ the probability limit of $\widehat{\boldsymbol{\beta}}$. If model (2.9) is correct, $\boldsymbol{\beta}^*=\boldsymbol{\beta}_0$,

otherwise $\boldsymbol{\beta}^*$ is the value that minimizes (2.12) at the population level, i.e. it minimizes

$$E_j\left(\left\|E_i\left[\left\{\frac{K_l(A_i-A_j)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta})}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right\|_2^2\{\sum_{i=1}^{n}K_l(A_i-A_j)\}\right) \tag{2.14}$$

with respect to $\boldsymbol{\beta}$. Here E_j means taking expectation over the jth observation. We list the following regularity conditions.

C0. $\boldsymbol{\beta}^*$ is the unique solution of $E\left[\left\{\frac{\pi_0(a,\mathbf{X})}{\pi(a,\mathbf{X},\boldsymbol{\beta})}-1\right\}\mathbf{B}(a,\mathbf{X})\right]=\mathbf{0}$.

C1. The kernel function $K(\cdot)\geqslant 0$ is bounded, twice differentiable with bounded first derivative, symmetric and has support on $(-1,1)$. It satisfies $\int_{-1}^{1}K(t)dt=1$.

C2. The bandwidth l satisfies $nl^4\to 0$ and $nl^2\to\infty$. The bandwidth h satisfies $h\to 0$ and $nh\to\infty$.

C3. The basis function $\mathbf{B}(a,\mathbf{x})$ is bounded.

C4. The propensity score $\pi(a,\mathbf{x},\boldsymbol{\beta})$ is differentiable with respect to $\boldsymbol{\beta}$ and a, is bounded away from zero, and its derivative with respect to a is bounded.

C5. $m(a,\mathbf{X}_i)$ is bounded, twice differentiable with respect to a, and the first derivative is bounded.

C6. $\sigma^2(A_i,\mathbf{X}_i)\equiv\text{var}(Y_i\mid A_i,X_i)$ is bounded.

These are typical regularity conditions. Similar to Condition A0 in the categorical treatment case, the uniqueness requirement in Condition C0 can be relaxed to local uniqueness. Moreover, with finite samples, C0 can be translated to: $\boldsymbol{\beta}^*$ is the unique solution of $E_i\left[\left\{\frac{K_l(A_i-A_j)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta})}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]=\mathbf{0}$ for $j=1,\ldots,n$, which is easier to fulfill. The existence of $\boldsymbol{\beta}^*$ is guaranteed when the propensity model $\pi(a,\mathbf{x},\boldsymbol{\beta})$ is correctly specified, and is a standard requirement when the number of equations qn is not larger than the length of $\boldsymbol{\beta}$. Thus, in the situation where we are not confident that a correct propensity model is used, we can always enrich the model to accommodate Condition C0. We start by giving the convergence rate of $\widehat{\boldsymbol{\beta}}$.

Lemma 2.3.1. *Denote by $\boldsymbol{\beta}^*$ the probability limit of $\widehat{\boldsymbol{\beta}}$. If model (2.9) is correct, $\boldsymbol{\beta}^*=\boldsymbol{\beta}_0$, otherwise $\boldsymbol{\beta}^*$ is the value that minimizes (2.14). Under regularity conditions C0 to C4, $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*=O_p(n^{-1/2})$.*

Condition C0 is not really necessary for Lemma 2.3.1. We can redefine $\boldsymbol{\beta}^*$ as the unique minimum of (2.14) and Lemma 2.3.1 still holds. Because the nonparametric estimation convergence rate is slower than $O_p(n^{-1/2})$, Lemma 2.3.1 indicates that we can fix $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^*$ in the following analysis as long as we let $nl^4 \to 0$, and the first order bias and variance property of $\widehat{\theta}(a)$ will not be affected.

Theorem 2.3.1. *Under regularity conditions C0 to C6, and if (2.9) holds, then the estimator $\widehat{\theta}(a)$ defined by (2.13) has asymptotic normal distribution with asymptotic bias and variance:*

$$\begin{aligned} E\{\widehat{\theta}(a)\} - \theta(a) &= \frac{h^2}{2} E\left[\frac{\partial^2\{\pi_0(a,\mathbf{X}_i)m(a,\mathbf{X}_i)\}}{\pi_0(a,\mathbf{X}_i)\partial a^2}\right]\int t^2 K(t)dt + O(h^4 + n^{-1/2}), \\ var\{\widehat{\theta}(a)\} &= \frac{\int K^2(t)dt}{nh} E\left\{\frac{m^2(a,\mathbf{X}_i) + \sigma^2(a,\mathbf{X}_i)}{\pi_0(a,\mathbf{X}_i)}\right\} + O(n^{-1}h^{-1/2}), \end{aligned}$$

where $\sigma^2(A_i, \mathbf{X}_i) = var(Y_i \mid A_i, \mathbf{X}_i)$.

Theorem 2.3.2. *Under regularity conditions C0 to C6, and if (2.10) holds, then the estimator $\widehat{\theta}(a)$ defined by (2.13) has asymptotic normal distribution with asymptotic bias and variance:*

$$\begin{aligned} E\{\widehat{\theta}(a)\} - \theta(a) &= \frac{h^2}{2} E\left[\frac{\partial^2\{\pi_0(a,\mathbf{X}_i)m(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)\partial a^2}\right]\int t^2 K(t)dt + O(h^4 + n^{-1/2}), \\ var\{\widehat{\theta}(a)\} &= \frac{\int K^2(t)dt}{nh} E\left[\frac{\pi_0(a,\mathbf{X}_i)\{m^2(a,\mathbf{X}_i) + \sigma^2(a,\mathbf{X}_i)\}}{\pi^2(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right] + O(n^{-1}h^{-1/2}). \end{aligned}$$

Theorems 2.3.1 and 2.3.2 together reflect a robust property of the proposed estimator, and give equivalent results when all nuisance models are correctly specified. Specifically, Theorem 2.3.1 describes the robustness to misspecification of the outcome models, in that as long as the propensity score is correctly specified, the estimation of the treatment response function is valid even if we do not assume a correct model for the outcome. This is because the propensity score balances any functions of the covariates. Theorem 2.3.2 allows for the misspecification of the propensity score, with the restriction that Condition C0 needs to hold. If we choose to ensure C0 through allowing sufficiently many model parameters, then $\boldsymbol{\beta}$ will have length $p = qn$, which practically means that the propensity score is non-parametrically estimated. For example, we can let $\pi(a_j, \mathbf{x}) = \boldsymbol{\beta}_{(j)}^{\mathrm{T}}\mathbf{B}(a_j, \mathbf{x})$, where $\boldsymbol{\beta}_{(j)}$ has dimension q. Then, solving (2.11) for all observed $a = a_j$ corresponds to

minimizing the loss function

$$\sum_{i=1}^{n}[K_l(A_i - a_j)\log\{\boldsymbol{\beta}_{(j)}^{\mathrm{T}}\mathbf{B}(a_j, \mathbf{X}_i)\} - \boldsymbol{\beta}_{(j)}^{\mathrm{T}}\mathbf{B}(a_j, \mathbf{X}_i)],$$

for $j = 1, \ldots, n$.

Finally, note here that the dose-response function $\theta(a)$ is estimated nonparametrically, and this estimation has bias of order h^2, although asymptotically vanishing, and there is the usual bias-variance trade-off. Next, we give a result useful for inference on a causal contrast $\theta(a) - \theta(b)$.

Theorem 2.3.3. *Under regularity conditions C0 to C6, and if either (2.9) or (2.10) hold, then $\widehat{\theta}(a) - \widehat{\theta}(b)$ defined by (2.13) is asymptotically a Gaussian process, and has asymptotic variance-covariance:*

$$\begin{aligned}
& cov\{\widehat{\theta}(a), \widehat{\theta}(b)\} \qquad (2.15)\\
= \; & (nh)^{-1}E\int_0^1 \frac{K(t)K(t+c)\{m^2(a, \mathbf{X}_i) + \sigma^2(a, \mathbf{X}_i)\}}{\pi(a, \mathbf{X}_i, \boldsymbol{\beta}^*)\pi(b, \mathbf{X}_i, \boldsymbol{\beta}^*)}\pi_0(a, \mathbf{X}_i)dt \\
& + n^{-1}E\int_0^1 K(t)K(t+c)\left\{2m(a, \mathbf{X}_i)m'_a(a, \mathbf{X}_i)\pi_0(a, \mathbf{X}_i) + m^2(a, \mathbf{X}_i)\pi'_{0a}(a, \mathbf{X}_i)\right. \\
& \left. + 2\sigma(a, \mathbf{X}_i)\sigma'_a(a, \mathbf{X}_i)\pi_0(a, \mathbf{X}_i) + \sigma^2(a, \mathbf{X}_i)\pi'_{0a}(a, \mathbf{X}_i)\right\} t/\{\pi(a, \mathbf{X}_i, \boldsymbol{\beta}^*)\pi(b, \mathbf{X}_i, \boldsymbol{\beta}^*)\}dt \\
& - n^{-1}\theta(a)\theta(b) + O(n^{-1}h + h^{-1}n^{-3/2}),
\end{aligned}$$

where $c \equiv (a - b)/h$.

Note that when $c \notin (-2, 1)$, $K(t)K(t + c) = 0$ for all t. Therefore, the covariance has order $O(n^{-1})$ if $c \notin (-2, 1)$ and $O\{(nh)^{-1}\}$ otherwise. Thus, comparing the term of order $O\{(nh)^{-1}\}$ in the covariance in Theorem 2.3.3 with the terms of the same order for the variances in Theorems 2.3.1 and 2.3.2, we see that when a and b are close to each other relative to h, the variance of the contrast $\widehat{\theta}(a) - \widehat{\theta}(b)$ is close to zero. However, when a and b are far apart, the variance of the contrast is dominated by the variance of $\widehat{\theta}(a)$ and $\widehat{\theta}(b)$.

Theorems 2.3.1, 2.3.2 and 2.3.3 provide theoretical properties of the leading orders of the bias, variance and covariance properties of the nonparametric estimators. In large samples, these results can be used to perform inference. Practically, however, inference based on these results is often insufficiently precise because the next order of the nonparametric analysis is only slightly smaller than the leading order. This phenomenon has been observed in many nonparametric or even semiparametric problems including

quantile regression, survival analysis, etc., and bootstrap is often used instead.

We now discuss how to practically implement our method. First, the bases functions can be chosen by noting that solving (2.11) is equivalent to balancing the mean of the bases $\mathbf{B}(a, \mathbf{X}_i)$ across all a. Hence, for example, letting $\mathbf{B}(a, \mathbf{x}) = [\mathbf{x}^{\mathrm{T}}, a, a^2, a^3]^{\mathrm{T}}$ yields that the mean of $\mathbf{X}_i$ is balanced while the dose-response function is the sum of a linear combination of the covariates and a cubic function of the treatment a. Further, to choose the bandwidth l in (2.12), note that $K_l(\cdot)$ is used to estimate the density of a, $\pi_0(\cdot)$. Hence, any data-driven bandwidth selection method for density estimation such as the plug-in method can be used; see Bowman et al. (1998) for in-depth discussions. Further, the bandwidth h in (2.13) can be chosen by using a standard bandwidth selection method, such as leave-one-out cross-validation. That is, the bandwidth h can be chosen as $\mathrm{argmin}_h \sum_{i=1}^n \{Y_i - \widehat{\theta}_{h,-i}(A_i)\}^2$, where $\widehat{\theta}_{h,-i}(A_i)$ is the estimate with the bandwidth h and the ith observation omitted. Lastly, for performing inference on the average causal effects, we provide the variance and the covariance estimator in Remark 2.3.2.

Remark 2.3.2. *The variance of $\widehat{\theta}(a)$ in Theorem 2.3.2 and the covariance of $\widehat{\theta}(a)$ and $\widehat{\theta}(b)$ in Theorem 2.3.3 can be estimated by*

$$\begin{aligned}
\widehat{var}\{\widehat{\theta}(a)\} &= \frac{\int_0^1 K^2(t)dt}{nh} \left\{ \sum_{i=1}^n \frac{K_h(A_i - a)}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})} \right\}^{-1} \sum_{i=1}^n \frac{K_h(A_i - a) Y_i^2}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})^2}, \\
\widehat{cov}\{\widehat{\theta}(a), \widehat{\theta}(b)\} &= \frac{\int_0^1 K(t)K(t + c)dt}{nh} \left\{ \sum_{i=1}^n \frac{K_h(A_i - a)}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})} \right\}^{-1} \sum_{i=1}^n \frac{K_h(A_i - a) Y_i^2}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}}) \pi(b, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})},
\end{aligned}$$

where $c \equiv (a - b)/h$.

2.4 Simulation experiments

2.4.1 Categorical treatments

To investigate the finite sample performance of our method for the categorical treatment case, we performed a first simulation study. We generate a five-dimensional covariate vector $\mathbf{X}$, where $X_1 = 1$, and X_2 to X_5 were generated independently from a normal distribution with mean 3 and variance 2. We set $K = 3$ and the propensity score $\pi_0(k, \mathbf{x}) = \exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_k)/\{1 + \sum_{k=1}^3 \exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_k)\}$ for $k = 1, 2, 3$, and let $\pi_0(0, \mathbf{x}) = 1 - \sum_{k=1}^3 \pi_0(k, \mathbf{x})$. Here, $\boldsymbol{\beta}_1 = (0, -0.3, 0.5, -0.25, -0.1)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1/2$, and $\boldsymbol{\beta}_3 = \boldsymbol{\beta}_1$. We set $m(k, \mathbf{x}) = \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{x}$, where $\boldsymbol{\alpha}_0 = (200, 0, 13.7, 13.7, 13.7)^{\mathrm{T}}$, $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0 + \mathbf{c}$ with $\mathbf{c} =$

$(0, 27.4, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_0 + \mathbf{c}$, and $\boldsymbol{\alpha}_3 = \boldsymbol{\alpha}_0 + \mathbf{c}/2$. Then we further generated Y_i^k's by adding a standard normal random noise to the true mean $m(k, \mathbf{x}_i)$.

In implementing the estimators, in addition to the ideal case where both the $\pi(\cdot)$ model and the basis for the $m(\cdot)$ model are correct, we also experiment with incorrectly specified models. In misspecifying the $\pi(\cdot)$ models, A was generated by $\pi(k, \mathbf{x}^*)$ where $\mathbf{x}^* = [1, \exp(x_2/3), x_3/\{1+\exp(x_2)\}+10, (x_2+x_4)/25+0.6, x_3+x_5+20]^{\mathrm{T}}$. In misspecifying the $m(\cdot)$ models, Y_i^k were generated using $m(k, \mathbf{x}_i^{**})$ where $\mathbf{x}^{**} = (1, x_2^2, x_3^2, x_4^2, x_5^2)$. However, we still used $\pi(k, \mathbf{x})$ as the propensity score or $\mathbf{x}$ as the basis. Note that this simulation design is identical to Fan et al. (2022) up to the treatments $k = 0, 1$. We set the groups with $k = 2, 3$ by modifying the outcome mean and the propensity model of the $k = 1$ group to provide further understanding on how such modifications affect the performance of the contrasts estimators.

We investigate four different scenarios: when both models are correct, when the $\pi(\cdot)$ model is misspecified, when the $m(\cdot)$ model is misspecified and when both models are misspecified. Note that our design is such that correctly specifying the basis for $m(\cdot)$ corresponds to balancing the first moments of the covariates. For comparison, in addition to our method (CB), we implemented the maximum likelihood estimation (ML) for the propensity score and compared the performance of the inverse probability weighting (IPW) estimators and the augmented inverse probability weighting (DR) estimators under the two approaches; for DR we used the R-package `PSweight` (Zhou et al. 2020). The results over 1000 replicates are displayed in Tables 2.1-2.4 for different sample sizes, where for each causal contrast $\theta_k - \theta_0$, $k = 1, 2, 3$, we provide absolute bias, standard deviation, mean squared errors (MSE) as well as average estimated standard deviation, and empirical coverage of the resulting 95% confidence interval. Remark 2.2.1 details how this inference can be performed.

The numerical results in Tables 2.1-2.4 confirm the theoretical robustness properties in the sense that smaller biases were observed when at least one of the models was correctly specified than when both models $\pi(\cdot)$ and $m(\cdot)$ are misspecified. Moreover, the estimators based on our covariate balancing (CB-IPW and CB-DR) yielded the same or lower variance and MSE than the maximum likelihood based estimators (ML-IPW and ML-DR). On the other hand, CB-IPW underperformed ML-DR, even though the two achieve the same kind of optimality in theory. That the combination of CB and DR showed the best performance in these experiments suggests that our covariate balancing strategy can improve the finite sample performance of estimators, which are asymptotically optimal. As expected from theory, empirical coverages matched the

Table 2.1: Results based on 1000 replicates for the estimation of contrasts $\theta_k - \theta_0$, $k = 1, 2, 3$. ML-IPW (DR) is the maximum likelihood based IPW (DR) estimator and CB-IPW (DR) the covariate balancing IPW (DR) method proposed. The basis of $m(\cdot)$ and the model $\pi(\cdot)$ are both correctly specified. Sample size $n = 1000$.

Method	|bias|	sd	MSE	$\widehat{\text{sd}}$	95%
		$\theta_1 - \theta_0$			
ML-IPW	1.729	2.175	7.719	2.113	0.949
CB-IPW	1.360	1.707	4.764	1.648	0.934
ML-DR	1.007	1.249	2.574	1.229	0.943
CB-DR	1.006	1.249	2.573	1.229	0.943
		$\theta_2 - \theta_0$			
ML-IPW	1.325	1.656	4.497	1.625	0.952
CB-IPW	1.221	1.513	3.780	1.510	0.953
ML-DR	1.007	1.253	2.586	1.228	0.938
CB-DR	1.007	1.253	2.586	1.228	0.938
		$\theta_3 - \theta_0$			
ML-IPW	1.217	1.495	3.714	1.475	0.957
CB-IPW	0.833	1.045	1.785	1.027	0.949
ML-DR	0.507	0.632	0.657	0.620	0.943
CB-DR	0.507	0.632	0.657	0.620	0.943

Table 2.2: Results based on 1000 replicates for the estimation of contrasts $\theta_k - \theta_0$, $k = 1, 2, 3$. The basis of $m(\cdot)$ is correctly and the model $\pi(\cdot)$ is wrongly specified. Sample size $n = 1000$.

Method	|bias|	sd	MSE	$\widehat{\text{sd}}$	95%
		$\theta_1 - \theta_0$			
ML-IPW	1.492	1.880	5.760	1.829	0.945
CB-IPW	1.541	1.692	5.239	1.773	0.923
ML-DR	1.005	1.252	2.579	1.229	0.937
CB-DR	1.005	1.252	2.579	1.229	0.937
		$\theta_2 - \theta_0$			
ML-IPW	1.583	1.997	6.491	1.963	0.957
CB-IPW	1.348	1.690	4.673	1.912	0.971
ML-DR	1.003	1.253	2.575	1.229	0.937
CB-DR	1.003	1.253	2.575	1.229	0.937
		$\theta_3 - \theta_0$			
ML-IPW	1.112	1.431	3.284	1.423	0.955
CB-IPW	0.997	1.152	2.322	1.326	0.953
ML-DR	0.505	0.633	0.657	0.621	0.939
CB-DR	0.505	0.633	0.657	0.621	0.939

Table 2.3: Results based on 1000 replicates for the estimation of contrasts $\theta_k - \theta_0$, $k = 1, 2, 3$. The basis of $m(\cdot)$ is wrongly and the model $\pi(\cdot)$ is correctly specified. Sample size $n = 1000$.

Method	|bias|	sd	MSE	$\widehat{\text{sd}}$	95%
		$\theta_1 - \theta_0$			
ML-IPW	14.173	18.113	528.941	17.166	0.942
CB-IPW	11.932	14.955	366.044	14.347	0.937
ML-DR	10.505	13.177	284.009	12.421	0.939
CB-DR	10.158	12.666	263.619	12.412	0.945
		$\theta_2 - \theta_0$			
ML-IPW	10.900	13.728	307.252	12.818	0.944
CB-IPW	10.320	12.958	274.417	12.207	0.939
ML-DR	9.692	12.054	239.234	11.186	0.937
CB-DR	9.537	11.862	231.657	11.200	0.939
		$\theta_3 - \theta_0$			
ML-IPW	10.074	12.717	263.204	12.180	0.941
CB-IPW	8.237	10.379	175.563	9.965	0.932
ML-DR	7.713	9.927	158.031	9.127	0.927
CB-DR	7.387	9.455	143.962	9.064	0.936

Table 2.4: Results based on 1000 replicates for the estimation of contrasts $\theta_k - \theta_0$, $k = 1, 2, 3$. The basis of $m(\cdot)$ and the model $\pi(\cdot)$ are both wrongly specified. Sample size $n = 1000$.

Method	|bias|	sd	MSE	$\widehat{\text{sd}}$	95%
		$\theta_1 - \theta_0$			
ML-IPW	11.616	14.086	333.360	13.869	0.940
CB-IPW	12.598	13.350	336.929	13.392	0.897
ML-DR	8.978	11.175	205.500	10.969	0.930
CB-DR	8.970	11.052	202.604	10.969	0.933
		$\theta_2 - \theta_0$			
ML-IPW	11.040	13.502	304.179	13.540	0.944
CB-IPW	11.077	13.703	310.473	13.789	0.947
ML-DR	9.723	11.969	237.792	11.800	0.937
CB-DR	9.515	11.762	228.878	11.782	0.946
		$\theta_3 - \theta_0$			
ML-IPW	8.730	10.865	194.267	10.884	0.960
CB-IPW	8.699	9.947	174.606	10.462	0.945
ML-DR	6.706	8.357	114.802	8.094	0.942
CB-DR	6.622	8.229	111.572	8.090	0.941

nominal level of 95%, except for when all models were misspecified. In most cases, the empirical coverage was slightly lower than the nominal level, partially due to the bias. Nevertheless, the coverages are still reasonably close to the nominal level because these biases are small compared to the length of the confidence intervals.

2.4.2 Continuous treatments

To assess the performance of the proposed methods under continuous treatment, we experiment with both linear and nonlinear outcome models. In the nonlinear design, we generate a five-dimensional covariate vector $\mathbf{X}$, where $X_1 = 1$ and $(X_2, X_3, X_4, X_5)^{\mathrm{T}}$ follows a multivariate standard normal distribution. Thus, these covariates have mean zero, variance 1 and are independent of each other. The true propensity score function is

$$\pi_0(a, \mathbf{x}) = \frac{\Gamma(15)}{\Gamma[15\lambda(\mathbf{x})]\Gamma[15\{1-\lambda(\mathbf{x})\}]}\Big(\frac{a}{20}\Big)^{15\lambda(\mathbf{x})-1}\Big(1-\frac{a}{20}\Big)^{15\{1-\lambda(\mathbf{x})\}-1}\frac{1}{20}.$$

Note that this is the probability density function of A when $A/20$ follows a beta distribution with parameters $15\lambda(\mathbf{x})$ and $15\{1-\lambda(\mathbf{x})\}$, where

$$\mathrm{logit}\{\lambda(\mathbf{x})\} = (-0.8, 0.1, 0.1, -0.1, 0.2)\mathbf{x}.$$

We further generate the response Y from a Bernoulli distribution with probability $m_1(A, \mathbf{X}) \equiv \mathrm{expit}\{\mu(A, \mathbf{X})\}$, where

$$\mu(a, \mathbf{x}) = (1, 0.2, 0.2, 0.3, -0.1)\mathbf{x} + a(0.1, -0.1, 0, 0.1, 0)\mathbf{x} - 0.13^3a^3.$$

This simulation design is identical to that of Kennedy et al. (2017). In the linear design, the response is generated from a normal distribution with mean $m_2(A, \mathbf{X})$ and variance 0.16, where $m_2(a, \mathbf{x}) = \{\mu(a, \mathbf{x}) + 15\}/20$.

Two different types of IPW estimators are implemented in the linear and nonlinear outcome cases. These are respectively a maximum likelihood based inverse probability weighting estimator and the proposed robust balancing estimator. For the former, we used a maximum likelihood approach to estimate the parameter of the propensity score. For the balancing estimator, (2.12) is minimized where the bandwidth l is set to $3n^{-1/3}$. In the nonparametric estimation of $\theta(a)$ in (2.13), both the local constant and local linear estimators given in Remark 2.3.1 are implemented and h is selected

by the leave-one-out cross-validation and the one-sided cross-validation (Hart & Yi 1998). For comparison, the inverse probability weighting and the doubly robust estimator given in Kennedy et al. (2017) are also implemented using the R-package `npcausal` (github.com/ehkennedy/npcausal).

For the linear outcome case, the estimators are assessed in four different scenarios; where both models are correct and where either (or all) of the models is (are) misspecified. We use the basis of $\mu(a, \mathbf{x})$ as the basis of the outcome model. In misspecifying either the $\pi(\cdot)$ or $m(\cdot)$ model, we replaced the covariates with $\mathbf{x}^*$ as in Kang & Schafer (2007), with

$$\mathbf{x}^* = \left\{1, e^{x_2/2}, \frac{x_3}{1+\exp(x_2)} + 10, (x_2 x_4/25 + 0.6)^3, (x_3 + x_5 + 20)^2\right\}^{\mathrm{T}}.$$

In addition, the misspecified $m_i(\cdot)\,(i = 1, 2)$ has no cubic term of a in its bases. In fact we used the same construction for the nonlinear outcome model. However, we point out that in this scenario the outcome model basis is never correctly specified, while the propensity score model is either correct or incorrect.

We generated the simulated data with sample sizes $n = 500, 1000, 2000$ and the result is based on 1000 replicates. Figure 2.1 illustrates the simulated data with the nonlinear outcome model and the empirical coverage of the proposed estimator under $n = 1000$. We assessed the performance of each estimator by calculating the integrated absolute bias and the integrated root-mean-squared error (RMSE), where

$$\begin{aligned} \text{bias} &= \int_{\mathcal{A}^*} \left| E\{\hat{\theta}(a)\} - \theta(a) \right| f_A(a) da, \\ \text{RMSE} &= \int_{\mathcal{A}^*} E\left[\{\widehat{\theta}(a) - \theta(a)\}^2\right]^{1/2} f_A(a) da, \end{aligned}$$

where $\mathcal{A}^*$ is a trimmed support of A which excludes 10% mass on the boundaries.

The results are given in Tables 2.5 and 2.6. The integrated absolute bias and the integrated RMSE were numerically calculated and presented with the integrated RMSE in parentheses. For ease of presentation, both measures are multiplied by 100. These results confirm that the proposed estimator is robust. In addition, as seen in Table 2.5, we found that our estimator showed robust performance even under the nonlinear outcome design where (2.10) did not hold, meaning that none of the four cases used the true basis of the outcome model. Among the balancing estimators, when all nuisance models were correctly specified, the variant using local linear fit and one-sided CV seemed

Table 2.5: Results based on 1000 replicates for continuous treatment case, and nonlinear outcome model. Integrated absolute bias and integrated RMSE (in parentheses). ML-IPW is the maximum likelihood based IPW estimator and CB-IPW the robust balancing-IPW method proposed (2.12-2.13).

$n = 500$

		π, m correct	π correct	m correct	none correct
IPW of Kennedy		na	3.33 (4.95)	na	3.00 (4.81)
DR of Kennedy		1.09 (3.31)	2.05 (3.75)	1.07 (3.31)	2.55 (4.02)
		π correct		none correct	
ML-IPW	Constant, CV	na	0.52 (4.52)	na	1.21 (4.40)
	Constant, OSCV	na	0.39 (4.23)	na	1.49 (4.42)
	Linear, OSCV	na	0.40 (4.08)	na)	1.99 (4.45)
CB-IPW	Constant, CV	0.38 (4.24)	0.26 (4.32)	1.15 (4.18)	1.23 (4.25)
	Constant, OSCV	0.28 (4.05)	0.31 (4.18)	1.41 (4.26)	1.52 (4.35)
	Linear, OSCV	0.69 (3.91)	0.82 (4.09)	1.86 (4.22)	1.99 (4.34)

$n = 1000$

		π, m correct	π correct	m correct	none correct
IPW of Kennedy		na	3.15 (4.11)	na	2.80 (3.91)
DR of Kennedy		0.97 (2.60)	1.88 (3.16)	0.94 (2.37)	2.36 (3.28)
		π correct		none correct	
ML-IPW	Constant, CV	na	0.39 (3.26)	na	1.32 (3.30)
	Constant, OSCV	na	0.46 (2.88)	na	1.42 (3.23)
	Linear, OSCV	na	0.48 (2.80)	na	1.96 (3.41)
CB-IPW	Constant, CV	0.27 (3.08)	0.20 (3.15)	1.27 (3.13)	1.34 (3.19)
	Constant, OSCV	0.29 (2.78)	0.20 (2.89)	1.37 (3.08)	1.46 (3.17)
	Linear, OSCV	0.68 (2.72)	0.69 (2.88)	1.85 (3.20)	1.97 (3.32)

$n = 2000$

		π, m correct	π correct	m correct	none correct
IPW of Kennedy		na	3.02 (3.62)	na	2.65 (3.44)
DR of Kennedy		0.79 (1.83)	1.76 (2.58)	0.78 (1.81)	2.37 (3.82)
		π correct		none correct	
ML-IPW	Constant, CV	na	0.33 (2.44)	na	1.45 (2.76)
	Constant, OSCV	na	0.56 (2.09)	na	1.41 (2.57)
	Linear, OSCV	na	0.54 (1.97)	na	2.00 (2.89)
CB-IPW	Constant, CV	0.22 (2.30)	0.19 (2.43)	1.41 (2.59)	1.47 (2.66)
	Constant, OSCV	0.39 (1.95)	0.26 (2.12)	1.36 (2.39)	1.44 (2.49)
	Linear, OSCV	0.66 (1.91)	0.70 (2.15)	1.91 (2.68)	2.00 (2.81)

Note: "na" stands for "not applicable".

Table 2.6: Results based on 1000 replicates for continuous treatment case, and linear outcome model. Integrated absolute bias and integrated RMSE (in parentheses). ML-IPW is the maximum likelihood based IPW estimator and CB-IPW the robust balancing-IPW method proposed (2.12-2.13).

$n = 500$

		π, m correct	π correct	m correct	none correct
IPW of Kennedy		na	3.02 (5.31)	na	2.58 (4.02)
DR of Kennedy		0.58 (2.60)	0.72 (2.69)	0.64 (2.55)	0.90 (2.64)
ML-IPW	Constant, CV	na	0.26 (3.55)	na	0.28 (3.55)
	Constant, OSCV	na	0.07 (3.64)	na	0.55 (3.74)
	Linear, OSCV	na	0.18 (3.36)	na	0.68 (3.44)
CB-IPW	Constant, CV	0.23 (3.29)	0.17 (3.34)	0.27 (3.21)	0.29 (3.28)
	Constant, OSCV	0.12 (3.55)	0.21 (3.58)	0.53 (3.56)	0.56 (3.58)
	Linear, OSCV	0.25 (3.23)	0.33 (3.27)	0.65 (3.26)	0.68 (3.30)

$n = 1000$

		π, m correct	π correct	m correct	none correct
IPW of Kennedy		na	2.96 (4.82)	na	2.55 (3.33)
DR of Kennedy		0.44 (1.92)	0.62 (1.97)	0.48 (1.85)	0.78 (1.98)
ML-IPW	Constant, CV	na	0.29 (2.55)	na	0.27 (2.52)
	Constant, OSCV	na	0.10 (2.52)	na	0.46 (2.61)
	Linear, OSCV	na	0.07 (2.31)	na	0.58 (2.43)
CB-IPW	Constant, CV	0.23 (2.34)	0.19 (2.39)	0.26 (2.28)	0.26 (2.32)
	Constant, OSCV	0.04 (2.43)	0.05 (2.46)	0.44 (2.46)	0.44 (2.48)
	Linear, OSCV	0.15 (2.21)	0.16 (2.27)	0.56 (2.27)	0.56 (2.31)

$n = 2000$

		π, m correct	π correct	m correct	none correct
IPW of Kennedy		na	2.93 (3.44)	na	2.43 (2.97)
DR of Kennedy		0.41 (1.45)	0.60 (1.55)	0.43 (1.40)	0.75 (1.57)
ML-IPW	Constant, CV	na	0.22 (1.84)	na	0.32 (1.84)
	Constant, OSCV	na	0.12 (1.79)	na	0.42 (1.85)
	Linear, OSCV	na	0.09 (1.70)	na	0.57 (1.80)
CB-IPW	Constant, CV	0.18 (1.72)	0.15 (1.74)	0.29 (1.66)	0.29 (1.67)
	Constant, OSCV	0.08 (1.72)	0.06 (1.76)	0.40 (1.74)	0.40 (1.76)
	Linear, OSCV	0.14 (1.61)	0.14 (1.65)	0.54 (1.68)	0.55 (1.71)

Note: "na" stands for "not applicable".

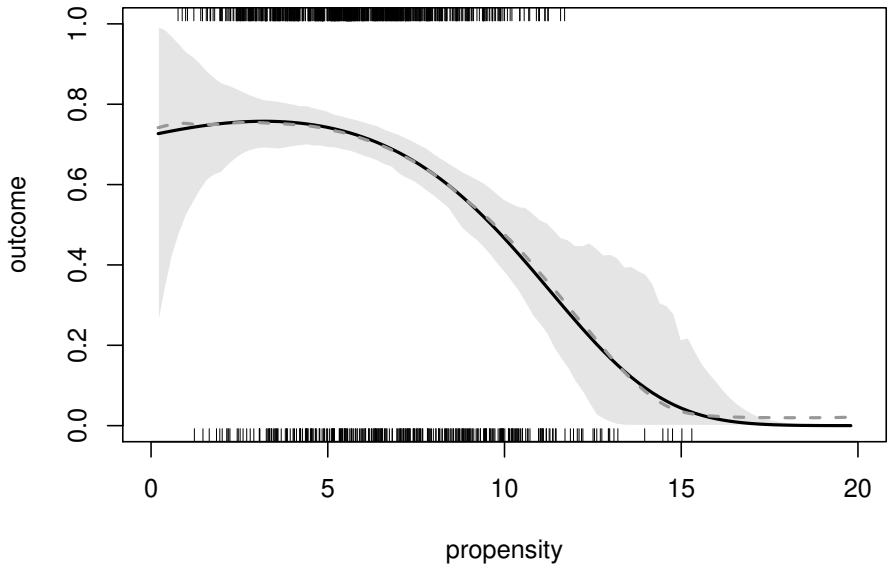


Figure 2.1: Simulation in the continuous nonlinear outcome case. Rug: one simulated data set with n=1000; solid: true outcome; dotted: mean of the estimates, i.e., $\frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t(a)$, using local constant estimation and CV, and $T = 1000$; filled: 5% and 95% quantiles of $\hat{\theta}_t(a)$.

to perform best in terms of bias and RMSE. The balancing method also had both lower bias and RMSE than the IPW estimators. We note that the bias was most sensitive to the specification of the propensity score model. In all cases, the proposed estimator outperformed the Kennedy et al. (2017) estimator in terms of bias, although RMSE Kennedy's double robust estimator had the lowest RMSE. In this regard, as for the categorical case, this estimator can be considered a benchmark since, unlike the introduced balancing estimators, it also fits outcome models.

2.5 Data application

As a case study, we investigate the effect of Body Mass Index (BMI) on self-reported health (SRH) decline. This analysis is based on data from the Survey of Health, Aging and Retirement in Europe (SHARE). This is an interview-based longitudinal survey of individuals of age 50 years or older (Börsch-Supan et al. 2013). Here we use data on women from three countries (Sweden, Netherlands, Italy) that participate in waves 1

and 5 of the SHARE study. Wave 1 data collected in 2004 serve as the baseline, and individuals are followed up at wave 5, collected in 2013. We are interested in estimating the average causal effect of BMI (a continuous valued treatment with range 15.62-49.60 in the data) on SRH decline between baseline and follow-up. SRH is measured by asking the question "Would you say your health is: excellent, very good, good, fair or poor?" Despite its unspecific nature, SRH has been found to predict mortality well in many studies (Idler & Benyamini 1997), and is thus considered an important health indicator. SRH decline is here defined as a binary variable which, for the respondents reporting "excellent, very good, or good health" at baseline, will take value one if they changed their answer to "fair or poor health" at follow-up, and 0 otherwise. The resulting sample of complete cases consists of 1530 participants. In Genbäck et al. (2018), predictors of SRH decline were investigated using logistic regression, and it was found that BMI measured at baseline was a significant (5% level) predictor of SRH decline. Here we aim at sharpening this analysis by studying BMI as a causal agent of SRH decline. We use the introduced covariate balancing procedure for causal inference. The covariates observed at baseline that we use for balancing are age (years), whether the participant responded to the SRH question at the beginning of the interview (or the end), socio-economic variables (education level, make ends meet easily), cognitive function variables (numeracy test, date orientation question), health variables (number of chronic diseases, number of mobility problems, depression measure, maximum grip strength, limitation in normal activities), and lifestyle variables (smoking habits, alcohol usage, physical activities). We refer to Genbäck et al. (2018) for a detailed description of these covariates. Encouraged by GBD 2015 Obesity Collaborators (2017) and Ng et al. (2016), our analysis is based on the following model for $A = (\text{BMI} - 15)/40$ given the covariate vector $\mathbf{x}$:

$$\begin{aligned} \pi_0(a, \mathbf{x}) &= \frac{\Gamma(\phi)}{\Gamma[\phi\lambda(\mathbf{x})]\Gamma[\phi\{1-\lambda(\mathbf{x})\}]} a^{\phi\lambda(\mathbf{x})-1}(1-a)^{\phi\{1-\lambda(\mathbf{x})\}-1}, \\ \text{logit}\{\lambda(\mathbf{x})\} &= \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}, \\ \boldsymbol{\beta} &= (\boldsymbol{\gamma}, \phi). \end{aligned}$$

The basis functions for the outcome model are chosen to be $\mathbf{B}(a, \mathbf{x}) = (\mathbf{x}, a, a^2, a^3)$. A value for $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\gamma}^{(0)}, \phi^{(0)})$ is obtained with maximum likelihood estimation, and used as the starting value when solving the balancing equations (2.12), with the bandwidth $l = 6n^{-1/3}$. For nonparametric estimation of $\theta(a)$ in (2.13), the local constant estimator given in Remark 2.3.1 is used for simplicity, where h was selected by one-sided cross-validation (Hart & Yi 1998).

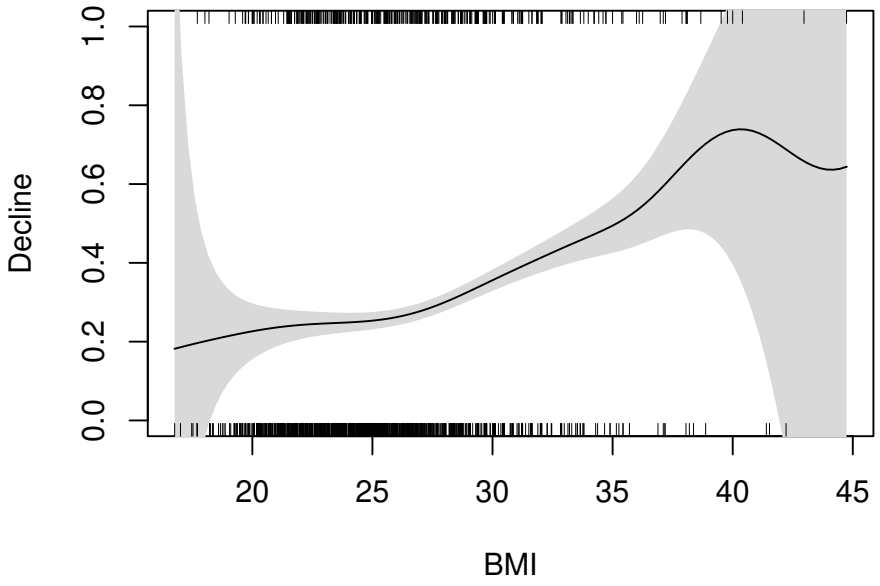


Figure 2.2: Effect of BMI on SRH decline. Rug: the observations; solid: the estimated average treatment effect curve; filled: the estimated pointwise confidence band.

Figure 2.2 displays the estimated effect curve of BMI on SRH decline. Confidence bands are obtained using the variance estimates described in A.2.4. Overall, the effect curve is nonlinear. Specifically, we observe that BMI has no significant effect for values of BMI considered as normal (i.e. below 25); the confidence band of the probability of decline contains the horizontal line. However, in the range of BMIs considered as overweight (BMI larger than 25), an increase in the probability of SRH decline is observed, reflecting a causal effect of BMI on the probability of SRH decline. The causal interpretation of this effect relies on the assumptions made. The principle amongst these assumptions are, that all confounders have been observed, and that a well-defined intervention on BMI corresponds to the effect measured (Hernán & Taubman 2008). Nevertheless, the results are in line with earlier studies pointing at a wide range of health risks of excess weight and obesity (GBD 2015 Obesity Collaborators 2017).

2.6 Discussion

We have introduced novel robust estimation and inference tools for multi-level treatments. For continuous treatments our proposal together with that of Kennedy et al. (2017) are, to the best of our knowledge, the only robust methods which model the causal dose-response curve nonparametrically. Our results expand the recent important developments given by Fan et al. (2022). For both the categorical and continuous treatment cases, we achieve robustness by balancing basis functions for the outcome models when fitting a generalized propensity score model which is either correct or sufficiently rich. While the estimator proposed is locally efficient for the categorical case, asymptotic efficiency is not relevant for the continuous case where the parameter of interest is a function of the dose and is estimated non-parametrically.

The proposal differs from earlier double robust estimation methods in that it does not need to estimate the parameters involved in the outcome model. This is an advantage when outcome is not observed or the outcome model is difficult to fit, due to model complexity, data size, numerical stability or other computational issues. Further, if fitting the outcome model is not an issue, then the proposed procedure can be done in combination with the classical augmented estimators. Our simulation results indicate that this results in improved finite sample performance. Our work contrasts with the widespread practice of using simple (e.g. linear or logistic linear) models for the propensity score with matching estimators, and assuming that balance in the joint distribution of the covariates is achieved (e.g., Waernbaum 2010, Rubin & Thomas 2000). However, balancing the joint distribution is not necessary, and in exchange, more elaborate demands are placed on the propensity score. This paper clarifies which functions of the covariates are sufficient to balance, if one is to achieve consistency and, in the categorical treatment case, local efficiency.

A natural development would be to extend the current methods to estimating average treatment effects on the treated or estimating quantile treatment effects. Also, in high-dimensional settings ($d \approx n$), it has recently been shown that bias due to regularization in estimating correctly specified linear outcome models can be corrected by using relevant weights which are not necessarily based on the true propensity score (Athey et al. 2018); see also, e.g., Farrell (2015) Dukes et al. (2020) and Antonelli et al. (2022) for double robust estimation with many covariates. An interesting future direction of research is whether one can generalize the results presented herein to high-dimensional situations, balancing many basis functions for the outcome models by using, e.g., regularized GMM

techniques (Belloni et al. 2018).

2.7 Acknowledgments

This chapter is based upon work supported by the National Science Foundation, the National Institute of Health, the Marianne and Marcus Wallenberg Foundation, and the Swedish Research Council. Any opinions, findings, and conclusions or recommendations expressed in this chapter are those of the authors, and do not necessarily reflect the views of the National Science Foundation, the National Institute of Health, the Marianne and Marcus Wallenberg Foundation, and the Swedish Research Council.

Chapter 3

Semiparametric approach to estimation of marginal and marginal quantile effects

3.1 Introduction

Generalized linear models (GLMs) (McCullagh & Nelder 1989) are arguably the most frequently used models in econometrics and statistics and their applications. With response variable $Y \in \mathbb{R}$ and covariates $\mathbf{X} \in \mathbb{R}^p$, GLMs have the familiar form

$$f_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}, \phi) = \exp\left\{\frac{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} - b(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{a(\phi)} + c(y, \phi)\right\}, \tag{3.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown regression coefficient vector, and ϕ is typically an unknown scalar parameter, if it is present. Here, the functions $a(\cdot), b(\cdot)$, and $c(\cdot)$ have pre-specified forms, and various choices of these functions lead to different GLMs extensively studied in the econometric and statistical literature.

While a large body of the classic literature is mainly concerned with the estimation and the inference of the model parameters $\boldsymbol{\beta}$ and ϕ in (3.1), an equally important target of research is to estimate marginal mean effect, say $\boldsymbol{\xi}$, defined as

$$\boldsymbol{\xi} \equiv E\left\{\frac{\partial E(Y \mid \mathbf{X})}{\partial \mathbf{X}}\right\}. \tag{3.2}$$

The marginal mean effect is a quantity of important meaning in several fields, in particular econometrics, and is used to capture the average rate of change of the regression mean with respect to the covariates (Greene 2000).

A complementary, less known but important quantity is the marginal quantile effect, defined as

$$\boldsymbol{\eta}_\tau \equiv E\left\{\frac{\partial Q_\tau(Y \mid \mathbf{X})}{\partial \mathbf{X}}\right\} \tag{3.3}$$

for a continuous response variable Y, where $\tau \in (0,1)$ is the quantile level and $Q_\tau(Y \mid \mathbf{x})$ is the τth conditional quantile of Y given $\mathbf{x}$. Here, we restrict our attention to the continuous response case only because the definition of conditional quantile in the discrete case is not universally agreed upon in the literature (Parzen 2004). The marginal quantile effect measures the average rate of change in the conditional quantiles, and can be equivalently written as $E\{\boldsymbol{\beta} Q'_\tau(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}$, where $Q'_\tau(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ is the derivative of $Q_\tau(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ with respect to $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. Unlike in the familiar quantile regression, the difference in marginal quantile effects at different quantile levels is caused not by quantile specific coefficients since $\boldsymbol{\beta}$ is fixed in the GLM, but by the baseline function $c(y,\phi)$. Obviously, our consideration of marginal quantile effect is a natural consequence of considering quantile regression (Koenker 2005), which is a key tool widely used as an alternative to the traditional mean regression models. Instead of modeling the conditional mean, the quantile regression estimates the relationship between the covariates and the response through the conditional quantile $Q_\tau(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, and considering its corresponding rate of change directly leads to the marginal quantile effect defined in (3.3). Comparing (3.2) and (3.3), it is clear that $\boldsymbol{\xi}$ and $\boldsymbol{\eta}_\tau$ are in fact the marginal mean effect and the marginal quantile effect respectively, while the marginal mean effect is customarily referred to as the marginal effect in the literature, which is what we will use from here on.

Marginal effects have a direct interpretation as measures of risk. For instance, in logistic regression, especially in the medical literature, it is customary to misinterpret odds ratios as measures of risk. Instead, marginal effects and marginal quantile effects represent changes in the probability of the occurrence of a binary event with respect to given risk factors and have a direct interpretation on the probability scale. This point has been stressed in the medical literature, e.g., by Norton et al. (2019). The marginal effect also captures how strongly the mean treatment outcome relies on different covariates on average. This further leads to the understanding on how the average treatment effect is affected by covariates. Treatment effect is directly linked to causal inference which is currently one of the trending topics in Economics and Statistics, as further confirmed by the 2021 Nobel Prize in Economics based on the seminal paper by Imbens & Angrist

(1994). In fact, the estimation of marginal effects has been studied by many earlier works in the literature, where the name "average mean derivatives estimation" was adopted. For example, Härdle & Stoker (1989) systematically studied the estimation problem, and Newey & Stoker (1993) studied semiparametric efficiency properties of weighted average mean derivative estimators. Likewise, the marginal quantile effect, formerly termed as average quantile derivatives, has also been proposed and studied as an alternative to the marginal effect by Chaudhuri et al. (1997). Following these earlier works, different approaches to estimating marginal effects under various settings have been proposed, see, e.g., Hristache et al. (2001), Cattaneo et al. (2010), and Cattaneo et al. (2013). The estimation of marginal effects is also quite widespread in medical economics concerning health outcomes. For instance, a two-year study performed at the University of Chicago focuses on the marginal effect on expenditures and length of stays in hospitals, rather than on parameter estimation of GLM; see Basu & Rathouz (2005). For the same study, Manning et al. (2005) estimate marginal effects on inpatient expenditures using a parametric family of distributions, a three parameter generalized Gamma distribution, which is an extension of the GLMs. Marginal effects are also frequently the target of study in economics. Parametric estimation of marginal effects in microeconometrics concerning the female labor force participation for 872 women from Switzerland can be found in Gerfin (1996) and Kleiber & Zeileis (2008). We will use this example as a benchmark to compare our nonparametric estimation in Section 3.5. In a GLM, the marginal effect happens to be identical to the regression coefficient $\boldsymbol{\beta}$ multiplied by average conditional variance, i.e. $\boldsymbol{\xi} = \boldsymbol{\beta} E\left\{\mathrm{var}(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right\}$. Similarly for the marginal quantile effect, we have $\boldsymbol{\eta}_\tau = \boldsymbol{\beta} E\left\{Q'_\tau(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right\}$. This allows for their straightforward estimation. Standard software packages provide direct estimation of marginal effects under GLMs, including the **margins** command in Stata and the package **mfx** in R (Fernihough 2019).

However, despite its central role in the literature and its wide application, GLM has its restrictions. Indeed, GLM prespecifies the three functions $a(\cdot), b(\cdot)$, and $c(\cdot)$ hence is purely parametric, therefore the model is susceptible to model misspecification. In addition, to facilitate computation, the choices of the three functions are often out of convenience, which further increases the chance of mis-modeling. To take into account the possibly large impact of deviation from the distributional assumptions on the parameter estimates and on the corresponding inference based on GLMs, a broad stream of the literature in the past decades has focused on robust methods for the estimation of the parameters; see e.g. Stefanski et al. (1986), Künsch et al. (1989), and Cantoni &

Ronchetti (2001) for a general class of optimally bounded influence estimators and tests for the parameters of GLMs. The corresponding formulas for the estimation of marginal effects and marginal quantile effects remain the same as the parametric case, but with the robust estimates of the parameters.

Here we go one step further in relaxing the parametric form of GLMs and consider a semiparametric generalized linear model (sGLM), where we assume

$$f_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}, c) = \frac{\exp\left\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + c(y)\right\}}{\int \exp\left\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + c(y)\right\} d\mu(y)}, \tag{3.4}$$

and leave $c(\cdot)$ unspecified. Here and in the following text, $\mu(\cdot)$ denotes the Lebesgue measure for a continuous variable y and the counting measure for a discrete y. For identifiability, we fix $c(0) = 0$ and for convenience, we assume each component of $\mathbf{x}$ to be centered so $E(\mathbf{X}) = \mathbf{0}$. Our focus is to estimate marginal effects and marginal quantile effects of (3.4), i.e. $\boldsymbol{\xi}$ defined in (3.2) and $\boldsymbol{\eta}_\tau$ in (3.3).

Model (3.4) is not entirely new, and has been proposed and studied in Luo & Tsai (2012) and Huang & Rathouz (2012) as a semiparametric proportional likelihood ratio model. Ning et al. (2017) studied high dimensional issue in (3.4), while Lin et al. (2021) studied the estimation of the parameter $\boldsymbol{\beta}$ in (3.4). Different from these existing works, our main focus is not in the model parameter $\boldsymbol{\beta}$, but in the marginal effect $\boldsymbol{\xi}$ and the marginal quantile effect $\boldsymbol{\eta}_\tau$.

In this context, we develop an approximate likelihood procedure to estimate both the model parameters $\boldsymbol{\beta}$ and $c(\cdot)$ in (3.4), which allows us to subsequently estimate the marginal effect $\boldsymbol{\xi}$ and the marginal quantile effect $\boldsymbol{\eta}_\tau$. We show that despite of the apparent difference of the estimation variability results from the theoretical efficiency bounds, our estimators are in fact semiparametricly efficient for both $\boldsymbol{\xi}$ and $\boldsymbol{\eta}_\tau$. We point out that our method differs from Lin et al. (2021) primarily in how to estimate the nonparametric component $c(y)$, which in turn has significant influence on estimation of the model parameters and functionals. Due to the special structure of model (3.4), kernel estimation for $c(y)$ is difficult to implement without giving up the model information and resorting to model-free estimation such as the Nadaraya-Watson estimator, hence the estimation of $c(y)$ in Lin et al. (2021) is not ideal. This leads to efficiency loss of the $\boldsymbol{\beta}$ estimation in finite samples, and greatly affects the efficiency of the $\boldsymbol{\xi}$ and $\boldsymbol{\eta}_\tau$ estimation. The efficiency gain from our method is demonstrated by simulation experiments in Section 3.4.

Specifically, our main contributions to the literature are the following. First, we

define and consider estimating marginal effects and marginal quantile effects in a semiparametric model, which have been studied by many earlier works in Economics and Statistics under various settings and are receiving more attention due to the recent interest in mean treatment effect or mean outcome estimation. Second, we discover that in this specific problem, kernel estimation may not be a good approach to estimate the function $c(y)$ while B-spline based estimation fits naturally and has a clear advantage. This is not usually the case in many semiparametric estimation problems. Third, we provide a clear and transparent estimation procedure, which is computationally simple and enjoys the advantage of convex optimization. Fourth, we establish the optimality of the resulting marginal effect and marginal quantile effect estimators as functionals of the model parameters. This optimality is not a straightforward result to obtain. It requires keen insight to link two seemingly unrelevant results and the right mathematical tools to achieve it. Putting our problem and results in a broader context, the results are sensible. For example, Shen & Wong (1994) and Zhou et al. (1998) studied the general convergence properties of sieve and more specifically regression spline estimators, and Chen & Liao (2014) further provided inference results for a general functional of the model parameters in sieve estimation. These works are done in a much more general context and agree with our findings.

The rest of the article is organized as the following. In Section 3.2, we provide the main methodological results of our work and describe the estimation procedures for both $\boldsymbol{\xi}$ and $\boldsymbol{\eta}_\tau$. In Section 3.3, we establish the asymptotic properties of our estimators, compare them to the theoretical efficiency bounds, and show that they in fact reach the efficiency bounds. Simulation studies are conducted in Section 3.4 to illustrate the finite sample performance of our method in comparison with existing methods. An interesting dataset concerning Swiss non-labor income is analyzed in Section 3.5, where our nonparametric analysis reveals new insights concerning model choice, marginal effect, and significant covariates that would remain hidden if using a strict parametric approach. We conclude our paper with a discussion in Section 3.6. All the technical details are relegated to the supplementary material.

3.2 Methodology

3.2.1 Efficiency bound of marginal effect estimation

Before engaging ourselves in the estimation and inference of the marginal effect $\boldsymbol{\xi}$ associated with the model (3.4), we would like first to understand the limit of our endeavor by establishing the efficiency bound of the $\boldsymbol{\xi}$ estimation. Let $v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \equiv E[\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}^2 \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}]$. The derivation detailed in Section B.1 leads to the efficient influence function as

$$\boldsymbol{\phi}_{\text{eff}}(y, \mathbf{x}) = \boldsymbol{\beta} v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) - \boldsymbol{\beta} E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} + \boldsymbol{\beta} y^2 + \mathbf{M}\mathbf{x}y + \mathbf{a}(y) - E\{\boldsymbol{\beta} Y^2 + \mathbf{M}\mathbf{x}Y + \mathbf{a}(Y) \mid \mathbf{x}\},$$

where $\mathbf{a}(y)$ satisfies

$$\begin{aligned} & E[E\{\mathbf{a}(Y) \mid \mathbf{X}\} \mid y] - \mathbf{a}(y) \\ = \; & 2\boldsymbol{\beta} E[yE(Y \mid \mathbf{X}) - E\{YE(Y \mid \mathbf{X}) \mid \mathbf{X}\} \mid y] + \mathbf{M}E[\mathbf{X}\{y - E(Y \mid \mathbf{X})\} \mid y]. \end{aligned}$$

and

$$\begin{aligned} \mathbf{M} \;=\; & \left(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\left[2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}} Y v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \mathbf{a}(Y)\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right) \\ & \times \left[E\{\mathbf{X}\mathbf{X}^{\mathrm{T}} v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\right]^{-1}, \end{aligned}$$

Obviously, the variance of the efficient influence function, i.e. $\mathrm{var}\{\boldsymbol{\phi}_{\text{eff}}(Y, \mathbf{X})\}$, is the efficiency bound in estimating $\boldsymbol{\xi}$.

3.2.2 Efficiency bound of marginal quantile effect estimation

In Section B.2, we further derive the efficient influence function of $\boldsymbol{\eta}_\tau$ under (3.4). Now for notational brevity, let $q(\nu) \equiv Q_\tau(Y \mid \nu)$, $q'(\nu) \equiv Q'_\tau(Y \mid \nu)$, $q''(\nu) \equiv Q''_\tau(Y \mid \nu)$, $\epsilon(\nu) \equiv \tau - I\{Y < q(\nu)\}$, and $\epsilon'(\nu) \equiv -\delta\{q(\nu) - Y\}q'(\nu)$. The efficient influence function turns out to be

$$\boldsymbol{\phi}_{\text{eff}}(y, \mathbf{x}) = \boldsymbol{\beta} q'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) - \boldsymbol{\beta} E\{q'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} + \mathbf{M}_1\mathbf{x}y + \mathbf{a}(y) - E\{\mathbf{M}_1\mathbf{x}Y + \mathbf{a}(Y)|\mathbf{x}\},$$

where $\mathbf{a}(y)$ satisfies

$$E[E\{\mathbf{a}(Y) \mid \mathbf{X}\} \mid y] - \mathbf{a}(y) = -\boldsymbol{\beta} E\{r(y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})|y\} + \mathbf{M}_1 E[\mathbf{X}\{y - E(Y|\mathbf{X})\}|y],$$

in which

$$r(Y,\nu) \equiv \frac{\epsilon(\nu)Y + \epsilon'(\nu) - \epsilon(\nu)[q(\nu) + q'(\nu)\nu + q'(\nu)c'\{q(\nu)\}]}{f\{q(\nu),\nu\}},$$

and

$$\begin{aligned}\mathbf{M}_1 &= (E\{q'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} + \boldsymbol{\beta}E\{\mathbf{X}^{\mathrm{T}}q''(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} - E[\mathbf{a}(Y)\mathbf{X}^{\mathrm{T}}\{Y - E(Y|\mathbf{X})\}]) \\ &\quad \times[E\{\mathbf{X}\mathbf{X}^{\mathrm{T}}v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}]^{-1}.\end{aligned}$$

Similar to the case for estimating $\boldsymbol{\xi}$, the variance of the efficient influence function is the efficiency bound for estimating $\boldsymbol{\eta}_\tau$.

3.2.3 Estimation procedure

We now propose an estimation procedure under the model (3.4). We first consider the case where the response Y is distributed continuously. In such case, we approximate $c(\cdot)$ of the conditional density $f_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta},c)$ by a B-spline curve $\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}$, where $\mathbf{B}(\cdot) \equiv \{B_1(\cdot),\ldots,B_m(\cdot)\}^{\mathrm{T}}$ is a B-spline basis vector and $\boldsymbol{\gamma} \equiv (\gamma_1,\ldots,\gamma_m)^{\mathrm{T}}$ is an unknown coefficient of the bases. Since we assume $c(0) = 0$ for identifiability, we ignore the first B-spline basis $B_0(\cdot)$ which corresponds to the intercept of the curve. Therefore, we replace $c(y)$ in (3.4) by $\mathbf{B}(y)^{\mathrm{T}}\boldsymbol{\gamma}$ and estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ through maximizing the approximate loglikelihood

$$l(\boldsymbol{\beta},\boldsymbol{\gamma}) \equiv \sum_{i=1}^{n}\left[y_i\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \mathbf{B}(y_i)^{\mathrm{T}}\boldsymbol{\gamma} - \log\int \exp\left\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \mathbf{B}(y)^{\mathrm{T}}\boldsymbol{\gamma}\right\} d\mu(y)\right]. \tag{3.5}$$

Note that the loglikelihood $l(\boldsymbol{\beta},\boldsymbol{\gamma})$ is a concave function, hence the optimizer is unique and can be readily obtained using off-the-shelf convex optimizers. In our implementation, we used the built-in R function **optim** for optimization, yet found the computation fast with satisfactory performance. For example, in our real data analysis where the dataset contains 871 observations and 7 predictors, it takes less than 10 seconds to obtain the optimizer within the relative tolerance of 10^{-6}.

Once we obtain $\widehat{\boldsymbol{\beta}}$ and $\widehat{c}(\cdot) \equiv \mathbf{B}(\cdot)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}$, we can easily estimate the marginal effect $\boldsymbol{\xi}$ using (3.2) through

$$\widehat{\boldsymbol{\xi}} \equiv \widehat{\boldsymbol{\beta}}\widehat{E}\left\{\widehat{\mathrm{var}}(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right\}$$

$$= \widehat{\boldsymbol{\beta}}\, n^{-1} \sum_{i=1}^{n} \left[\int y^2 f_{Y|\mathbf{X}}(y, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{c}) d\mu(y) - \left\{ \int y f_{Y|\mathbf{X}}(y, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{c}) d\mu(y) \right\}^2 \right].$$

Similarly, we estimate the marginal quantile effect $\boldsymbol{\eta}_\tau$ using (3.3) through

$$\begin{aligned} \widehat{\boldsymbol{\eta}}_\tau &\equiv \widehat{\boldsymbol{\beta}} \widehat{E} \left\{ \widehat{Q}'_\tau(Y \mid \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) \right\} \\ &= \widehat{\boldsymbol{\beta}}\, n^{-1} \sum_{i=1}^{n} \left. \frac{\tau \int_0^1 y f_{Y|\mathbf{X}}(y, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{c}) d\mu(y) - \int_0^{q_i} y f_{Y|\mathbf{X}}(y, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{c}) d\mu(y)}{f_{Y|\mathbf{X}}(q_i, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{c})} \right|_{q_i = \widehat{Q}_\tau(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)}, \end{aligned}$$

where $\widehat{Q}_\tau(Y \mid \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i), i = 1, \ldots, n$, the estimated conditional τth quantiles of Y given $\mathbf{x}_i$, are obtained by solving for q_i from

$$\int_0^{q_i} f_{Y|\mathbf{X}}(y, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{c}) d\mu(y) = \tau.$$

When the response Y is a categorical variable taking values $\{0, ..., M\}$, then the goal is to estimate $\boldsymbol{\beta}$ and $c(0), \ldots, c(M)$. In this case, (3.4) is a purely parametric model and we can proceed with maximum likelihood estimation (MLE). An alternative way of viewing this is that we replace $c(\cdot)$ by $\mathbf{D}(\cdot)^{\mathrm{T}} \boldsymbol{\gamma}$ where $\mathbf{D}(\cdot) \equiv \{I(\cdot = 0), \ldots, I(\cdot = M)\}^{\mathrm{T}}$ with $\gamma_1 = 0$, and maximize the approximate loglikelihood given in (3.5) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Having obtained $\widehat{\boldsymbol{\beta}}$ and $\widehat{c}(\cdot)$, we use

$$\mathrm{pr}(Y = y | \mathbf{x}; \boldsymbol{\beta}, c) = \frac{\exp\{y \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x} + c(y)\}}{\sum_{y=0}^{M} \exp\{y \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x} + c(y)\}},$$

$y = 0, \ldots, M$ to estimate the marginal effect $\boldsymbol{\xi}$ by

$$\widehat{\boldsymbol{\xi}} \equiv \widehat{\boldsymbol{\beta}}\, n^{-1} \sum_{i=1}^{n} \left[\sum_{y=0}^{M} y^2 \,\mathrm{pr}(Y = y | \mathbf{x}_i; \widehat{\boldsymbol{\beta}}, \widehat{c}) - \left\{ \sum_{y=0}^{M} y \,\mathrm{pr}(Y = y | \mathbf{x}_i; \widehat{\boldsymbol{\beta}}, \widehat{c}) \right\}^2 \right].$$

Because there is no generally accepted unique way of defining quantiles for discrete data, we do not further study the marginal quantile effect estimation in this case.

3.3 Theoretical properties

We now establish the theoretical properties of our proposed estimators for both the marginal effect and the marginal quantile effect in (3.4).

3.3.1 Continuous response

First we analyze the properties of our estimators under the continuous response case. We simplify $d\mu(y)$ as dy below. To set notation, let

$$\begin{aligned}
f^*_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma}) &\equiv \frac{\exp\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}+\mathbf{B}(y)^{\mathrm{T}}\boldsymbol{\gamma}\}}{\int\exp\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}+\mathbf{B}(y)^{\mathrm{T}}\boldsymbol{\gamma}\}dy},\\
E^*\left\{\mathbf{g}(Y)|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma}\right\} &\equiv \int\mathbf{g}(y)f^*_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma})dy,\\
\mathrm{var}^*\left(Y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma}\right) &\equiv E^*(Y^2|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma})-\left\{E^*(Y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma})\right\}^2,\\
\mathrm{cov}^*\left\{\mathbf{g}(Y),\mathbf{h}(Y)|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma}\right\} &\equiv E^*\left\{\mathbf{g}(Y)\mathbf{h}(Y)^{\mathrm{T}}|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma}\right\}\\
&\quad -E^*\left\{\mathbf{g}(Y)|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma}\right\}E^*\left\{\mathbf{h}(Y)^{\mathrm{T}}|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma}\right\},
\end{aligned}$$

where we use E^* instead of E to emphasize that the expectation is computed under the approximate model, which has the same form as (3.4), but with $c(\cdot)$ replaced by $\mathbf{B}^{\mathrm{T}}(\cdot)\boldsymbol{\gamma}$.

For real numbers a_n and b_n, $a_n \asymp b_n$ denotes $a_n = O(b_n)$ and $b_n = O(a_n)$ simultaneously. Similarly, for random variables A_n and B_n, $A_n \asymp_p B_n$ denotes $A_n = O_p(B_n)$ and $B_n = O_p(A_n)$ simultaneously. For a vector $\mathbf{a} = (a_1,\ldots,a_d)^{\mathrm{T}} \in \mathbb{R}^d$, we denote the l_p-norm of $\mathbf{a}$ as $\|\mathbf{a}\|_p \equiv (|a_1|^p + \cdots + |a_d|^p)^{1/p}$, $1 \leqslant p \leqslant \infty$. For a matrix $\mathbf{A} \in \mathbb{R}^{r\times c}$, we denote the induced l_p-norm of $\mathbf{A}$ as $\|\mathbf{A}\|_p \equiv \sup_{\mathbf{u}\in\mathbb{R}^c,\|\mathbf{u}\|_p=1}\|\mathbf{A}\mathbf{u}\|_p$, $1 \leqslant p \leqslant \infty$. For a function $g(\cdot)$ in the L^2 space, we denote its L_p-norm as $\|g(\cdot)\|_p \equiv \{\int_0^1 |g(y)|^p dy\}^{1/p}$. We denote the set of the qth order smooth functions as $C^q([0,1]) \equiv \{g : g^{(q)} \in C([0,1])\}$.

To facilitate the theoretical derivation, we view the estimation procedure described in Section 3.2.3 alternatively as a profile procedure. Specifically, treating $\boldsymbol{\beta}$ as a fixed parameter, estimate $c(\cdot)$ by the spline curve $\widehat{c}(\cdot,\boldsymbol{\beta}) \equiv \mathbf{B}(\cdot)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta})$ via maximizing the approximate loglikelihood

$$l(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i=1}^n\left(y_i\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \mathbf{B}(y_i)^{\mathrm{T}}\boldsymbol{\gamma} - \log\left[\int_0^1\exp\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \mathbf{B}(y)^{\mathrm{T}}\boldsymbol{\gamma}\}dy\right]\right)$$

with respect to $\boldsymbol{\gamma}$. Then estimate $\boldsymbol{\beta}$ by maximizing

$$\widehat{l}(\boldsymbol{\beta}) \equiv \sum_{i=1}^n\left(y_i\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \widehat{c}(y_i,\boldsymbol{\beta}) - \log\left[\int_0^1\exp\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \widehat{c}(y,\boldsymbol{\beta})\}dy\right]\right).$$

We point out that profiling and performing maximization jointly with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ yield the same result.

In the following, we first aim at proving the convergence property of $\widehat{c}(y,\boldsymbol{\beta})$ when

$\boldsymbol{\beta} = \boldsymbol{\beta}_0$, where we let $\boldsymbol{\beta}_0$ denote the regression coefficient of the true conditional density.

We assume the following regularity conditions.

(C1) The function $c(\cdot) \in C^q([0,1])$ where $q \geqslant 2$ and $c(0) = 0$. The true conditional density of Y given $\mathbf{X}$, $f_{Y|\mathbf{X}}(y|\mathbf{x})$, has a compact support $[0,1]$, is positive and bounded on its support. The marginal density of $\mathbf{X}$, $f_{\mathbf{X}}(\mathbf{x})$, has compact support $\mathcal{X}$ and is bounded on its support.

(C2) The spline order $r \geqslant q$.

(C3) Define the knots $t_{-r+1} = \cdots = t_0 = 0 < t_1 < \cdots < t_N < 1 = t_{N+1} = \cdots = t_{N+r}$ where N is the number of interior knots and [0,1] is divided into $(N+1)$ subintervals. Let $m = N + r - 1$. N satisfies $N \to \infty$, $N^{-2q+1}n \to 0$, and $N^{-2}n(\log n)^{-1} \to \infty$ as $n \to \infty$.

(C4) Let h_p be the distance between the pth and $(p+1)$th knots. There exists a constant C_h such that $0 < C_h < \infty$ and $\max_{r \leqslant p \leqslant N+r} h_p / \min_{r \leqslant p \leqslant N+r} h_p < C_h$. Therefore, $\max_{r \leqslant p \leqslant N+r} h_p \asymp N^{-1}$ and $\min_{r \leqslant p \leqslant N+r} h_p \asymp N^{-1}$.

(C5) (De Boor 1978) Under Conditions (C1)-(C4), there exists a spline coefficient $\boldsymbol{\gamma}_0$ such that

$$\sup_{y \in [0,1]} \left| \mathbf{B}(y)^{\mathrm{T}} \boldsymbol{\gamma}_0 - c(y) \right| = O(N^{-q}).$$

Condition (C1) imposes the smoothness of the functional component $c(\cdot)$, and the boundedness of the densities involved in the model, which are standard requirements. The compact support requirement of the densities is also a standard requirement in the B-spline literature. Condition (C2) requires that the order of the B-spline basis is sufficiently large for the B-spline curve to converge to the true function fast enough. We further assume that there are an appropriate number of interior knots by Condition (C3), and that the knots are uniformly distributed in the asymptotic sense by Condition (C4). Lastly, we point out that Condition (C5) does not further impose any additional requirement. It is a direct result given Conditions (C1)-(C4), and is only stated as a condition for convenience.

Proposition 3.3.1. *Under Conditions (C1)-(C5),* $\|\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0\|_2 = O_p(n^{-1/2}N)$ *and*

$$\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0 \;=\; \boldsymbol{\Sigma}_{22}^{-} n^{-1} \sum_{i=1}^{n} [\mathbf{B}(y_i) - E\{\mathbf{B}(Y)|\mathbf{x}_i\}] + \mathbf{r}_1,$$

where $\|\mathbf{r}_1\|_2 = o_p(n^{-1/2}N)$. *Furthermore,*

$$\sup_{y\in[0,1]} |\widehat{c}(y,\boldsymbol{\beta}_0) - c(y)| = O_p\left\{n^{-1/2}(N\log N)^{1/2}\right\}.$$

Proposition 3.3.1 states that given the true regression coefficient $\boldsymbol{\beta}_0$, $\widehat{\boldsymbol{\gamma}}$ converges to the B-spline basis coefficient $\boldsymbol{\gamma}_0$ at the nonparametric convergence rate $n^{-1/2}N$, hence the B-spline curve approximates $c(\cdot)$. We further obtain that the estimator $\widehat{c}(\cdot,\boldsymbol{\beta}_0)$ converges to the true function $c(\cdot)$ at the $O_p\{n^{-1/2}(N\log N)^{1/2} + N^{-q}\}$ rate. Based on Proposition 3.3.1, below we further establish the asymptotic property of the estimator $\widehat{\boldsymbol{\beta}}$. Note $\boldsymbol{\beta}$ is not our research interest, hence this property is stated as a by-product. We first impose one additional regularity condition, which is a standard requirement.

(C6) The expectation of the conditional covariance of $\{\mathbf{X}^{\mathrm{T}}Y, \mathbf{B}^{\mathrm{T}}(Y)\}^{\mathrm{T}}$ given $\mathbf{X}$, i.e.

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \equiv E\begin{bmatrix} \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathrm{var}(Y|\mathbf{X}) & \mathbf{X}\mathrm{cov}\{Y,\mathbf{B}(Y)|\mathbf{X}\} \\ \mathrm{cov}\{\mathbf{B}(Y),Y|\mathbf{X}\}\mathbf{X}^{\mathrm{T}} & \mathrm{var}\{\mathbf{B}(Y)|\mathbf{X}\} \end{bmatrix},$$

is invertible.

It is easy to see that the conditional covariance of $\{\mathbf{X}^{\mathrm{T}}Y, \mathbf{B}^{\mathrm{T}}(Y)\}^{\mathrm{T}}$ given $\mathbf{X}$ is positive semidefinite. Thus, $\boldsymbol{\Sigma}$ is also positive semidefinite. Condition (C6) further requires that $\boldsymbol{\Sigma}$ is positive definite, which is very mild. Note that this guarantees $\boldsymbol{\Sigma}_{11}$ and the Schur complement of $\boldsymbol{\Sigma}_{22}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \equiv \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}\boldsymbol{\Sigma}_{21}$, are both positive definite.

Proposition 3.3.2. *Under Conditions (C1)-(C6),* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ *and*

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}} n^{-1}\sum_{i=1}^{n} \mathbf{x}_i\{y_i - E(Y|\mathbf{x}_i)\} \\ &\quad -\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-} n^{-1}\sum_{i=1}^{n}[\mathbf{B}(y_i) - E\{\mathbf{B}(Y)|\mathbf{x}_i\}] + \mathbf{r}_2, \end{aligned}$$

where $\|\mathbf{r}_2\|_2 = o_p(n^{-1/2})$. *Furthermore,*

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}_p, \mathbf{I}_p)$$

in distribution as $n \to \infty$.

Proposition 3.3.2 establishes how the regression coefficient estimator $\widehat{\boldsymbol{\beta}}$ is asymptotically distributed, and allows to perform inference since the asymptotic variance of $\widehat{\boldsymbol{\beta}}$

is estimable based on $\widehat{\boldsymbol{\beta}}$ and $f_{Y|\mathbf{X}}(y, \mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{c})$. Below, we study how the estimator of the marginal effect $\boldsymbol{\xi}$ is asymptotically distributed in Theorem 3.3.1. We point out that the asymptotic variance of $\widehat{\boldsymbol{\xi}}$ is estimable as well so that we can perform inference on the marginal effect.

Theorem 3.3.1. *Let* $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} \equiv \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}} var\{var(Y|\mathbf{X})\}$, *where* $\mathbf{A} \equiv [\mathbf{A}_1, \mathbf{A}_2]$ *and*

$$\begin{aligned}\mathbf{A}_1 &\equiv E\{var(Y|\mathbf{X})\}\mathbf{I} + \boldsymbol{\beta}_0 E[\{Y - E(Y|\mathbf{X})\}^3\mathbf{X}^{\mathrm{T}}],\\ \mathbf{A}_2 &\equiv \boldsymbol{\beta}_0 E(\{Y - E(Y|\mathbf{X})\}^2[\mathbf{B}(Y) - E\{\mathbf{B}(Y)|\mathbf{X}\}])^{\mathrm{T}}.\end{aligned}$$

Under Conditions (C1)-(C6),

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0) \to N(\mathbf{0}_p, \mathbf{I}_p)$$

in distribution as $n \to \infty$.

To facilitate the analysis of the properties of the marginal quantile effect estimator $\widehat{\boldsymbol{\eta}}_\tau$, we further discuss the conditions stated above. Note that for an arbitrary twice differentiable function $g(\cdot)$, $\|g'(\cdot)\|_\infty \leqslant 2(\|g(\cdot)\|_\infty\|g''(\cdot)\|_\infty)^{1/2}$ by the Landau-Kolmogorov inequality. Setting $g(\cdot) = \mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}_0 - c(\cdot)$, we have $\|g(\cdot)\|_\infty = O(N^{-q})$ by Condition (C5) and $\|g''(\cdot)\|_\infty$ is bounded since $g''(\cdot)$ is continuous on a compact support by Conditions (C1) and (C2). Therefore, $\|\mathbf{B}'(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}_0 - c'(\cdot)\|_\infty = O(N^{-q/2})$, i.e., $\mathbf{B}'(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}_0$ converges to $c'(\cdot)$ uniformly at the rate $O(N^{-q/2})$. Theorem 3.3.2 below provides the asymptotic normality of the marginal quantile effect estimator $\widehat{\boldsymbol{\eta}}_\tau$.

Theorem 3.3.2. *Let* $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_\tau} \equiv \mathbf{C}\boldsymbol{\Sigma}^{-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}} var\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}$, *where* $\mathbf{C} \equiv [\mathbf{C}_1, \mathbf{C}_2]$ *and*

$$\begin{aligned}\mathbf{C}_1 &\equiv E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\mathbf{I} + \boldsymbol{\beta}_0 E\Bigg[\mathbf{X}^{\mathrm{T}}\Bigg\{\frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y^2|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}}\\ &\quad -2q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) - \{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2[\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + c'\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\}],\\ \mathbf{C}_2 &\equiv \boldsymbol{\beta}_0 E\Bigg\{\frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y\mathbf{B}^{\mathrm{T}}(Y)|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}}\\ &\quad -q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}^{\mathrm{T}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} - \frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}^{\mathrm{T}}(Y)|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}}\\ &\quad \times[q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\}.\end{aligned}$$

Under Conditions (C1)-(C6),

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}_\tau}^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\eta}}_\tau - \boldsymbol{\eta}_{\tau 0}) \to N(\mathbf{0}_p, \mathbf{I}_p)$$

in distribution as $n \to \infty$.

To utilize Proposition 3.3.2 and Theorems 3.3.1 and 3.3.2 for inference, we can estimate $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_\tau}$ respectively by $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\Sigma}}_{11} - \widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-}\widehat{\boldsymbol{\Sigma}}_{21})^{-1}$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} \equiv \widehat{\mathbf{A}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{A}}^{\mathrm{T}} + \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\mathrm{var}}\{\widehat{\mathrm{var}}(Y|\mathbf{X})\}$, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_\tau} \equiv \widehat{\mathbf{C}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}}^{\mathrm{T}} + \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\mathrm{var}}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}$, where

$$\widehat{\boldsymbol{\Sigma}} \equiv \{(\widehat{\boldsymbol{\Sigma}}_{11}, \widehat{\boldsymbol{\Sigma}}_{12})^{\mathrm{T}}, (\widehat{\boldsymbol{\Sigma}}_{21}, \widehat{\boldsymbol{\Sigma}}_{22})^{\mathrm{T}}\}^{\mathrm{T}},$$

$\widehat{\mathbf{A}} \equiv [\widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2]$, $\widehat{\mathbf{C}} \equiv [\widehat{\mathbf{C}}_1, \widehat{\mathbf{C}}_2]$, and

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{11} &\equiv \widehat{E}\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\widehat{\mathrm{var}}(Y|\mathbf{X})\},\\
\widehat{\boldsymbol{\Sigma}}_{12} &\equiv \widehat{E}[\mathbf{X}\widehat{\mathrm{cov}}\{Y, \mathbf{B}(Y)|\mathbf{X}\}],\\
\widehat{\boldsymbol{\Sigma}}_{21} &\equiv \widehat{E}[\widehat{\mathrm{cov}}\{\mathbf{B}(Y), Y|\mathbf{X}\}\mathbf{X}^{\mathrm{T}}],\\
\widehat{\boldsymbol{\Sigma}}_{22} &\equiv \widehat{E}[\widehat{\mathrm{var}}\{\mathbf{B}(Y)|\mathbf{X}\}],\\
\widehat{\mathbf{A}}_1 &\equiv \widehat{E}\{\widehat{\mathrm{var}}(Y|\mathbf{X})\}\mathbf{I} + \widehat{\boldsymbol{\beta}}\widehat{E}[\{Y - \widehat{E}(Y|\mathbf{X})\}^3\mathbf{X}^{\mathrm{T}}],\\
\widehat{\mathbf{A}}_2 &\equiv \widehat{\boldsymbol{\beta}}\widehat{E}(\{Y - \widehat{E}(Y|\mathbf{X})\}^2[\mathbf{B}(Y) - \widehat{E}\{\mathbf{B}(Y)|\mathbf{X}\}])^{\mathrm{T}},\\
\widehat{\mathbf{C}}_1 &\equiv \widehat{E}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}\mathbf{I} + \widehat{\boldsymbol{\beta}}\widehat{E}\Bigg[\mathbf{X}^{\mathrm{T}}\Bigg\{\frac{\widehat{E}([\tau - I\{Y \leqslant \widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}]Y^2|\mathbf{X})}{\widehat{f}_{Y|\mathbf{X}}\{\widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})|\mathbf{X}\}}\\
&\quad -2\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}^2[\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{c}'\{\widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}]\}],\\
\widehat{\mathbf{C}}_2 &\equiv \widehat{\boldsymbol{\beta}}\widehat{E}\Bigg\{\frac{\widehat{E}([\tau - I\{Y \leqslant \widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}]Y\mathbf{B}^{\mathrm{T}}(Y)|\mathbf{X})}{\widehat{f}_{Y|\mathbf{X}}\{\widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})|\mathbf{X}\}}\\
&\quad -\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\mathbf{B}^{\mathrm{T}}\{\widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\} - \frac{\widehat{E}([\tau - I\{Y \leqslant \widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}]\mathbf{B}^{\mathrm{T}}(Y)|\mathbf{X})}{\widehat{f}_{Y|\mathbf{X}}\{\widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})|\mathbf{X}\}}\\
&\quad \times[\widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) + \widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\widehat{c}'\{\widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}]\}.
\end{aligned}$$

Here, $\widehat{E}(\cdot \mid \mathbf{x})$, $\widehat{\mathrm{cov}}(\cdot \mid \mathbf{x})$, $\widehat{\mathrm{var}}(\cdot \mid \mathbf{x})$, $\widehat{f}_{Y|\mathbf{X}}(\cdot \mid \mathbf{x})$, $\widehat{q}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$, and $\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$ are computed under the estimated model of (3.4) with $\widehat{\boldsymbol{\beta}}$, $\widehat{c}(\cdot) = \mathbf{B}(y)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}$, and $\widehat{c}'(\cdot) = \mathbf{B}'(\cdot)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}$. Also, the approximate marginal expectation and variance, $\widehat{E}(\cdot)$ and $\widehat{\mathrm{var}}(\cdot)$, are computed via the corresponding sample moments over the n observations. Below, we show that the estimators of the variances are consistent.

Theorem 3.3.3. *Under Conditions (C1)-(C6),* $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\|_2 \to 0$ *and* $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}}\|_2 \to 0$ *in*

probability when $n \to \infty$. *In addition, under Conditions (C1)-(C6),* $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_\tau} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}_\tau}\|_2 \to 0$ *in probability when* $n \to \infty$.

The asymptotic properties we established, especially the estimation variances of $\widehat{\boldsymbol{\xi}}$ and $\widehat{\boldsymbol{\eta}}_\tau$, have very different forms from the efficiency bounds we derived in Sections 3.2.1 and 3.2.2. Nevertheless, closer inspection, together with some basic but less frequently adopted linear algebra tools reveal that these two sets of results have much closer connections, and the efficiency bounds in estimating both $\boldsymbol{\xi}$ and $\boldsymbol{\eta}_\tau$ are actually reached by our B-spline based approximate maximum likelihood estimators. As a by-product, we also state the efficiency property of $\widehat{\boldsymbol{\beta}}$ as a proposition, even though our interest is not in $\boldsymbol{\beta}$.

Proposition 3.3.3. *Under Conditions (C1)-(C6), the approximate maximum likelihood estimator* $\widehat{\boldsymbol{\beta}}$ *is efficient.*

Theorem 3.3.4. *Under Conditions (C1)-(C6), the estimator* $\widehat{\boldsymbol{\xi}}$ *based on the approximate maximum likelihood estimator* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{c}(\cdot)$ *is efficient.*

Theorem 3.3.5. *Under Conditions (C1)-(C6), the estimator* $\widehat{\boldsymbol{\eta}}_\tau$ *based on the approximate maximum likelihood estimator* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{c}(\cdot)$ *is efficient.*

3.3.2 Discrete response

We now analyze the properties of our estimators under the discrete response case. For notational simplicity, we denote $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}) \equiv \{\mathrm{pr}(Y = 1 \mid \mathbf{x}; \boldsymbol{\beta}, \mathbf{c}), \ldots, \mathrm{pr}(Y = M \mid \mathbf{x}; \boldsymbol{\beta}, \mathbf{c})\}^{\mathrm{T}}$, and $\mathbf{p}_k(\mathbf{x}, \boldsymbol{\theta}) \equiv \{1^k \mathrm{pr}(Y = 1 \mid \mathbf{x}; \boldsymbol{\beta}, \mathbf{c}), \ldots, M^k \mathrm{pr}(Y = M \mid \mathbf{x}; \boldsymbol{\beta}, \mathbf{c})\}^{\mathrm{T}}$. Also in this section, we use $\mathbf{D}(y)$ to denote a vector of indicator functions, i.e., $\mathbf{D}(y) \equiv \{I(y = 1), \ldots, I(y = M)\}^{\mathrm{T}}$. Note that we allow M to grow with the sample size n. Here we present theoretical results when M grows to infinity. Note that in the finite M case, the analysis can be done easily through incorporating the classical maximum likelihood approach. We first list a set of regularity conditions.

(D1) The true conditional mass function of Y given $\mathbf{X}$, $\mathrm{pr}(Y = y \mid \mathbf{x}; \boldsymbol{\theta}_0)$, has a support set $\{0, \ldots, M\}$. $E\{\mathrm{pr}(Y = 0 \mid \mathbf{X}; \boldsymbol{\theta}_0)\} \leqslant 1 - \delta$ for some constant $0 < \delta < 1$ and $E(Y^4 \mid \mathbf{x})$ is bounded. The marginal density of $\mathbf{X}$, $f_{\mathbf{X}}(\mathbf{x})$, has compact support $\mathcal{X}$ and is bounded on its support.

(D2) There exist constants L_k such that $|E(Y^k \mid \mathbf{x}, \boldsymbol{\theta}^*) - E(Y^k \mid \mathbf{x}, \boldsymbol{\theta}_0)| \leqslant L_k \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2$ for $k = 1, 2, 3$.

(D3) $M \to \infty$, $n^{-1}M^3 \to 0$, and $E\{\|\mathbf{p}(\mathbf{X}, \boldsymbol{\theta}_0)\|_2^2\} \to 0$ as $n \to \infty$.

(D4) $\boldsymbol{\Sigma}_{22} \equiv E[\mathrm{var}\{\mathbf{D}(Y) \mid \mathbf{X}\}]$ is invertible. $\|\boldsymbol{\Sigma}_{22}\|_2^{-1}\boldsymbol{\Sigma}_{22}$ has all eigenvalues bounded above a constant $C_l > 0$.

(D5) The expectation of the conditional covariance of $\{\mathbf{X}^{\mathrm{T}}Y, \mathbf{D}^{\mathrm{T}}(Y)\}^{\mathrm{T}}$ given $\mathbf{X}$, i.e.

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \equiv E \begin{bmatrix} \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathrm{var}(Y \mid \mathbf{X}) & \mathbf{X}\mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{X}\} \\ \mathrm{cov}\{\mathbf{D}(Y), Y \mid \mathbf{X}\}\mathbf{X}^{\mathrm{T}} & \mathrm{var}\{\mathbf{D}(Y) \mid \mathbf{X}\} \end{bmatrix},$$

is invertible.

Conditions (D1) and (D2) require boundedness and Lipschitz continuity on the conditional moments of Y, which are standard requirements. Condition (D3) requires M to tend to infinity at the rate slower than $n^{1/3}$, and the mass function not to concentrate on a finite subset of the support. Further, since $\boldsymbol{\Sigma}_{22}$ and $\boldsymbol{\Sigma}$ are positive semidefinite by their definitions, the invertibility imposed by Conditions (D4) and (D5) is very mild. Lastly, we point out some results on $\boldsymbol{\Sigma}_{22}$ in Remark 3.3.1 below.

Remark 3.3.1. *Note that the sum of the eigenvalues of $\boldsymbol{\Sigma}_{22}$ is of constant order by Conditions (D1) and (D3), because trace$(\boldsymbol{\Sigma}_{22}) = E\{1 - pr(Y = 0 \mid \mathbf{X}; \boldsymbol{\theta}_0) - \|\mathbf{p}(\mathbf{X}, \theta_0)\|_2^2\}$. Thus we get $\|\boldsymbol{\Sigma}_{22}\|_2 \asymp M^{-1}$ and $\|\boldsymbol{\Sigma}_{22}^{-1}\|_2 \asymp M$ because the eigenvalues are of the same order by Condition (D4).*

We now state the convergence rate and the asymptotic properties of the estimators of the model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Our analysis shows that the estimator of the regression coefficient $\widehat{\boldsymbol{\beta}}$ achieves the parametric convergence rate under the regularity conditions stated above. Further, we establish the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ as a by-product, by which one can perform inference on the regression coefficient. We formally state the results in Proposition 3.3.4 below.

Proposition 3.3.4. *Under Conditions (D1)-(D5), $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$, $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 = O_p(n^{-1/2}M^{1/2})$, and*

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \end{bmatrix} = \boldsymbol{\Sigma}^{-1}n^{-1}\sum_{i=1}^{n} \begin{bmatrix} \mathbf{x}_i\{y_i - E(Y \mid \mathbf{x}_i)\} \\ \mathbf{D}(y_i) - E\{\mathbf{D}(Y) \mid \mathbf{x}_i\} \end{bmatrix} + \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix},$$

where $\|\mathbf{r}_1\|_2 = o_p(n^{-1/2})$ and $\|\mathbf{r}_2\|_2 = o_p(n^{-1/2}M^{1/2})$. Furthermore, let $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} \equiv (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1}$, then

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}_p, \mathbf{I}_p)$$

in distribution as $n \to \infty$.

Based on Proposition 3.3.4, we further establish a theoretical result on the marginal effect estimator, which states the convergence rate and the asymptotic distribution of $\widehat{\boldsymbol{\xi}}$. Theorem 3.3.6 shows that $\widehat{\boldsymbol{\xi}}$, the functional of both the parametric and the nonparametric components, achieves the parametric convergence rate. We also provide the closed form of the asymptotic variance of $\widehat{\boldsymbol{\xi}}$, which can be used to infer the marginal effect of a population.

Theorem 3.3.6. *Let* $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} \equiv \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}} var\{var(Y \mid \mathbf{X})\}$, *where* $\mathbf{A} \equiv [\mathbf{A}_1, \mathbf{A}_2]$ *and*

$$\begin{aligned} \mathbf{A}_1 &\equiv E\{var(Y \mid \mathbf{X})\}\mathbf{I} + \boldsymbol{\beta}_0 E[\{Y - E(Y \mid \mathbf{X})\}^3 \mathbf{X}^{\mathrm{T}}], \\ \mathbf{A}_2 &\equiv \boldsymbol{\beta}_0 E(\{Y - E(Y \mid \mathbf{X})\}^2 [\mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\}])^{\mathrm{T}}. \end{aligned}$$

Under Conditions (D1)-(D5),

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0) \to N(\mathbf{0}_p, \mathbf{I}_p)$$

in distribution as $n \to \infty$.

The estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ have exactly the same expressions as those of $\boldsymbol{\Sigma}^{*-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ defined in Section 3.3.1, while $\mathbf{B}(\cdot)$ replaced by $\mathbf{D}(\cdot)$. In Theorem 3.3.7, we further establish the consistency of the variance estimators.

Theorem 3.3.7. *Under Conditions (D1)-(D5),* $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\|_2 \to 0$ *and* $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}}\|_2 \to 0$ *in probability when* $n \to \infty$.

We now show that our proposed estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\xi}}$ for the discrete case also achieve the efficiency bounds. Although the asymptotic variances of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\xi}}$ established in Proposition 3.3.4 and in Theorem 3.3.6 appear very different from the efficiency bounds derived in Appendix B.10 and Section B.1, our analysis shows that these two seemingly different variance structures are actually identical. Below, we formally state the efficiency of $\widehat{\boldsymbol{\xi}}$ as Theorem 3.3.8, and that of $\widehat{\boldsymbol{\beta}}$ as Proposition 3.3.5.

Proposition 3.3.5. *Under Conditions (D1)-(D5), the approximate maximum likelihood estimator* $\widehat{\boldsymbol{\beta}}$ *is efficient.*

Theorem 3.3.8. *Under Conditions (D1)-(D5), the estimator* $\widehat{\boldsymbol{\xi}}$ *based on the approximate maximum likelihood estimator* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{c}(\cdot)$ *is efficient.*

3.4 Simulation experiments

We conduct simulation studies to investigate the finite sample performance of the proposed methods. We consider the case where the response follows a GLM or a truncated GLM. All results are based on 1000 replicates with sample size $n = 1000$. For comparison, we implemented three different estimators, our proposed approximate maximum likelihood estimator (aMLE), the pairwise marginal likelihood estimator (pMLE) by Lin et al. (2021), and the maximum likelihood estimator (MLE) under the non-truncated GLM. Note that in our simulation scenarios, GLM is a true model when the response Y is generated without truncation, hence MLE provides the most efficient estimator among all the implemented methods. But GLM is a misspecified model when the response Y is generated from a truncated GLM, hence MLE leads to biased estimates in this case.

For our proposed aMLE for the continuous response, we used the cubic B-spline basis with the number of interior knots equal to the smallest integer larger than $0.5n^{1/4}$, i.e., $N = \lceil 0.5n^{1/4} \rceil$, where the knots are the quantiles of $\{y_i : i = 1, \ldots, n\}$ of length $(N + 2)$ whose levels are evenly spaced in $[0, 1]$. On the other hand, because Lin et al. (2021) only studied the estimation of $\boldsymbol{\beta}$ and $c(\cdot)$, we implemented the estimators $\widehat{\boldsymbol{\xi}}$ and $\widehat{\boldsymbol{\eta}}_\tau$ of pMLE in the same manner as for aMLE, based on the pMLE estimated $\boldsymbol{\beta}$ and $c(\cdot)$.

We report the average of the absolute bias, the sample standard error σ_{sim}, the average of the asymptotic standard error $\widehat{\sigma}_{\text{est}}$, and the empirical coverage of the estimated confidence interval at 95% confidence level (CI). The $\widehat{\sigma}_{\text{est}}$ and CI of pMLE are omitted because Lin et al. (2021) did not provide them.

3.4.1 Normal distribution

We first examine the normal regression model. A three dimensional covariate vector $\mathbf{X}_i$ was independently drawn from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\Sigma = (\sigma_{kl})$ where $\sigma_{kl} = 0.1^{|k-l|}, k = 1, \ldots, 3, l = 1, \ldots, 3$. Then for the truncated case, we further generated a response Y_i independently from a truncated normal distribution on $[a, b]$, which has the density

$$f_Y(y; \theta_i, \sigma, a, b) = \frac{\frac{1}{\sigma}\phi(\frac{y-\theta_i}{\sigma})}{\Phi(\frac{b-\theta_i}{\sigma}) - \Phi(\frac{a-\theta_i}{\sigma})},$$

where $\theta_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \sigma = 1, a = -5, b = 5$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (pdf) and the cumulative distribution function (cdf) of the standard normal

Figure 3.1: $c(\cdot)$ estimation results. Red: the true curve $c(\cdot)$; Black: the median curve of $\widehat{c}(\cdot)$; Filled curves: the 2.5% and 97.5% quantiles of $\widehat{c}(\cdot)$.

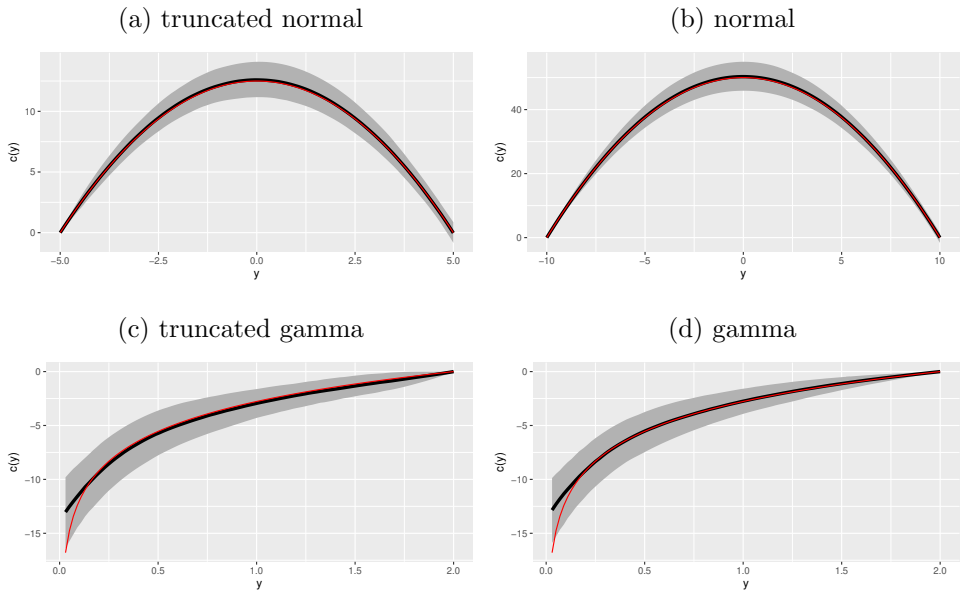


Table 3.1: $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ estimation results under the truncated normal distribution.

	|bias|			σ_{sim}			$\hat{\sigma}_{\text{est}}$			C.I.		
	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE
β_1	.049	.049	.263	.061	.061	.035	.060	-	.035	.947	-	.000
β_2	.083	.084	.530	.103	.103	.041	.101	-	.036	.951	-	.000
β_3	.118	.118	.793	.145	.145	.044	.145	-	.035	.949	-	.000
ξ_1	.022	.023	.028	.028	.029	.035	.029	-	.035	.948	-	.951
ξ_2	.028	.027	.033	.035	.034	.041	.035	-	.036	.941	-	.912
ξ_3	.032	.031	.036	.039	.038	.044	.043	-	.035	.964	-	.881

distribution. We set $\boldsymbol{\beta} = (1, 2, 3)^{\mathrm{T}}$ in the simulation procedure. For the non-truncated case, we generated replicates with the same parameters but without truncation. The non-truncated simulation design is identical to that of Lin et al. (2021).

To evaluate the performance under the truncated case, we illustrate in Figure 3.1a how the estimated $c(\cdot)$ performed by our method while fixing $c(-5) = 0$. As can be seen from the plot, $\widehat{c}(\cdot)$ approximated $c(\cdot)$ with satisfactory bias and variance at sample size 1000. This reflects the theoretical properties described in Lemma B.6.1 of the supplementary material. In addition, in Table 3.1, we illustrate the estimation properties for $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ via aMLE, pMLE and MLE. The aMLE and pMLE methods were numerically almost identical in terms of both bias and standard error, reflecting the theoretical properties established in Proposition 3.3.3 as well as those in Theorem 1 of Lin et al. (2021). In terms of inference of aMLE, the estimated standard error was very close to the sample standard error, and the coverage rate of the confidence intervals was close to

Table 3.2: $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ estimation results under the normal distribution.

	|bias|			σ_{sim}			$\hat{\sigma}_{\text{est}}$			C.I.		
	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE
β_1	.046	.049	.026	.057	.064	.032	.055	-	.032	.938	-	.954
β_2	.081	.088	.025	.099	.114	.032	.096	-	.032	.946	-	.957
β_3	.116	.127	.025	.143	.166	.032	.139	-	.032	.940	-	.941
ξ_1	.026	.029	.026	.032	.038	.032	.032	-	.032	.953	-	.954
ξ_2	.025	.030	.025	.032	.038	.032	.035	-	.032	.961	-	.957
ξ_3	.026	.030	.025	.032	.038	.032	.038	-	.032	.964	-	.941

the nominal level 95%. These indicate that our asymptotic properties are already useful at sample size $n = 1000$ in this model. In contrast, the inference results of MLE were very bad, with the coverage rate almost 0. This is a direct consequence of the estimation bias caused by model misspecification. Table B.1 in the supplementary material provides the estimation results of $\boldsymbol{\eta}_\tau$ at the quantile levels $\tau = 0.05, 0.25, 0.5, 0.75$ and 0.95 respectively. Again, the estimation and inference properties of aMLE were satisfactory, suggesting that the properties described in Theorem 3.3.2 is reflected in this model for $n = 1000$. Also, MLE performed poorly in estimating $\boldsymbol{\eta}_\tau$ especially at the low or high quantile level.

For the non-truncated case, the corresponding results are shown in Figure 3.1b and Tables 3.2 and B.2 in the supplementary material. We can see from Figure 3.1b that $\widehat{c}(\cdot)$ estimated the curve $c(\cdot)$ sufficiently well, even though the response is infinitely supported hence our compact support assumption is violated. We also note the small biases of the estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\xi}}$, and $\widehat{\boldsymbol{\eta}}_\tau$ from Tables 3.2 and B.2. Moreover, interestingly, our estimators of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}_\tau$ performed as well as MLE in this simulation setting, even when τ is near 0 or 1. In terms of inference, the empirical coverage of the estimated confidence interval by our method was close to the nominal level. These seem to suggest an empirical robustness property of our aMLE against the compact support assumption. Lastly, we point out that our method aMLE was more efficient than pMLE in this example in terms of estimating $\boldsymbol{\beta}, \boldsymbol{\xi}$ as well as the marginal quantile effect $\boldsymbol{\eta}_\tau$ at various quantile levels τ.

3.4.2 Gamma distribution

Next, we consider the situation when Y given $\mathbf{x}$ has a gamma distribution. We first generated a covariate vector $\mathbf{X}_i = (X_{1i}, X_{2i})^{\mathrm{T}}$, where $X_{ki}, k = 1, 2$ are independently and identically distributed (iid) as a uniform random variable on $[0.5, 1]$. Then for the truncated case, a response Y_i was generated independently from a truncated gamma

Table 3.3: $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ estimation results under the truncated gamma distribution.

	|bias|			σ_{sim}			$\hat{\sigma}_{\text{est}}$			C.I.		
	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE
β_1	.097	.114	.074	.124	.141	.074	.122	-	.076	.944	-	.889
β_2	.101	.117	.062	.128	.145	.077	.128	-	.077	.942	-	.945
ξ_1	.064	.284	.108	.081	.154	.059	.080	-	.061	.942	-	.589
ξ_2	.064	.523	.117	.080	.126	.061	.079	-	.061	.947	-	.531

distribution with the pdf

$$f_{Y|\mathbf{X}}(y;\alpha,\theta_i,b) = \frac{y^{\alpha-1}e^{-y/\theta_i}}{\Gamma(\alpha)\theta_i^{\alpha}} \left\{ \int_0^b \frac{y^{\alpha-1}e^{-y/\theta_i}}{\Gamma(\alpha)\theta_i^{\alpha}} \right\}^{-1},$$

where $\alpha = 5$, $\alpha\theta_i = 1/\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$, $b = 2$, and $\boldsymbol{\beta} = (0.5, 1)^{\mathrm{T}}$. For the non-truncated case, we simply carried out the same data generation mechanism while applying the usual gamma distribution with the same parameters.

For the truncated case, Figure 3.1c illustrates the performance of the estimator $\widehat{c}(\cdot)$ with $c(2) = 0$. The estimation had very small bias on most part of the support except when y is close to 0 due to the boundary effect. Indeed, the true curve $c(y) = (\alpha - 1)\log(y)$ is unbounded near 0 and very few observations are available since the density converges to 0. Nevertheless, we can see from Tables 3.3 and 3.4 that aMLE estimated $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\eta}_\tau$ well with small bias, and its inference was also sufficiently precise with a good match between the sample variance and the estimated variance, and the 95% confidence intervals had coverage close to the nominal level. In contrast, MLE showed bias and did not perform well in general, and deteriorated further when τ is near 1. In our application, we also find that pMLE was computationally unstable in this setting and did not lead to reasonable results.

For the case when the response is not truncated, Figure 3.1d shows that similar performance was observed as in the truncated gamma regression case. In addition, Tables 3.5 and B.3 in the supplementary material suggest that aMLE still provided good results in estimating the corresponding parameters and can be used for reliable inference. In this situation, aMLE still outperformed pMLE numerically even though both should be inconsistent in theory. Because MLE assumes a fully parametric model which happens to be correct, it had the best performance among all three methods as we expected.

Table 3.4: $\boldsymbol{\eta}_\tau$ estimation results under the truncated gamma distribution.

τ		Method	|bias|	σ_{sim}	$\widehat{\sigma}_{\text{est}}$	C.I.
0.05	η_1	aMLE	0.032	0.041	0.041	0.945
		pMLE	0.120	0.077	-	-
		MLE	0.030	0.025	0.025	0.825
	η_2	aMLE	0.038	0.048	0.048	0.945
		pMLE	0.217	0.090	-	-
		MLE	0.023	0.026	0.027	0.943
0.25	η_1	aMLE	0.052	0.066	0.065	0.942
		pMLE	0.395	0.163	-	-
		MLE	0.048	0.041	0.041	0.826
	η_2	aMLE	0.056	0.071	0.070	0.938
		pMLE	0.750	0.127	-	-
		MLE	0.036	0.042	0.044	0.947
0.50	η_1	aMLE	0.068	0.086	0.085	0.943
		pMLE	0.436	0.192	-	-
		MLE	0.071	0.055	0.056	0.799
	η_2	aMLE	0.071	0.089	0.089	0.939
		pMLE	0.824	0.136	-	-
		MLE	0.055	0.058	0.059	0.919
0.75	η_1	aMLE	0.083	0.106	0.105	0.950
		pMLE	0.296	0.181	-	-
		MLE	0.118	0.073	0.074	0.669
	η_2	aMLE	0.087	0.109	0.107	0.945
		pMLE	0.529	0.144	-	-
		MLE	0.112	0.077	0.079	0.761
0.95	η_1	aMLE	0.089	0.113	0.111	0.936
		pMLE	0.120	0.109	-	-
		MLE	0.336	0.106	0.107	0.113
	η_2	aMLE	0.089	0.112	0.113	0.943
		pMLE	0.220	0.141	-	-
		MLE	0.496	0.111	0.115	0.008

Table 3.5: $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ estimation results under the gamma distribution.

	|bias|			σ_{sim}			$\hat{\sigma}_{\text{est}}$			C.I.		
	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE
β_1	.088	.372	.061	.112	.578	.077	.111	-	.076	.948	-	.954
β_2	.093	.372	.064	.119	.541	.080	.119	-	.078	.944	-	.943
ξ_1	.073	1.040	.051	.093	.986	.064	.092	-	.065	.941	-	.951
ξ_2	.079	1.551	.055	.099	.782	.068	.098	-	.065	.944	-	.942

Table 3.6: $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ estimation results under the Bernoulli distribution.

	|bias|			σ_{sim}			$\hat{\sigma}_{\text{est}}$			C.I.		
	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE
β_1	.089	.089	.088	.111	.111	.111	.109	-	.109	.950	-	.949
β_2	.088	.088	.088	.112	.112	.111	.110		.110	.947		.946
β_3	.100	.099	.099	.123	.123	.123	.124	-	.124	.956	-	.954
ξ_1	.015	.015	.015	.019	.019	.019	.019	-	.019	.950	-	.949
ξ_2	.015	.015	.015	.020	.020	.020	.019		.019	.945		.944
ξ_3	.014	.014	.014	.017	.017	.017	.017	-	.017	.947	-	.948

3.4.3 Bernoulli distribution

Our first simulation for the discrete response was carried out under a conditional Bernoulli distribution with the intention to investigate the performance of aMLE in the discrete response case. A three dimensional covariate vector $\mathbf{X}_i$ was generated independently from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\Sigma = (\sigma_{kl})$ where $\sigma_{kl} = 0.1^{|k-l|}, k = 1, \ldots, 3, l = 1, \ldots, 3$. Then a binary response Y_i was drawn independently from a Bernoulli distribution with a success rate $1/\{1 + \exp(-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}$. We set $\boldsymbol{\beta} = (-0.5, 0.5, 1)^{\mathrm{T}}$.

The performances of the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\xi}}$ by the three methods are given in Table 3.6. All methods performed similarly to each other, because the problem is fully parametric as explained in Section 3.2.3.

3.4.4 Poisson and negative binomial distributions

We now conduct simulation studies for the case where Y is discrete with infinite support. Specifically, Y has Poisson and negative binomial distributions respectively. We first generated a covariate vector $\mathbf{X}_i = (X_{1i}, X_{2i})^{\mathrm{T}}$, where $X_{ki}, k = 1, 2$ are independently and identically distributed (iid) uniform random variables on $[0.5, 1]$. Then for the Poisson regression model, a response Y_i was generated from a Poisson distribution with the rate $\theta_i = \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ where $\boldsymbol{\beta} = (0, 1)^{\mathrm{T}}$. For the negative binomial case, we generated Y_i independently from a negative binomial distribution with the mass function

$$\mathrm{pr}(Y = y; r, \theta_i) = \binom{y + r - 1}{r - 1} \theta_i^y (1 - \theta_i)^r,$$

where $r = 2$, $\theta_i = \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$, and $\boldsymbol{\beta} = (0, -1)^{\mathrm{T}}$. In estimation, MLE assumes $\theta_i = \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ and $c(y) = -\log(y!)$ hence results in the most efficient estimator under the Poisson regression among the three methods, but wrongly specifies the structure when the conditional distribution is the negative binomial.

Table 3.7: $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ estimation results under the Poisson distribution.

	|bias|			σ_{sim}			$\hat{\sigma}_{\text{est}}$			C.I.		
	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE
β_1	.121	.116	.088	.154	.146	.110	.151	-	.109	.939	-	.944
β_2	.130	.130	.087	.162	.157	.109	.158	-	.106	.947	-	.938
ξ_1	.258	.268	.187	.327	.337	.236	.320	-	.233	.938	-	.944
ξ_2	.267	.296	.195	.334	.365	.243	.326	-	.237	.940	-	.934

Table 3.8: $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ estimation results under the negative binomial distribution.

	|bias|			σ_{sim}			$\hat{\sigma}_{\text{est}}$			C.I.		
	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE	aMLE	pMLE	MLE
β_1	.088	.075	1.388	.112	.095	.171	.113	-	.114	.956	-	.000
β_2	.103	.158	.383	.130	.115	.176	.128	-	.120	.951	-	.202
ξ_1	.336	.395	2.535	.426	.502	.340	.430	-	.223	.958	-	.000
ξ_2	.424	.886	2.708	.528	1.008	.322	.518	-	.217	.947	-	.000

Tables 3.7 and 3.8 show the $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ estimation results under the Poisson and the negative binomial regression models respectively. In both cases, we found that aMLE estimated $\boldsymbol{\xi}$ more efficiently than pMLE even though our method aMLE showed similar performance to pMLE in estimating $\boldsymbol{\beta}$. Especially for the estimation of ξ_2 in Table 3.8, we can see that aMLE was about twice as efficient as pMLE in terms of both bias and standard error. This supports that our estimation procedure is more suitable for estimating $\boldsymbol{\xi}$ in finite sample situations than the method proposed by Lin et al. (2021). Also, considering that the empirical coverage of our approximate confidence interval reached the nominal level of 95%, the results confirm the asymptotic properties of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\xi}}$ proposed in Proposition 3.3.4 and Theorem 3.3.6. On the other hand, MLE in Table 3.7 indeed had the best performance in terms of both bias and variance, benefiting from the perfectly correct parametric model setting. However, we can observe from Table 3.8 that the estimators from MLE were severely biased when $c(y)$ was misspecified.

3.5 Data application

We now analyze a data set concerning non-labor income situation in Switzerland. The data set is publicly available from the R-package `AER` (Kleiber & Zeileis 2008), and consists of information collected from 871 married women randomly drawn from the representative health survey for Switzerland (SOMIPOPS) in 1981 with income not abnormally low.

Our goal is to estimate the marginal effect and the marginal quantile effect of covariates on non-labor income (such as husband's income). Thus, the non-labor income in the log scale is used as response Y. All the other variables in the dataset are considered

Figure 3.2: $c(\cdot)$ estimation in the Swiss non-labor income data. Black: the curve $\widehat{c}(\cdot)$; Filled curves: the estimated pointwise confidence band of $c(\cdot)$.

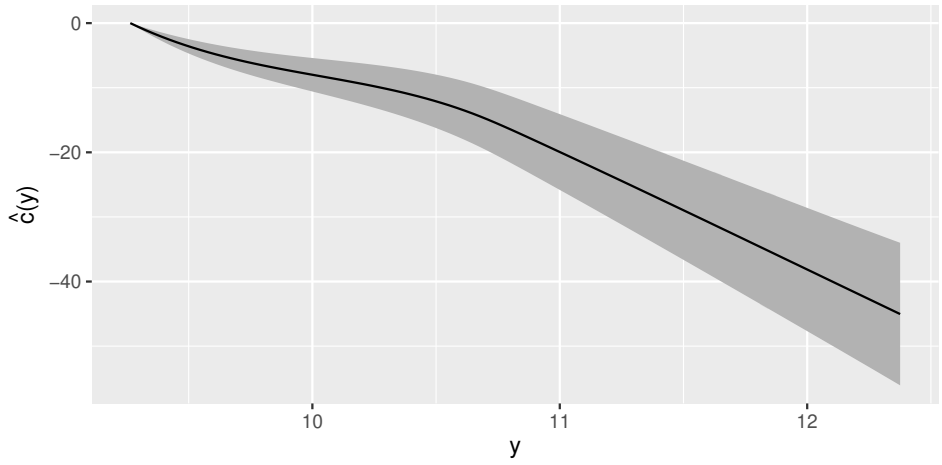


Table 3.9: AIC and BIC in the Swiss non-labor income data.

Method	AIC	BIC
aMLE	538.800	600.806
Normal	667.201	710.128
Gamma	657.090	700.016

as explanatory variables, and they include Participation (taking the value 1 if the individual participated in the labor market, 0 otherwise), Age (age in years), Age2 (squared age in years then divided by 10), Education (years of formal education), Foreign (taking the value 1 if the individual is a permanent foreign resident, 0 otherwise), Youngkids (number of young children), Oldkids (number of older children). The age category for children was decided by whether the age is under 7. For more detailed description, see Gerfin (1996).

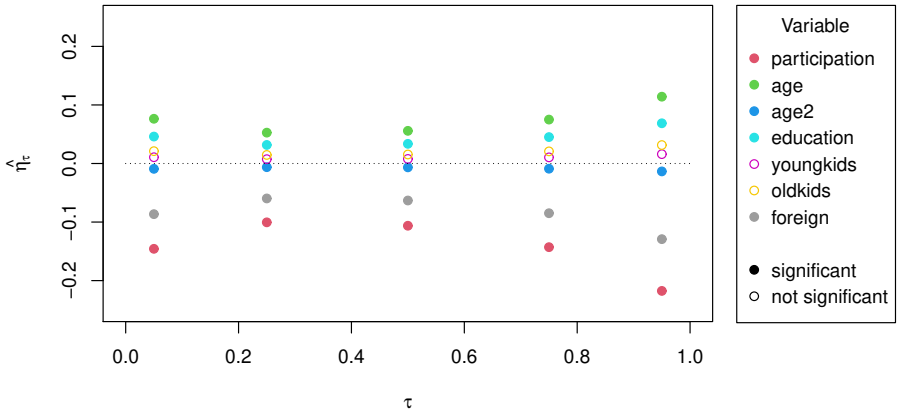
We implemented our method aMLE. For comparison, we also implemented the normal and the gamma regression. We fitted the models without any transformation on or interaction between the covariates, while the number of knots of the B-spline basis is chosen in the same way as explained in Section 3.4.

In Figure 3.2, we can see that the fitted distribution from aMLE does not resemble that of the normal regression, whose $c(y) = -y^2/2\sigma^2$ with $\sigma > 0$. It also does not resemble that of the gamma distribution whose $c(y) = (\alpha - 1)\log(y)$ with $\alpha > 0$. In Table 3.9, we compared the information criteria AIC and BIC of the three models, and found that our aMLE on (3.4) has the lowest AIC and BIC compared to the normal and

Table 3.10: $\boldsymbol{\xi}$ estimation results in the Swiss non-labor income data. "*" indicates the significance of the corresponding predictor at 5% significance level.

Variable	$\widehat{\xi}_{\text{aMLE}}$	p-value	$\widehat{\xi}_{\text{Normal}}$	p-value	$\widehat{\xi}_{\text{Gamma}}$	p-value
Participation	-0.133	<1e-3*	-0.130	<1e-3*	-0.130	<1e-3*
Age	0.070	<1e-3*	0.065	<1e-3*	0.065	<1e-3*
Age2	-0.008	<1e-3*	-0.008	<1e-3*	-0.008	<1e-3*
Education	0.042	<1e-3*	0.042	<1e-3*	0.042	<1e-3*
Youngkids	0.010	0.695	0.011	0.657	0.011	0.654
Oldkids	0.019	0.147	0.023	0.091	0.022	0.095
Foreign	-0.079	0.017*	-0.060	0.064	-0.062	0.053

Figure 3.3: The result of $\boldsymbol{\eta}_\tau$ estimation in the Swiss non-labor income data.



the gamma regression model, hence is the most suitable modeling choice. This suggests that our model in (3.4) is the most suitable to use in analyzing the swiss non-labor income data.

Table 3.10 shows the inference result in estimating the marginal effect. All except one explanatory variables have similar levels of significance based on all three methods. The exception is "Foreign", which was selected as a significant variable at 5% level in the analysis based on aMLE, but was considered non-significant in both the normal and the gamma regression. In other words, only by using (3.4) in combination with aMLE, we can conclude that being a permanent foreign resident has a negative effect on the non-labor income in Switzerland.

In Figure 3.3, we further illustrate the estimated $\widehat{\boldsymbol{\eta}}_\tau$ at the quantile level $\tau = 0.05$, 0.25, 0.5, 0.75 and 0.95. We fixed the significance level at 5%. The p-values of $\widehat{\boldsymbol{\eta}}_\tau$ were largely similar to those of $\widehat{\boldsymbol{\xi}}$. However, as we can see from the plot, the magnitude of the

marginal quantile effect increases when τ is near 0 or 1. That is, our model suggests that when the non-labor income level is relatively low or high, the effect of the covariates on the non-labor income is more extreme.

3.6 Discussion

We have proposed a B-spline based approximate maximum likelihood estimation procedure to estimate both the marginal effect and the marginal quantile effect in a semiparametric generalized linear model. Compared to the classical GLM, our model is more flexible hence less susceptible to model misspecification. The estimators we proposed are shown to reach the semiparametric efficiency bounds, hence are optimal.

Following the spirit of GLMs, we have worked with a linear summary of covariates $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ in our work. It is easy to see that we can replace $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ by a more general form $m(\mathbf{x}, \boldsymbol{\beta})$, where m is a known function which can be nonlinear. All our procedures can be carried through while replacing $\mathbf{x}$ by $\partial m(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$, where we only need to ensure identifiability, sufficient smoothness, and boundedness of the corresponding quantities to facilitate the almost identical derivation to the linear case. Throughout the article, we have assumed the data to be iid. This assumption can be relaxed to a certain extent. Specifically, even if the observations are correlated, the developed estimation procedures are still valid and the resulting estimators for $\boldsymbol{\beta}, c(y), \boldsymbol{\xi}$, and $\boldsymbol{\eta}_\tau$ are still consistent and have asymptotically normal distributions under suitable conditions. This can be understood from a composite likelihood point of view, in that the likelihood treating the observations as iid is in fact a composite likelihood when the observations are truly correlated. However, the asymptotic variance and inference results no longer hold in general. To this end, Chen et al. (2014) established that the dependency is ignorable when the correlation is weak, while in general, the Martingale central limit theorem needs to be employed to establish the asymptotic results and inference tools.

An interesting but difficult further extension of our work worth mentioning is the possibility of allowing non-compactly supported distribution of Y. Although truncation is routinely done in practice, we found it very difficult to rigorously extend our theoretical analysis to encompass this scenario. We suspect that more fundamental work is needed, possibly in a model simpler than the semiparametric GLM proposed in (3.4), for example, a purely nonparametric model.

3.7 Acknowledgments

This chapter is based upon work supported by the National Science Foundation and the National Institute of Health. Any opinions, findings, and conclusions or recommendations expressed in this chapter are those of the authors, and do not necessarily reflect the views of the National Science Foundation and the National Institute of Health.

Chapter 4
Optimal sampling for positive only electronic health record data

4.1 Introduction

Identification of patients with the phenotype of interest is an essential step in utilizing electronic health records (EHR) data for research. Frequently, EHR are "positive-only", where we only have a small group of confirmed positive cases, while the rest of patients are left with unknown phenotype statuses. For example, patients with depression symptoms often need special treatment and management during their hospitalizations. At their admissions, physicians may use their EHR to assess their risk of depression, so they could choose proper interventions earlier to prevent deterioration and improve their chance of survival and recovery. The EHR of a patient diagnosed with depression may contain ICD codes relevant to depression, which can help identify the patient as a case of having depression symptoms. However, patients whose EHRs do not contain depression-related ICD codes may still have the symptoms. In Gehrmann et al. (2018), a group of medical experts carefully reviewed clinical notes of 415 ICU patients from the MIMIC-III EHR data and identified 148 patients manifesting depression symptoms. Among the 148 patients, only 48 patients had depression diagnosis in their EHRs. This kind of positive only data, consisting of a small number of cases ("positives") and a large number of patients with unknown phenotypes ("unlabeled"), is common in EHR-based research (Zhang et al. 2020).

Because traditional manual chart review is labor-intensive hence infeasible for millions of patients' records in modern EHR data, it is common to rely on automatic algorithms to infer patients' phenotype statuses from their EHR (e.g., Hripcsak & Albers 2013, Hou et al. 2020, Zhang et al. 2020). To develop and train such algorithms, gold

standard phenotype statuses for a group of patients are needed. Due to the positive-only nature of EHR, the essential task here is to ascertain the phenotype statuses for a group of the unlabeled patients, and combine them with existing true-positive cases to form a training sample. Furthermore, the size of the training sample with complete labels will be substantially limited because phenotyping the unlabeled patients does require cost and labor intensive manual chart-review. These obstacles motivate the question of central interest to us — what is the optimal strategy for selecting the unlabeled patients? Here, the ultimate goal of the optimal sampling is to form a training dataset based on which the trained model achieves the optimal classification accuracy.

The subsampling technique has been widely used to relax computational burden or labor and experimental cost (Drineas et al. 2006). It aims to obtain a solution that is both efficient and practical by sampling a subset from a full dataset and performing estimation based on it. Recently, the optimal sampling design approaches have received more attention, which not only reduce the costs but further improve the efficiency of subsequent estimators. See Wang et al. (2018) and Wang & Ma (2021) for more detailed discussion. However, these procedures are not applicable in our context because the selection probabilities depend on both covariates and responses, while in our problem context, most of the responses are not available. On the other hand, Tan & Heagerty (2020) and Yin et al. (2022) proposed cost-effective procedures to handle health record data that minimize certain estimation criteria. However, these two papers focus on estimator augmentation after stratified sampling rather than optmal sampling design itself, and are specific to the logistic regression framework.

Unlike these existing literatures, we focus on devising a design that optimizes classification performance through rigorously investigating the sampling design based on the partially available information in combination with proper estimation procedure. Specifically, the key contributions of our work are the following. First, our proposed method determines optimal sampling weights by incorporating the surrogate indicator in the sampling scheme, when a given data set has the positive-only nature. Second, our method is model-free in the sense that it is built on an arbitrary working classification model that does not have to be correctly specified. The method is flexible in that it can be applied to any working models. As a by-product, our method also yields the optimal model parameter estimation in terms of prediction MSE when the working model happens to be correct. In these perspectives, our method generalizes and improves the developments of Tan & Heagerty (2020) and Yin et al. (2022).

We now summarize the research question into a general statistical problem. Let

$i = 1, \ldots, N$. Let Y_i denote the latent binary phenotype of interest of patient i, such as a certain disease status. We denote $Y_i = 1$ if the patient has the disease, while $Y_i = 0$ if the patient does not have the disease. We equivalently name the diseased patient a case, and name the disease-free patient a control. Let $\mathbf{X}_i$ denote a vector of clinical variables that are predictive of Y_i and is fully observed. Let S_i be an observed binary variable. We name S_i an "anchor variable". S_i is linked to Y_i as the following. If $S_i = 1$, then $Y_i = 1$. If $S_i = 0$, then Y_i can be either 0 or 1. Note that this implies that there is no false alarm, i.e. $\text{pr}(S_i = 1 \mid Y_i = 0) = 0$. We refer to the patients with $S_i = Y_i = 1$ as "anchor-positive cases", and those "$S_i = 0$" as "unlabeled" patients. In the motivating example, S_i is the binary indicator whether the patient has the diagnosis record of depression in EHR. Finally, we define R_i as an indicator of whether the ith patient will be selected into the training data and ascertain his or her disease status. By the definition of S_i, $S_i = 1$ implies that the ith patient has the disease, hence we automatically have $R_i = 1$ and $R_i Y_i = 1$. When $S_i = 0$, i.e., if the ith observation is unlabeled and Y_i is missing, then R_i will be either 0 or 1 depending on whether the ith observation is selected into the training data set for identifying its label Y_i. Accordingly, the observations with $R_i = 1$ will form a subsample with complete labels, and will hence be used to train the classification model. In summary, the observed data is $(\mathbf{X}_i, S_i, R_i, R_i Y_i), i = 1, \ldots, N$. Without loss of generality, we assume $S_i = 0$ for $i = 1, \ldots, N_0$ and $S_i = 1$ for $i = N_0 + 1, \ldots, N$. Due to the nature of EHR, N is very large so we can view the data as the whole population, and view the quantities estimated based on the N observations as completely known.

Let a sampling probability be $\pi(\mathbf{X}_i) = \text{pr}(R_i = 1 \mid \mathbf{X}_i, S_i = 0)$. Our goal is to find a good $\pi(\cdot)$ so that we can randomly sample a subset according to $\pi(\mathbf{X}_i)$'s from the data $(\mathbf{X}_i, S_i = 0), i = 1, \ldots, N_0$, obtain the corresponding Y_i's, combine with those with $S_i = 1$ hence $Y_i = 1$ to form a training data set. Note that when $S_i = 1$, we do not need subsample since we know $Y_i = 1$. In a way, we can treat all the observations with $S_i = 1$ as sampled with probability 1 since they are included in the training data automatically. On the other hand, we do not have any Y_i values for those with $S_i = 0$ when perform sampling, so $\pi(\mathbf{X}_i)$ can only depend on $\mathbf{X}_i$ since this is all we have. To this end, we also define $\pi_T(\mathbf{X}_i, S_i) \equiv \text{pr}(R_i = 1 \mid \mathbf{X}_i, S_i) = (1 - S_i)\pi(\mathbf{X}_i) + S_i$, the probability of selecting any Y_i in the overall population (i.e. any patient in the EHR).

Recall that our eventual target is to perform classification on patients' phenotypes. For this purpose, let $p_0(\mathbf{X}) = \text{pr}(Y = 1 \mid \mathbf{X})$ and $\text{pr}(S = 1 \mid \mathbf{X}) = a(\mathbf{X})$. We can view $a(\mathbf{X})$ as known because it can be estimated based on the entire data set with size N. Also, let $c(\mathbf{X}) \equiv \text{pr}(S = 1 \mid \mathbf{X}, Y = 1)$ be the anchor sensitivity. Note that $a(\mathbf{X}) = c(\mathbf{X}) p_0(\mathbf{X})$.

To perform classification we will need a working model for $p_0(\mathbf{X})$, say $p(\mathbf{X}, \boldsymbol{\beta})$ and we will need to estimate $\boldsymbol{\beta}$, say the estimator is $\widehat{\boldsymbol{\beta}}$. Now, denote a new observation be $(\mathbf{X}^*, S^*, Y^*)$. We then classify Y^* using $p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}})$ for a new observation if only $\mathbf{X}^*$ is available, and classify using $\text{pr}(Y^* = 1 \mid \mathbf{X}^*, S^* = 0) = \{p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}) - a(\mathbf{X}^*)\}/\{1 - a(\mathbf{X}^*)\}$ if in addition we have $S^* = 0$. In this paper, we propose an optimal sampling strategy $\pi(\mathbf{X}_i)$ that leads to a subsequent classification model with the highest classification accuracy.

4.2 Main contributions

4.2.1 Methodology

For simplicity, we first consider classifying the Y^* value of a future observation $\mathbf{X}^*$ without knowing the corresponding S^* value. Consider the case when $p(\mathbf{x}, \boldsymbol{\beta})$ is an approximate model in the whole EHR, i.e., we allow $p_0(\mathbf{x}) \neq p(\mathbf{x}, \boldsymbol{\beta})$ for any $\boldsymbol{\beta}$. Our goal is to find a sampling strategy to minimize the mean squared error (MSE) of $p(\mathbf{x}^*, \widehat{\boldsymbol{\beta}})$, i.e. $E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}})\}^2]$. We can decompose $E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}})\}^2]$ into two parts, $E[\{Y^* - p_0(\mathbf{X}^*)\}^2]$ and $E[\{p_0(\mathbf{X}^*) - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}})\}^2]$. Since the first term is not relevant to the estimator $\widehat{\boldsymbol{\beta}}$, we aim at minimizing $E[\{p_0(\mathbf{X}^*) - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}})\}^2]$ which entails the discrepancy between the true model and the estimated model. We will show that this is an equivalent problem to minimizing

$$E\left\{\frac{p^2(\mathbf{X}, \boldsymbol{\beta}^*) + p_0(\mathbf{X}) - 2p_0(\mathbf{X})p(\mathbf{X}, \boldsymbol{\beta}^*)}{(1-S)\pi(\mathbf{X}) + S}\mathbf{p}'^{\,\mathrm{T}}_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*)\mathbf{A}^{-1}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*)\right\}$$

with respect to a sampling strategy $\pi(\mathbf{X})$, where $\boldsymbol{\beta}^*$ is the unique minimizer of $E[\{p_0(\mathbf{X}) - p(\mathbf{X}, \boldsymbol{\beta})\}^2]$ with respect to $\boldsymbol{\beta}$, $\mathbf{A} \equiv E[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*)^{\otimes 2} + \mathbf{p}''_{\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}}(\mathbf{X}, \boldsymbol{\beta}^*)\{p(\mathbf{X}, \boldsymbol{\beta}^*) - p_0(\mathbf{X})\}]$, and $\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}, \boldsymbol{\beta})$ and $\mathbf{p}''_{\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}}(\mathbf{x}, \boldsymbol{\beta})$ are the gradient and the hessian of $p(\mathbf{x}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$.

Without any surprise, the optimal choice of $\pi(\mathbf{x})$ depends on the true $p_0(\mathbf{x})$ and $\boldsymbol{\beta}^*$, both are unknown. Hence we need reasonable approximations of $p_0(\mathbf{x})$ and $\boldsymbol{\beta}^*$. Hereby, we assume that we are given a consistent estimator of $\boldsymbol{\beta}^*$, $\widetilde{\boldsymbol{\beta}}$. For example, the estimator $\widetilde{\boldsymbol{\beta}}$ can be obtained by minimizing the sample MSE of $p(\mathbf{x}, \boldsymbol{\beta})$, $m^{-1}\sum_{i=1}^{m}\{y_i - p(\mathbf{x}_i, \boldsymbol{\beta})\}^2$ with respect to $\boldsymbol{\beta}$ based on either external data or pilot study, $(\mathbf{x}_i, y_i), i = 1, \ldots, m$.

Once we have $\widetilde{\boldsymbol{\beta}}$, the simplest thing to do is to replace $\boldsymbol{\beta}^*$ and $p_0(\mathbf{x})$ with $\widetilde{\boldsymbol{\beta}}$ and

$p(\mathbf{x}, \widetilde{\boldsymbol{\beta}})$ respectively, and carry out the approximate optimization problem

$$\pi_{\text{opt}}(\mathbf{x}) \quad = \quad \text{argmin}_{\pi} E\left[\frac{p(\mathbf{X}, \widetilde{\boldsymbol{\beta}})\{1-p(\mathbf{X}, \widetilde{\boldsymbol{\beta}})\}}{(1-S)\pi(\mathbf{X})+S}\mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{X}, \widetilde{\boldsymbol{\beta}})\mathbf{A}^{-1}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \widetilde{\boldsymbol{\beta}})\right]$$

subject to the constraints $0 < \pi(\mathbf{x}) \leqslant 1$ and $E\{\pi(\mathbf{X})|S=0\} \leqslant \pi_0$, where $\mathbf{A}$ is simplified to $E\{\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \widetilde{\boldsymbol{\beta}})^{\otimes 2}\}$. Here, π_0 is a prefixed constant determined by the limit on resources. Because we consider the N observations as the whole population, writing $\mathbf{w} = (w_1, \ldots, w_{N_0})^{\mathrm{T}}$, the above reduces to

$$\mathbf{w}_{\text{opt}} \quad = \quad \text{argmin}_{\mathbf{w}} \sum_{i=1}^{N_0} \frac{p(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})\{1-p(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})\}}{w_i}\mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})\mathbf{A}_N^{-1}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}}) \tag{4.1}$$

subject to the constraints $0 < w_i \leqslant 1$ for $i = 1, \ldots, N_0$ and $\sum_{i=1}^{N_0} w_i \leqslant N_0\pi_0$, where $\mathbf{A}_N \equiv N^{-1}\sum_{i=1}^{N} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})^{\otimes 2}$. Note that in (4.1), the minimizer clearly is obtained at $E\{\pi(\mathbf{X})|S=0\} = \pi_0$, i.e., $\sum_{i=1}^{N_0} w_i = N_0\pi_0$. So we can restrict the search region on the boundary.

Once we have $\mathbf{w}_{\text{opt}}$, we sample R_i from Bernoulli(w_i) for $i = 1, \ldots, N_0$ and set $R_i = 1$ for $i = N_0+1, \ldots, N$ in order to obtain the observations $(\mathbf{X}_i, S_i, R_i, R_iY_i), i = 1, \ldots, N$. In other words, we measure the Y_i values of the observations with $(S_i, R_i) = (0, 1)$. Then we solve the estimating equation

$$\sum_{i=1}^{N} \frac{R_i}{(1-S_i)w_i + S_i}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta})\left\{p(\mathbf{x}_i, \boldsymbol{\beta}) - Y_i\right\} = \mathbf{0} \tag{4.2}$$

with respect to $\boldsymbol{\beta}$.

Note that given any fixed sampling weights $w_i, i = 1, \ldots, N_0$, solving (4.2), i.e., minimizing the inverse probability weighted sample MSE, provides the most efficient estimator for $\boldsymbol{\beta}^*$ under the situation $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$ (Qin et al. 2017). The most efficient estimator $\widehat{\boldsymbol{\beta}}$ will also lead to the minimum MSE of $p(\mathbf{x}, \widehat{\boldsymbol{\beta}})$. Indeed, theoretical analysis in Section 4.2.2 will show that our proposed procedure guarantees the optimal classification MSE among all possible estimators of $\boldsymbol{\beta}$. Note that an equivalent representation of the sample MSE in the binary classification context is $N^{-1}\sum_{i=1}^{N} R_i/\{(1-S_i)w_i + S_i\}[[Y_i\{1-p(\mathbf{x}_i, \boldsymbol{\beta})\}^2 + (1-Y_i)p(\mathbf{x}_i, \boldsymbol{\beta})^2]$, hence we can interpret minimizing MSE as minimizing the total squared mis-classification probabilities. In addition, our estimation procedure can also be viewed as a likelihood-based approach where a normal working model $Y \sim N\{p(\mathbf{x}, \boldsymbol{\beta}), \sigma^2\}$ is adopted. Although the response variable is binary,

we can adopt a working model for a general response because our procedure does not require the model to be correct. Empirical evidences will be demonstrated through simulation studies and a data example in Sections 4.3 and 4.4 to support the theory. Even if $p(\mathbf{x}, \boldsymbol{\beta}^*) \neq p_0(\mathbf{x})$, using $p(\mathbf{x}, \boldsymbol{\beta}^*)$ as a working model for $p_0(\mathbf{x})$ is the most sensible choice, which leads to (4.2) as well, and it still yields a consistent estimator for $\boldsymbol{\beta}^*$.

We now write out the algorithm explicitly.

1. Form $\mathbf{A}_N = N^{-1} \sum_{i=1}^{N} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})^{\otimes 2}$ and obtain the sampling weights as

$$\widehat{\mathbf{w}}_{\text{opt}} \quad = \quad \text{argmin}_{\mathbf{w}} \sum_{i=1}^{N_0} \frac{p(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})\{1 - p(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})\}}{w_i} \mathbf{p}'^{\,\mathrm{T}}_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}}) \mathbf{A}_N^{-1} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})$$

   subject to the constraints $0 < w_i \leqslant 1$ for $i = 1, \ldots, N_0$ and $\sum_{i=1}^{N_0} w_i = N_0 \pi_0$.

2. Sample R_i from Bernoulli($\widehat{w}_i$) for $i = 1, \ldots, N_0$ and set $R_i = 1$ for $i = N_0 + 1, \ldots, N$. Then determine the value of Y_i through extensive manual chart reviews if $(S_i, R_i) = (0, 1)$.

3. Obtain $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{N} \frac{R_i}{(1 - S_i)\widehat{w}_i + S_i} \left\{ y_i - p(\mathbf{x}_i, \boldsymbol{\beta}) \right\}^2 . \tag{4.3}$$

4. Perform classification of Y^* at a new observation $\mathbf{X}^*$ as $\widehat{Y}^* = I\{p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}) > 0.5\}$.

Remark 4.2.1. *We point out that under the constraint* $\sum_{i=1}^{N_0} w_i = N_0 \pi_0$ *alone, the solution* $\mathbf{w}_{\text{opt}}$ *in* (4.1) *is*

$$w_i = N_0 \pi_0 \frac{\sqrt{c_i}}{\sum_{i=1}^{N_0} \sqrt{c_i}} \tag{4.4}$$

for $i = 1, \ldots, N_0$, *where* $c_i \equiv p(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})\{1 - p(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})\} \mathbf{p}'^{\,\mathrm{T}}_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}}) \mathbf{A}_N^{-1} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})$. *Considering the additional constraint* $0 < w_i \leqslant 1$, *we note that* (4.1) *has a closed form solution (4.4) if* $0 < N_0 \pi_0 \sqrt{c_i} / \sum_{i=1}^{N_0} \sqrt{c_i} \leqslant 1$ *for all* $i = 1, \ldots, N_0$.

We end this section with some notes on practical implementation of our method. First, the cutoff value of 0.5 in step 4 of the algorithm can be modified to any other appropriate value. We may take into account additional information such as the commonness/rareness of disease in the population to adjust the cutoff to be larger/smaller

than 0.5. For example, we can set the cutoff value to be the average disease rate in the sample, i.e., $N^{-1}\sum_{i=1}^{N} R_i Y_i/\{(1-S_i)w_i + S_i\}$ (Agresti 2003). Furthermore, the cutoff value may be guided by the intended application of the classification model. For instance, if the purpose of the study is for identifying high risk individuals, the choice can be guided by the desired level of accuracy. In addition, if a pilot data is available to us, it is beneficial to include it. To this end, we can form a pooled dataset combining the pilot data and the data we have been considering so far. We then conduct the similar procedure on the combined data, while assigning $\pi_T(\mathbf{X}_i, S_i) = 1$ and $R_i = 1$ for each observation in the pilot data.

4.2.2 Theory

Now we study the theoretical properties of the estimation procedure stated above. All proofs are relegated to Web Appendices. We require the following regularity conditions.

(C1) $\boldsymbol{\beta}^*$ is the unique minimizer of $E[\{p_0(\mathbf{X}) - p(\mathbf{X}, \boldsymbol{\beta})\}^2]$.

(C2) $\mathbf{A} \equiv E[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}^*, \boldsymbol{\beta}^*)^{\otimes 2} + \{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}\partial^2 p(\mathbf{X}^*, \boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}]$ is nonsingular.

Condition (C1) is very mild and it implies that $\boldsymbol{\beta}^*$ is a solution to $E[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta})\{p_0(\mathbf{X}) - p(\mathbf{X}, \boldsymbol{\beta})\}] = \mathbf{0}$. It often holds naturally under appropriate parameterization when the model $p(\mathbf{x}, \boldsymbol{\beta})$ is true. Condition (C2) naturally holds when the model $p(\mathbf{X}, \boldsymbol{\beta}^*)$ is not drastically misspecified, and is a reasonable requirement in both theory and practice.

From the description in Section 4.2.1, it is clear that the estimator of $\boldsymbol{\beta}$ relies on the sampling weights. Hence, in the following, when we emphasize this dependence, we write $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$, and reserve $\widehat{\boldsymbol{\beta}}$ for $\widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\mathrm{opt}}}$.

Theorem 4.2.1. *Let $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ be the solution of (4.2) and $\mathbf{w}_{\mathrm{opt}}$ be as defined in (4.1). Then under any given $\mathbf{w}$, $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ minimizes the estimated classification MSE. Furthermore, when $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$ and $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, $\mathbf{w}_{\mathrm{opt}}$ minimizes the MSE of $p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})$, i.e., $E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\mathrm{opt}}})\}^2] \leqslant E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}^2]$ for any $\mathbf{w}$ satisfying $0 < w_i \leqslant 1$ for $i = 1, \ldots, N_0$ and $\sum_{i=1}^{N_0} w_i \leqslant N\pi_0$.*

Theorem 4.2.2. *Assume the regularity conditions specified in Robins et al. (1994) and let $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ be the solution of (4.2). When $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$, $E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}^2] \leqslant E[\{Y^* - p(\mathbf{X}^*, \breve{\boldsymbol{\beta}}_{\mathbf{w}})\}^2]$ for any estimator $\breve{\boldsymbol{\beta}}_{\mathbf{w}}$ and $\mathbf{w}$ satisfying $0 < w_i \leqslant 1$ for $i = 1, \ldots, N_0$ and $\sum_{i=1}^{N_0} w_i \leqslant N\pi_0$.*

Theorem 4.2.1 states that our optimal sampling choice $\mathbf{w}_{\text{opt}}$ yields the estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\text{opt}}}$ with the minimum MSE among the class of estimators $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ from (4.2) associated with an arbitrary sampling weight $\mathbf{w}$. On the other hand, Theorem 4.2.2 states that given a sampling strategy $\mathbf{w}$, an estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ obtained by solving (4.2) has a smaller classification MSE compared to any other estimator $\breve{\boldsymbol{\beta}}_{\mathbf{w}}$. Therefore, we conclude that $\widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\text{opt}}}$ provides the minimum MSE among all possible estimators $\breve{\boldsymbol{\beta}}_{\mathbf{w}}$. We highlight this conclusion as a remark below.

Remark 4.2.2. *Theorem 4.2.1 illustrates the optimality of the sampling weights* $\mathbf{w}_{\text{opt}}$ *in terms of minimizing the classification MSE when the working model is true. Obviously, in practice, a working model almost always serves as an approximate model, hence the optimality is also only approximate. This is inevitable as we always need to pay a price for what we are unable to know. On the other hand, the MSE also relies on the choice of the model paramater estimator. To this end, as established in Theorem 4.2.2, our estimator* $\widehat{\boldsymbol{\beta}}$ *is also optimal in the sense that it minimizes the MSE to the leading order, and is efficient when the working model is correct, and is consistent when the working model is misspecified. The optimality of* $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ *translates to the local optimality of the MSE under any weight choice* $\mathbf{w}$*, and to the global optimality of the MSE under* $\mathbf{w}_{\text{opt}}$*. We summarize this result as Corollary 4.2.1.*

Corollary 4.2.1. *Let* $\mathbf{w}_{\text{opt}}$ *be as defined in* (4.1) *and* $\widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\text{opt}}}$ *be the solution of* (4.2)*. Among all the choices of sampling weights* $\mathbf{w}$ *and all the consistent estimators of* $\boldsymbol{\beta}^*$ *that does not require the working model* $p(\mathbf{x}, \boldsymbol{\beta})$ *to be correct,* $\widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\text{opt}}}$ *minimizes the classification MSE if* $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$ *and* $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$*.*

We finally also point out the role that the external information based estimator $\widetilde{\boldsymbol{\beta}}$ plays in Corollary 4.2.2. Essentially, as long as $\widetilde{\boldsymbol{\beta}}$ is a consistent estimator, it is as good as the true $\boldsymbol{\beta}$.

Corollary 4.2.2. *Let* $\mathbf{w}_{\text{opt}}$ *be as defined in* (4.1) *and* $\widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\text{opt}}}$ *be the solution of* (4.2)*. Given that* $\widetilde{\boldsymbol{\beta}}$ *converges to* $\boldsymbol{\beta}^*$ *in probability,* $\widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\text{opt}}}$ *minimizes the classification MSE among all choices of* $\mathbf{w}$ *and* $\breve{\boldsymbol{\beta}}_{\mathbf{w}}$ *if* $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$*.*

Since the optimal sampling weights $\mathbf{w}_{\text{opt}}$ aims at minimizing (4.1) which includes $\widetilde{\boldsymbol{\beta}}$, it is obvious that $\mathbf{w}_{\text{opt}}$ depends on the estimate $\widetilde{\boldsymbol{\beta}}$ from external data or pilot study. However, Corollary 4.2.2 points out that even when $\widetilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*$, our choice of $\mathbf{w}_{\text{opt}}$ and the corresponding $\widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\text{opt}}}$ still results in the minimum classification MSE as long as $\widetilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}^*$.

4.2.3 Classification given $S^* = 0$

We now extend the classification based on $\mathbf{X}^*$ alone to the case when the anchor variable is known to be 0. By conditioning on S, we can decompose $p_0(\mathbf{x}) = \text{pr}(Y = 1|\mathbf{x})$ as $\text{pr}(Y = 1|\mathbf{x}, S = 1)\text{pr}(S = 1|\mathbf{x}) + \text{pr}(Y = 1|\mathbf{x}, S = 0)\text{pr}(S = 0|\mathbf{x})$, where $\text{pr}(Y = 1|\mathbf{x}, S = 1) = 1$ and $\text{pr}(S = 1|\mathbf{x}) = a(\mathbf{x})$. In addition, as we have pointed out, $a(\mathbf{x})$ is known. Hence, $\text{pr}(Y = 1|\mathbf{x}, S = 0)$ can be estimated using $p(\mathbf{x}, \widehat{\boldsymbol{\beta}})$ by

$$\widehat{\text{pr}}(Y = 1|\mathbf{x}, S = 0) = \frac{p(\mathbf{x}, \widehat{\boldsymbol{\beta}}) - a(\mathbf{x})}{1 - a(\mathbf{x})}. \tag{4.5}$$

Then perform classification of Y^* at a new observation $(\mathbf{X}^*, S^* = 0)$ as $\widehat{\mathbf{Y}}^* = I\{\widehat{\text{pr}}(Y^* = 1|\mathbf{X}^*, S^* = 0) > 0.5\}$. In other words, we follow the same procedure in the algorithm described in Section 4.2.1, while replace the probability in the sixth step with (4.5).

4.3 Simulation experiments

We conduct simulation studies to assess the finite sample performance of our method. For comparison, we implemented our proposed optimal sampling in combination with optimal estimator of $\boldsymbol{\beta}$ (OPT), the estimation procedure proposed by Yin et al. (2022) (OSCA), the simple random sampling in combination with solving (4.3) (SRS), and the maximum likelihood estimator of the model $p(\mathbf{x}, \boldsymbol{\beta})$ under simple random sampling (MLE). Note that MLE mimics a naive method not considering the missingness probability, and SRS mimics the usual inverse probability weighting approach, both under the simple random sampling over the observations with $S = 0$. Summary statistics are also calculated under the true model (TRUE) to serve as a benchmark.

All results below are based on 1000 replicates with sample size $N = 10000$ for each replicate. In addition, we considered different constraints by setting $N_0\pi_0 = 200, 300,$ and 400 to further compare the performances under different limits of resources. For the entire replicates, we fixed $\widetilde{\boldsymbol{\beta}}$ estimated based on 400 observations drawn from the same population and independent of the N observations. Lastly, we set the average disease rate $E\{p_0(\mathbf{X})\} = 0.05$ and 0.2, and let the anchor sensitivity to be about 40%, i.e., $\text{pr}(S = 1 \mid Y = 1) = 0.4$. The detail of the simulation design in each experiment is individually described below.

We report the mean squared error (MSE) of the full dataset, calculated as $N^{-1}\sum_{i=1}^{N}\{y_i - p(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})\}^2$, and the mean squared error in the subset of the data with $S = 0$ (MSE_0),

calculated as $N_0^{-1}\sum_{i=1}^{N_0}\{y_i - \text{pr}(Y_i = 1 \mid \mathbf{x}_i, S_i = 0, \widehat{\boldsymbol{\beta}})\}^2$, where $\text{pr}(Y = 1|\mathbf{x}, S = 0)$ is estimated using (4.5) or is set to zero if (4.5) leads to a negative value. Although our estimator is designed to optimize MSE, we also report deviance (DEV), the full data specificity (TN) and sensitivity (TP) where the classification is $I\{p(\mathbf{x}, \widehat{\boldsymbol{\beta}}) > 0.5\}$, and the area under the ROC curve (AUC). The deviance is calculated based on N_ϵ observations whose predicted probabilities are within $[10^{-10}, 1 - 10^{-10}]$ and multiplied by N/N_ϵ.

4.3.1 Correct model specification

We first consider the situation when the classification model is correctly specified. The covariate $\mathbf{X}$ is of dimension 12 with X_1, X_5, X_9 generated independently from $N(0, 1)$, X_2, X_6, X_{10} from discrete uniform on $\{1, 2, 3, 4, 5\}$, X_3, X_7, X_{11} from Bernoulli with a success rate $p = 0.5$, and X_4, X_8, X_{12} from Bernoulli with $p = 0.1$ being categorical variables with low prevalence. We set the regression coefficient

$$\boldsymbol{\beta} = (0.2, 0.3, 0.4, -0.5, 0.8, 1.0, -1.2, 1.4, 1.7, -2.0, 2.3, 2.6)^{\mathrm{T}},$$

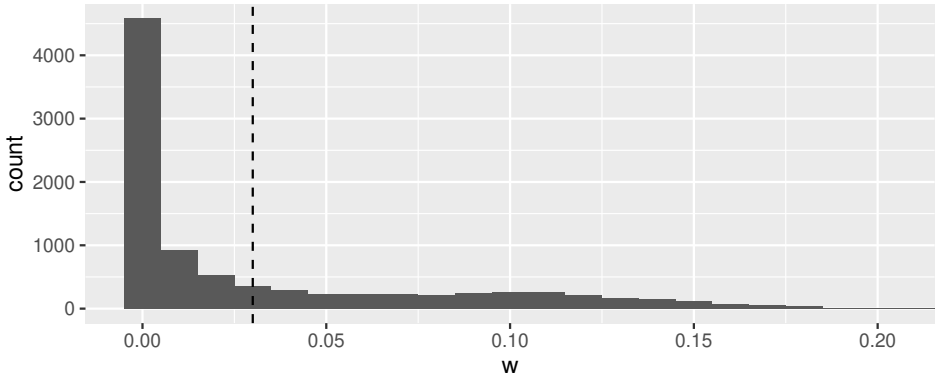
so that X_1, X_2, X_3, X_4 represent weak, X_5, X_6, X_7, X_8 moderate, and $X_9, X_{10}, X_{11}, X_{12}$ strong covariates.

Further, we draw Y_i from a logistic regresson model with $p_0(\mathbf{x}_i) = p(\mathbf{x}_i, \boldsymbol{\beta}) = \text{expit}(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$, where $\text{expit}(\eta) \equiv 1/(1 + e^{-\eta})$ and β_0 is chosen to yield the average disease rates 5% and 20% respectively. Finally, the anchor variable S_i is generated as the product of Y_i and a random variable distributed as Bernoulli with $p = \text{expit}(\gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_i)/\text{expit}(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$, where $\boldsymbol{\gamma} = \boldsymbol{\beta} + 0.1\text{sign}(\boldsymbol{\beta})$ and γ_0 is chosen to ensure that approximately 40% of the patients are detected by the anchor variable S. This implies that S given only $\mathbf{x}$ follows a Bernoulli distribution with $p = \text{expit}(\gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})$, thus we use $a(\mathbf{x}) = \text{expit}(\gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})$ to perform classification as described in Section 4.2.3.

To illustrate the estimation procedure of OPT, in Figure 4.1, we first plot the histogram of the sampling weights estimated by solving (4.1) in one simulation. As can be seen, OPT leads to very different sampling weights for different observations. This contrasts with the simple sampling weight, which is fixed as π_0. As described in Theorem 4.2.1 and Corollary 4.2.2, this sampling strategy will result in forming validation data that leads to a subsequent estimator that is optimal in MSE, which will be illustrated next.

In Table 4.1, we summarize how the implemented methods perform under various

Figure 4.1: Histogram of sampling weights $w_i, i = 1, \ldots, N_0$ in the simulation experiments with $N_0\pi_0 = 300$. Dotted: sampling weight under simple random sampling.



simulation settings. We can observe that OPT indeed shows the best performance in terms of both MSE and MSE_0. Compared to SRS, the OPT weight choice leads to improved performance of $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$, in line with Theorem 4.2.1. The optimal performance of OPT is more clear to see when $N_0\pi_0 = 200$, which suggests that the improvement of OPT over the estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ is more evident if the available resource is more restricted. We can also see that OPT provides the most similar result to TRUE in terms of both TN and TP among all the methods implemented, even though we do not expect OPT to show any superior performance in terms of sensitivity and specificity. In contrast, OSCA and MLE perform the worst in terms of MSE and MSE_0 even though the model is correctly specified, while they are better in DEV and AUC.

4.3.2 Wrong model specification

To further assess the performance where the classification model is misspecified, we generate the covariates $\mathbf{X}^* = (\mathbf{X}^{\mathrm{T}}, X_1^2, X_2X_6, X_7X_{11}, X_8X_{12})^{\mathrm{T}}$, where $\mathbf{X}$ is drawn from the same generation mechanism described in Section 4.3.1. In addition, we set the regression coefficient $\boldsymbol{\beta}^*$ to be a concatenation of $\boldsymbol{\beta}$ given in Section 4.3.1 and additional coefficients corresponding to higher order terms, i.e., $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^{\mathrm{T}}, -0.4, -0.6, 0.8, 1.0)^{\mathrm{T}}$. Then we generate Y_i from a Bernoulli distribution with the success rate $p_0(\mathbf{x}_i) = \text{expit}(\beta_0 + \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{x}_i^*)$ where β_0 is again chosen to achieve the disease rate 5% and 20%. To reflect the misspecified classification model scenario, the model $p(\mathbf{x}, \boldsymbol{\beta})$ is fitted without the higher order covariates, i.e., it only contains $\mathbf{x}$. Finally, the anchor variable S_i is generated as a multiplication of Y_i and a random variable distributed as Bernoulli with $\text{expit}(\gamma_0 + \boldsymbol{\gamma}^{*\mathrm{T}}\mathbf{x}_i^*)/\text{expit}(\beta_0 + \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{x}_i^*)$, so that $S|\mathbf{x}^*$ follows Bernoulli

with $p = \text{expit}(\gamma_0 + \boldsymbol{\gamma}^{*\mathrm{T}}\mathbf{x}^*)$. Here $\boldsymbol{\gamma}^* = \boldsymbol{\beta}^* + 0.1\text{sign}(\boldsymbol{\beta}^*)$ and γ_0 is chosen to ensure that 40% of patients are detected by the anchor variable S. In this setting, we use $a(\mathbf{x}^*) = \text{expit}(\gamma_0 + \boldsymbol{\gamma}^{*\mathrm{T}}\mathbf{x}^*)$ for classification as proposed in Section 4.2.3.

In Table 4.2, we summarize the performances of the implemented methods. Even though the model is misspecified hence the assumption of Corollary 4.2.2 is violated, the result still shows that our method OPT estimates the model with the smallest MSE and MSE_0, thus provides the best possible classification among all the methods. Compared to SRS, the MSE of $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ is better for OPT, which possibly suggests that the optimal sampling weight choice is robust to the model specification. Again, we can see that when $N_0\pi_0$ is smaller, the gain of OPT over the other methods is larger in terms of MSE. In addition, TN and TP results suggest that OPT has a similar level of specificity to the true model, though sensitivity is not as good, possibly caused by the wrong model choice. Lastly, as in Section 4.3.1, MSE and MSE_0 of OSCA and MLE are worse than that of OPT, while their DEV and AUC are competitive. Interestingly, OPT achieves the best AUC when the sampling size is 400.

4.3.3 Robustness to violation of positive-only assumption

To assess the sensitivity of our proposal to the positive-only assumption, we now consider the case where true data structure does not satisfy $P(Y = 1 \mid S = 1) = 1$. The covariates $\mathbf{X}_i$ and the response Y_i are drawn in the same manner as in Section 4.3.1. Then the anchor variable S_i is defined as $Y_i S_{1i} + (1 - Y_i)S_{0i}$, where S_{1i} follows the Bernoulli distribution with $p_i = \text{expit}(\gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_i)/\text{expit}(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ and S_{0i} is drawn from Bernoulli with probability $q = 0.025$. Here, β_0 is chosen to yield the disease rate 20%, and $\boldsymbol{\gamma} = \boldsymbol{\beta} + 0.1\text{sign}(\boldsymbol{\beta})$ where γ_0 is chosen to satisfy $P(S = 1|Y = 1) = 0.4$. This data generation procedure leads to 20% false alarm, i.e., $P(Y = 1 \mid S = 1) = 0.8$.

We implement all methods under the false assumption $P(Y = 1 \mid S = 1) = 1$. That is, we mistakenly label all the Y's to be 1 when the corresponding anchor variable $S = 1$. We then implement all estimation procedures using the logistic regression model of Y given $\mathbf{X}$, and measure the prediction performance using the true Y.

Table 4.3 summarizes the performances of implemented methods under violation of the positive only assumption. The results show that OPT has the best performance among the implemented methods in both MSE and MSE_0. Compared to the results in Table 4.1, both OPT and SRS are robust to the mislabeling, while the performances of two likelihood-based approaches, OSCA and MLE, severely deteriorated due to the mislabeling in terms of MSE and MSE_0. We find that this is due to the weight difference

between observations corresponding to $S = 1$ and $S = 0$. In OPT and SRS, the weights assigned to observations with $S = 1$ are much larger than those assigned to observations with $S = 0$, hence the mislabeled observations, which must have $S = 1$, have almost no impact on the final estimation. Interestingly, OPT also achieves the best DEV and AUC when the sampling size is above 300, even though its goal is to minimize the mean squared error.

4.4 Data application

To further illustrate the numerical performance of our method, we apply it to the motivating example, where we try to identify patients with depression symptoms at their hospital admissions using their EHRs. The data was part of the Medical Information Mart for Intensive Care III database version 1.4 (MIMIC-III v1.4) (Johnson et al. 2016), a freely-available database on health-related information of over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. We used a subset of MIMIC-III that consists of 767 admission entries from 415 frequent ICU patients. The same set of patients and admission entries has been analyzed in Gehrmann et al. (2018). We emphasize that our method in practice can be conducted on bigger databases, as demonstrated in Section 4.3 with larger sample size. Sampling weights in (4.1) can be obtained by off-the-shelf algorithms for constrained optimization, and have a closed form under suitable conditions as given in Remark 4.2.1.

Our goal is to classify admission entries into two groups, whether patients present any depression symptoms during hospital admissions. Hence we defined a binary label $Y = 1$ if a patient had any depression symptoms. In addition, for the anchor variable S, we extracted the ICD-9 diagnosis codes to determine whether a patient was diagnosed as having depression. In addition, a panel of medical experts in Gehrmann et al. (2018) have conducted cross-validated chart reviews of those entries, and identified 30.7% of the entries presenting depression symptoms (i.e. $Y = 1$) . Among these patients with $Y = 1$, 24.4% were diagnosed as having depression (i.s. $S = 1$), hence anchor-positive. There were 5 observations with $S = 1$ but $Y = 0$, which violates the positive-only assumption. These were deleted from the dataset in our analysis. We then used additional information from the records as covariates, which includes gender, age, ethnicity, marital status, insurance type, admission type, previous length of stays in hospital, number of procedures done on patients during the admission in each category (evaluation and man-

agement, medicine, radiology, surgery), and the number of prescribed medications for depressions (Amitriptyline, Clomipramine, Desipramine, Doxepin, Duloxetine, Escitalopram, Fluoxetine, Imipramine, Levomilnacipran, Nortriptyline, Paroxetine, Sertraline, Venlafaxine).

Using the above information, we implemented our method (OPT) as if the labels Y were not observed. For comparison, we also implemented the procedure proposed by Yin et al. (2022) (OSCA), the simple random sampling in combination with solving (4.3) (SRS), and the maximum likelihood estimator under the simple random sampling (MLE). As a benchmark, the minimizer of MSE using the full dataset (FULL) was also implemented. For all methods, we adopted the logistic regression model with all the covariates mentioned above and age squared. No additional higher order terms or interaction terms of the covariates are included. For our method, we used $m = 100$ and 200 observations sampled from the full dataset to obtain the initial estimate $\widetilde{\boldsymbol{\beta}}$. We then implemented all the methods on the remaining data.

To compare the performance of the methods, we report the prediction mean squared error (MSE) and the mean squared error on the observations with $S = 0$ (MSE_0). Although our method is designed to minimize MSE, to assess the performance of our method under other criteria, deviance (DEV), the specificity (TN), the sensitivity (TP), and the area under the ROC curve (AUC) were also reported. The criteria are calculated based on the data excluding the m observations used for initial estimation. Details on how the measurements were calculated can be found in Section 4.3, and the anchor positive probability $a(\mathbf{x})$ was fitted by the logistic regression on S given the same covariates. The summaries are based on 1000 replicates, and we consider different sampling sizes $N_0\pi_0 = 50, 100, 200, 300$.

Figure 4.2 illustrates the sampling weights estimated by our method when $m = 100$ and $N_0\pi_0 = 200$. The figure shows that our procedure yields various sampling weights for different observations, in contrast to the simple random sampling, which would assign the same sampling weight 0.326 to every observation. This weighting strategy indeed improved MSE and MSE_0 of the following estimates, as can be seen from Figures 4.3 and 4.4. These two figures show the boxplots of MSE and MSE_0 under each allowed subsampling size. We can see that our method indeed results in estimators with smaller MSE and MSE_0 than other methods in general, regardless of the subsampling sizes.

For more accurate comparison, we summarized the average and the standard deviation of each criterion in Table 4.4. From the table we can see that OPT indeed has the smallest MSE and MSE_0 among all methods under reasonably large subsample sizes.

Figure 4.2: Histogram of sampling weights $w_i, i = 1, \ldots, N_0$ in the data example with $m = 100$ and $N_0\pi_0 = 200$. Dotted: sampling weight under simple random sampling.

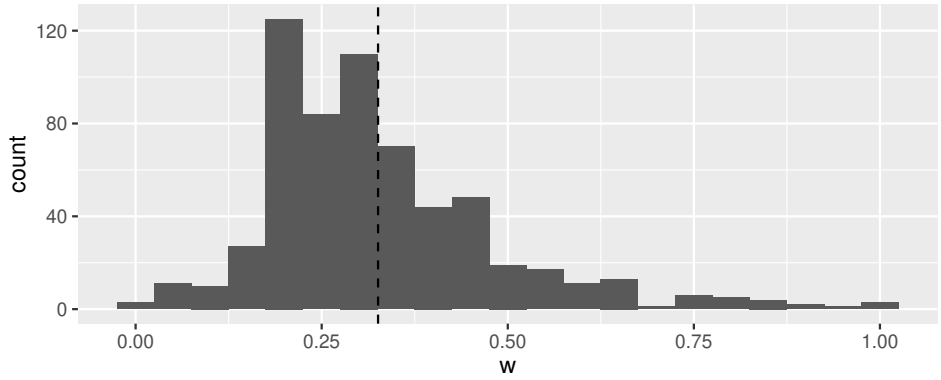


Figure 4.3: Boxplots of MSE in the data example with $m = 100$. Dashed: MSE based on the full dataset.

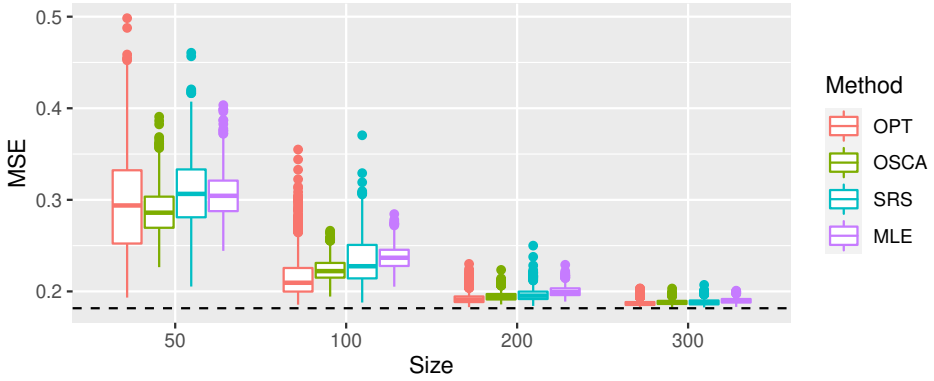
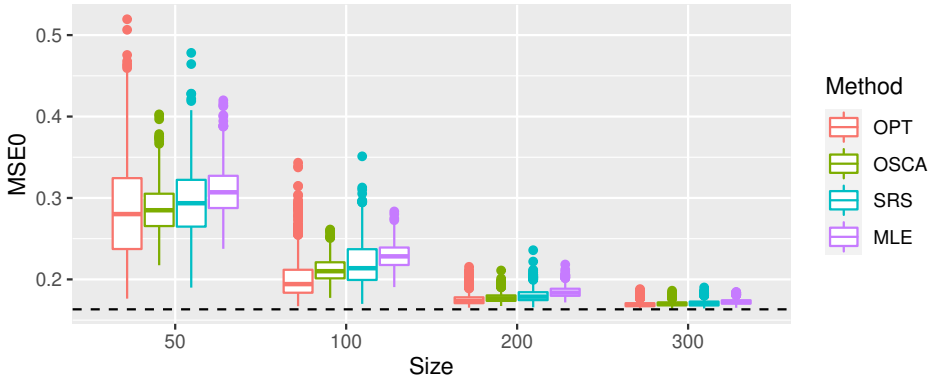


Figure 4.4: Boxplots of MSE_0 in the data example with $m = 100$. Dashed: MSE_0 based on the full dataset.



Indeed, when the subsample size is very small such as $N_0\pi_0 = 50$, every method performs worse than a pure random guess, which yields MSE 0.25. Also, it can be seen from Table 4.4 that under smaller subsample size, the gain from our method is more prominent. This suggests that our method is especially valuable when resource is very limited. On the other hand, in terms of specificity and sensitivity, OPT provides the most similar results to the benchmark results by FULL in all situations. In terms of DEV and AUC, although OPT performs less satisfactory when the sampling size is very small (50), it is very similar to the best performance when the size increases to 100, and almost has the best performance when it further increases to 200 and 300. Lastly, the pilot size m has almost no impact on the competitiveness of OPT, suggesting that a consistent $\widetilde{\boldsymbol{\beta}}$ is sufficient as shown in Corollary 4.2.2.

4.5 Discussion

We have proposed an optimal subsampling procedure in terms of prediction MSE. The procedure does not require a correct prediction model, and can be conducted when the allowed subsampling size is restricted. Although the detailed derivation aims to minimize MSE, the idea and general approach are applicable to any other criteria. Under different criterions, the optimal sampling weights are also likely different. Theoretical derivations are also criterion dependent, and may be challenging especially when the chosen criterion is non-smooth or discontinuous. Our derivation is conducted under a single prediction model. If several models are of interest, we recommend to construct an ensemble model which includes these models as special cases to increase model flexibility and carry out the proposed procedure.

We also point out that our method can be directly applied to the simpler case where no anchor variable S is available. To do so, we simply treat the no anchor variable case as a special case with an anchor variable, but all anchor variables are zero, i.e., $S_i = 0$ for all $i = 1, \ldots, N$, and $N_0 = N$. Then our method is immediately applicable. Because our method is appliable regardless whether there is a genuine positive-only anchor variable, in the case when we are given an anchor variable that is not necessarily positive-only, we recommend dismissing it or treating it as a covariate to remain in our framework, even though the method appears to be robust in our numerical experiements. Throughout the article, we have considered a binary response Y. However, our analysis procedure also can be applied to the case when the response is continuous. On the other hand, in the case of a binary response, Criteria other than MSE are also of interest, such as

misclassification rate. We believe each criterion requires a specifically tailored procedure and poses its own challenge in establishing the optimality.

4.6 Acknowledgments

This chapter is based upon work supported by the National Science Foundation and the National Institute of Health. Any opinions, findings, and conclusions or recommendations expressed in this chapter are those of the authors, and do not necessarily reflect the views of the National Science Foundation and the National Institute of Health.

Table 4.1: Simulation results under correct model specification. Mean and standard deviation (in parentheses) of the corresponding summaries.

$E\{p_0(\mathbf{X})\} = .05$

Size	Method	MSE	MSE_0	DEV	TN	TP	AUC
	TRUE	.025 (.001)	.021 (.001)	1685.168	.991	.510	.970
200	OPT	.043 (.009)	.039 (.009)	36826.695	.975	.549	.947
	OSCA	.067 (.012)	.066 (.013)	4602.412	.913	.853	.958
	SRS	.050 (.008)	.046 (.009)	58735.478	.967	.586	.942
	MLE	.059 (.007)	.059 (.007)	4011.114	.922	.869	.966
300	OPT	.039 (.008)	.035 (.007)	31249.481	.979	.539	.951
	OSCA	.051 (.007)	.050 (.008)	3363.587	.936	.821	.963
	SRS	.044 (.006)	.040 (.007)	47680.739	.974	.567	.947
	MLE	.049 (.005)	.048 (.005)	3267.891	.937	.840	.967
400	OPT	.036 (.006)	.032 (.006)	23624.540	.981	.542	.955
	OSCA	.043 (.005)	.041 (.006)	2809.119	.948	.793	.965
	SRS	.040 (.006)	.036 (.006)	38676.695	.978	.549	.951
	MLE	.044 (.004)	.042 (.004)	2870.093	.946	.817	.967

$E\{p_0(\mathbf{X})\} = .20$

Size	Method	MSE	MSE_0	DEV	TN	TP	AUC
	TRUE	.063 (.002)	.063 (.002)	4094.214	.957	.732	.958
200	OPT	.084 (.015)	.086 (.015)	24863.828	.942	.721	.945
	OSCA	.131 (.030)	.140 (.027)	9129.010	.805	.898	.930
	SRS	.098 (.014)	.100 (.014)	51564.708	.935	.720	.940
	MLE	.103 (.007)	.109 (.008)	6830.639	.846	.917	.954
300	OPT	.075 (.010)	.077 (.011)	12224.434	.946	.725	.949
	OSCA	.104 (.011)	.110 (.012)	6775.164	.849	.889	.945
	SRS	.081 (.012)	.083 (.013)	20707.036	.945	.721	.947
	MLE	.091 (.005)	.096 (.006)	5959.228	.869	.899	.955
400	OPT	.070 (.006)	.072 (.007)	6414.891	.950	.725	.951
	OSCA	.090 (.006)	.094 (.007)	5776.053	.874	.876	.950
	SRS	.072 (.007)	.074 (.008)	7771.140	.949	.724	.951
	MLE	.084 (.004)	.088 (.005)	5491.735	.883	.885	.956

Table 4.2: Simulation results under model misspecification. Mean and standard deviation (in parentheses) of the corresponding summaries.

$E\{p_0(\mathbf{X})\} = .05$

Size	Method	MSE	MSE_0	DEV	TN	TP	AUC
	TRUE	.023 (.001)	.020 (.001)	1575.514	.991	.548	.975
200	OPT	.046 (.009)	.041 (.009)	47652.551	.974	.525	.937
	OSCA	.073 (.011)	.073 (.011)	4930.425	.903	.858	.956
	SRS	.051 (.009)	.046 (.009)	60229.612	.968	.552	.932
	MLE	.067 (.008)	.067 (.008)	4483.136	.911	.864	.960
300	OPT	.042 (.006)	.037 (.006)	43313.776	.978	.511	.941
	OSCA	.056 (.007)	.055 (.008)	3682.277	.928	.820	.959
	SRS	.046 (.007)	.041 (.007)	53151.853	.974	.530	.938
	MLE	.055 (.005)	.055 (.006)	3640.138	.928	.831	.961
400	OPT	.039 (.006)	.035 (.006)	35424.364	.980	.502	.944
	OSCA	.048 (.006)	.046 (.006)	3109.903	.941	.787	.961
	SRS	.043 (.006)	.038 (.006)	45619.777	.978	.515	.941
	MLE	.049 (.004)	.048 (.005)	3204.222	.939	.805	.962

$E\{p_0(\mathbf{X})\} = .20$

Size	Method	MSE	MSE_0	DEV	TN	TP	AUC
	TRUE	.058 (.002)	.059 (.002)	3750.032	.959	.761	.965
200	OPT	.089 (.015)	.093 (.013)	17583.095	.938	.685	.933
	OSCA	.142 (.014)	.153 (.014)	9778.256	.784	.914	.932
	SRS	.105 (.017)	.107 (.016)	44193.354	.930	.687	.928
	MLE	.119 (.008)	.128 (.008)	7994.974	.819	.916	.945
300	OPT	.080 (.008)	.085 (.007)	7823.384	.943	.690	.939
	OSCA	.115 (.008)	.124 (.009)	7542.984	.827	.897	.940
	SRS	.086 (.012)	.090 (.011)	14564.369	.940	.689	.936
	MLE	.106 (.006)	.113 (.006)	6935.105	.844	.896	.946
400	OPT	.077 (.005)	.082 (.005)	5519.307	.946	.690	.941
	OSCA	.101 (.006)	.108 (.007)	6506.819	.852	.879	.943
	SRS	.078 (.006)	.083 (.006)	6339.479	.944	.691	.940
	MLE	.098 (.005)	.104 (.005)	6357.604	.861	.880	.946

Table 4.3: Simulation results under violation of the positive-only assumption. Mean and standard deviation (in parentheses) of the corresponding summaries.

Size	Method	MSE	MSE_0	DEV	TN	TP	AUC
	TRUE	.063 (.002)	.065 (.002)	4094.214	.957	.732	.958
200	OPT	.082 (.015)	.086 (.016)	19821.625	.938	.734	.945
	OSCA	.355 (.017)	.374 (.019)	18803.531	.202	1.000	.923
	SRS	.095 (.016)	.099 (.016)	41650.395	.932	.729	.940
	MLE	.365 (.016)	.385 (.017)	20007.290	.269	1.000	.942
300	OPT	.073 (.009)	.076 (.009)	8205.711	.944	.736	.950
	OSCA	.269 (.014)	.280 (.016)	14712.217	.402	.998	.939
	SRS	.078 (.012)	.081 (.012)	15439.427	.942	.733	.948
	MLE	.288 (.013)	.301 (.015)	15885.078	.408	.998	.947
400	OPT	.069 (.005)	.072 (.005)	5172.247	.947	.738	.952
	OSCA	.218 (.012)	.225 (.013)	12387.180	.538	.993	.946
	SRS	.071 (.006)	.073 (.007)	6453.707	.946	.736	.951
	MLE	.239 (.011)	.248 (.012)	13411.293	.510	.995	.950

Table 4.4: Data example results. Mean and standard deviation (in parentheses) of the corresponding summaries.

$m = 100$

Size	Method	MSE	MSE_0	DEV	TN	TP	AUC
	FULL	.182 (.000)	.163 (.000)	739.200	.926	.300	.714
50	OPT	.295 (.056)	.284 (.061)	3511.233	.772	.431	.613
	OSCA	.288 (.026)	.287 (.030)	1521.553	.520	.656	.624
	SRS	.309 (.039)	.297 (.042)	4369.568	.775	.435	.609
	MLE	.306 (.025)	.309 (.029)	1629.172	.475	.686	.625
100	OPT	.217 (.024)	.202 (.026)	1173.383	.860	.385	.661
	OSCA	.224 (.013)	.212 (.015)	1069.759	.763	.500	.659
	SRS	.234 (.026)	.219 (.026)	1658.405	.841	.410	.651
	MLE	.238 (.014)	.229 (.016)	1128.305	.704	.554	.656
200	OPT	.192 (.006)	.175 (.007)	802.642	.907	.350	.688
	OSCA	.195 (.005)	.178 (.006)	826.873	.887	.374	.686
	SRS	.196 (.007)	.180 (.008)	860.564	.896	.368	.686
	MLE	.200 (.006)	.185 (.007)	847.391	.847	.441	.682
300	OPT	.187 (.003)	.169 (.003)	768.319	.921	.330	.696
	OSCA	.188 (.003)	.170 (.003)	768.789	.919	.319	.698
	SRS	.188 (.003)	.171 (.004)	783.098	.914	.346	.699
	MLE	.190 (.003)	.172 (.003)	776.052	.899	.379	.695

$m = 200$

Size	Method	MSE	MSE_0	DEV	TN	TP	AUC
	FULL	.170 (.000)	.153 (.000)	590.416	.929	.361	.743
50	OPT	.282 (.054)	.271 (.058)	3140.056	.788	.450	.635
	OSCA	.262 (.026)	.261 (.030)	1186.123	.618	.637	.665
	SRS	.291 (.036)	.280 (.039)	3691.941	.792	.454	.632
	MLE	.276 (.025)	.279 (.029)	1272.175	.576	.678	.669
100	OPT	.204 (.025)	.190 (.027)	980.448	.872	.411	.688
	OSCA	.204 (.012)	.192 (.015)	823.034	.814	.492	.696
	SRS	.222 (.024)	.209 (.025)	1460.681	.854	.441	.676
	MLE	.215 (.013)	.206 (.015)	867.045	.765	.558	.698
200	OPT	.178 (.005)	.162 (.006)	634.448	.911	.390	.723
	OSCA	.181 (.004)	.164 (.005)	656.411	.903	.386	.720
	SRS	.184 (.007)	.169 (.008)	704.734	.903	.406	.713
	MLE	.184 (.004)	.168 (.005)	668.724	.873	.455	.722
300	OPT	.173 (.002)	.156 (.002)	604.041	.922	.377	.734
	OSCA	.175 (.002)	.157 (.002)	613.060	.923	.347	.733
	SRS	.176 (.002)	.159 (.003)	623.377	.918	.392	.729
	MLE	.176 (.002)	.158 (.002)	616.538	.907	.398	.733

Chapter 5
Doubly flexible estimation under label shift

5.1 Introduction

In studies ranging from clinical medicine to policy research, there often exist data and information from a population $\mathcal{P}$, while the quantity of interest is defined on a particular target $\mathcal{Q}$, relevant but different from $\mathcal{P}$. For instance, in a clinical trial setting, physicians may be left interpreting evidence from a randomized controlled trial consisting of patients who have demographics and comorbidities that are quite different from those of their own patients (population $\mathcal{Q}$). As another example, to build a predictive model on pneumonia outbreak for the flu season (population $\mathcal{Q}$), researchers might find a similar model during the non-flu season relevant and useful. In these scenarios, there is a discrepancy between the distributions of $\mathcal{P}$ and $\mathcal{Q}$, termed distribution shift throughout. Distribution shift can also refer to the fact that the distribution of the training sample is different from that of the testing sample, in the evaluation of a learning algorithm.

In all of these situations, it is of vital interest to propose methods that can appropriately leverage the information from $\mathcal{P}$ into the statistical tasks for $\mathcal{Q}$. Our methodology will use the information from both outcome (output, response, label) Y and covariate (input, predictor, feature) $\mathbf{X}$ in population $\mathcal{P}$ as well as covariate $\mathbf{X}$ in population $\mathcal{Q}$. This setting is also named unsupervised domain adaptation (Quinonero-Candela et al. 2008, Moreno-Torres et al. 2012, Kouw & Loog 2019).

Without any assumptions on the nature of shift, it is certainly impossible to leverage information between two heterogeneous populations. Two major types of distribution shifts have been defined in the literature. The first is called covariate shift where the shift happens between the marginal distributions of $\mathbf{X}$ while the conditional distribution

of Y given $\mathbf{X}$ does not change; i.e., $p_{\mathbf{X}}(\mathbf{x}) \neq q_{\mathbf{X}}(\mathbf{x})$ and $p_{Y|\mathbf{X}}(y, \mathbf{x}) = q_{Y|\mathbf{X}}(y, \mathbf{x})$. The difference between $\mathcal{P}$ and $\mathcal{Q}$ can be summarized as a density ratio $q_{\mathbf{X}}(\mathbf{x})/p_{\mathbf{X}}(\mathbf{x})$, which is, fortunately, estimable since covariate $\mathbf{X}$ is available from both populations. Covariate shift aligns with the causal learning setting (Schölkopf et al. 2012) where $\mathbf{X}$ is the cause and Y is the effect. Covariate shift has attracted a great deal of attention and has been investigated in many literatures, such as Shimodaira (2000), Huang et al. (2006), Sugiyama et al. (2008), Gretton et al. (2009), Sugiyama & Kawanabe (2012), Kpotufe & Martinet (2021) and the references therein.

The second type, which is the focus of this paper, is named label shift, because it assumes that the shift is induced by the marginal distributions of Y while the process generating $\mathbf{X}$ given Y is identical in both populations. Formally, it assumes

$$p_Y(y) \neq q_Y(y), \text{ and } p_{\mathbf{X}|Y}(\mathbf{x}, y) = q_{\mathbf{X}|Y}(\mathbf{x}, y) \equiv g(\mathbf{x}, y).$$

Label shift is also called prior probability shift (Storkey 2009, Tasche 2017), target shift (Zhang et al. 2013, Nguyen et al. 2016), or class prior change (Du Plessis & Sugiyama 2014, Iyer et al. 2014). Label shift aligns with the anticausal learning setting in which the outcome/label Y causes the covariate/feature $\mathbf{X}$; for example, diseases cause symptoms or objects cause sensory observations. Consider the situation that one fits a model to predict whether a patient has pneumonia based on observed symptoms, and that this model predicts reliably when deployed in the clinic during the non-flu season. When the flu season starts, there is a sudden surge of pneumonia cases and the probability of patients developing pneumonia given that they show symptoms rises, while the mechanism of showing symptoms of pneumonia is rather stable. Label shift also exists in many computer vision applications, such as predicting object locations and directions, and human poses; see Martinez et al. (2017), Yang et al. (2018), Guo et al. (2020).

In the label shift framework, one fundamental problem (Garg et al. 2020) is determining whether the shift has occurred and estimating the label distribution $q_Y(y)$, or equivalently, assessing the density ratio $q_Y(y)/p_Y(y) \equiv \rho(y)$. In contrast to estimating the density ratio $q_{\mathbf{X}}(\mathbf{x})/p_{\mathbf{X}}(\mathbf{x})$ under covariate shift, estimating $\rho(y)$ is a daunting task due to the absence of the Y observations in population $\mathcal{Q}$. Works in the label shift framework are mainly limited to the classification problems in the machine learning literature. Saerens et al. (2002) proposed a simple Expectation-Maximization (EM) (Dempster et al. 1977) procedure, named maximum likelihood label shift (MLLS), to estimate $q_Y(y)$ assuming access to a classifier that outputs the true conditional proba-

bilities of the population $\mathcal{P}$, $p_{Y|\mathbf{X}}(y, \mathbf{x})$. Later on, Chan & Ng (2005) proposed an EM algorithm that requires the estimation of $g(\mathbf{x}, y)$, which is unfortunately difficult for high-dimensional $\mathbf{X}$ and moreover, it does not apply to regression problems. Alternatively, Lipton et al. (2018) and Azizzadenesheli et al. (2019) proposed moment-matching based estimators, named black box shift learning (BBSL) and regularized learning under label shift (RLLS), that make use of the invertible confusion matrix of a classifier learned from population $\mathcal{P}$. The connection and comparison of these two lines of research, either empirical or theoretical, remain unclear. To our best knowledge, neither BBSL nor RLLS has been benchmarked against EM. Alexandari et al. (2020) showed that, in combination with a calibration named bias-corrected temperature scaling, MLLS outperforms BBSL and RLLS empirically; whereas MLLS underperforms BBSL when applied naively. Under label shift, Maity et al. (2022) also studied the minimax rate of convergence for nonparametric classification.

For continuous Y in regression problems, estimating $q_Y(y)$ becomes the problem of estimating a function instead of a finite number of parameters. Not surprisingly, its literature is quite scarce. Zhang et al. (2013) proposed a nonparametric method to estimate the density ratio by kernel mean matching of distributions. However, this approach does not scale to large data as the computational cost is quadratic in the sample size. Nguyen et al. (2016) considered continuous label shift adaptation and studied an importance weight estimator, but their approach relies on a parametric model for $p_{Y|\mathbf{X}}(y, \mathbf{x})$ hence can only be applied in supervised learning.

In this paper, we take a completely distinct approach from all of the existing literature. Different from the current majority, our methodology is devised to accommodate both classification and regression. Whether the outcome Y is discrete or continuous is not essential in our proposal. We directly estimate a characteristic of the population $\mathcal{Q}$. Specifically, we estimate the parameter $\boldsymbol{\theta}$ such that $\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$ where $\mathbf{U}(\cdot)$ is a user specified function and $\mathrm{E}_q(\cdot)$ stands for the expectation with respect to $q_Y(y)g(\mathbf{x}, y)$ or equivalently to $q_{Y|\mathbf{X}}(y, \mathbf{x})q_{\mathbf{X}}(\mathbf{x})$. This is a general framework, including estimating the mean of Y or the t-th quantile of Y as special cases. According to how the nuisance components are estimated, detailed in the next three paragraphs, we propose various estimators for $\boldsymbol{\theta}$, and develop large sample theory for these estimators to quantify the estimation uncertainties and to conduct statistical inference.

To estimate $\boldsymbol{\theta}$, three nuisance components are involved. First and foremost is the density ratio $\rho(y)$, which is almost infeasible to estimate based on the observed data due to the lack of Y-observations in population $\mathcal{Q}$. Our intention is to bypass the challenging

task of estimating $\rho(y)$. This turns out achievable through careful manipulation of other components of the influence function. In fact, a unique feature of our work is that, we do not need to estimate $\rho(y)$ throughout the estimation procedure. Instead, only a working model, denoted as $\rho^*(y)$, is needed.

The second one is $p_{Y|\mathbf{X}}(y, \mathbf{x})$, or some dependent quantities such as $\mathrm{E}_p(\cdot \mid \mathbf{x})$. In contrast to $\rho(y)$, estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$ is blessed with the observed data in population $\mathcal{P}$. Indeed, we can use off-the-shelf machine learning methods or nonparametric regression methods to obtain the corresponding estimator $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$. Nonetheless, we can also choose to give up estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$ even though we can do it. This means that we can misspecify the conditional distribution $p_{Y|\mathbf{X}}(y, \mathbf{x})$ while we also misspecify the density ratio $\rho(y)$. We call such an estimator $\widehat{\boldsymbol{\theta}}$ doubly flexible—the working density ratio model $\rho^*(y)$ is flexible, so is the working conditional distribution model $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$. Note that our superscripts here are different: superscript * stands for the working model of the density ratio whereas * is for the conditional distribution model. This doubly flexible property is much more favorable than the classic "doubly robust" in the literature. The standard double robustness means that one can misspecify either one of two models but not both, while here, we can misspecify both models. As an alternative, if one chooses to estimate $\mathrm{E}_p(\cdot \mid \mathbf{x})$, say, $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$, we name the corresponding estimator $\widetilde{\boldsymbol{\theta}}$ singly flexible—only flexible in working model $\rho^*(y)$.

The third nuisance is the conditional density function $g(\mathbf{x}, y)$, whose estimation might be subject to the curse of dimensionality. Fortunately, in our estimation procedure, $g(\mathbf{x}, y)$ only affects quantities of the form $\mathrm{E}(\cdot \mid y)$, which are one dimensional regression problems hence can be easily solved via the most basic nonparametric regression procedure such as the Nadaraya-Watson estimation.

The remaining of the paper is structured as follows. In Section 5.2, we first outline our strategy of how to incorporate samples from two heterogeneous populations. The proposed doubly flexible estimator is presented in Section 5.3, and the alternative singly flexible estimator is contained in Section 5.4. For easier understanding and improved readability, we present both the methodology and the theory for a special parameter $\theta = \mathrm{E}_q(Y)$ in the main text, while defer the general results for $\boldsymbol{\theta}$ such that $\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$ to the Supplement. Section 5.5 contains empirical results for extensive simulation studies. We present an application to the MIMIC-III database in Section 5.6. The paper is concluded with discussions in Section 5.7. All the technical details are also included in the Supplement.

5.2 Model structure

We consider independent and identically distributed (iid) observations $\{Y_i, \mathbf{X}_i\}, i = 1, \ldots, n_1$ from population $\mathcal{P}$, and iid observations $\mathbf{X}_j, j = n_1 + 1, \ldots, n_1 + n_0 = n$ from population $\mathcal{Q}$. To use the information in population $\mathcal{P}$ under label shift, we stack the two random samples together and assemble a new data set of size n, which represents a random sample for an imaginary population consisting of $100\pi\%$ population $\mathcal{P}$ members and $100(1-\pi)\%$ population $\mathcal{Q}$ members. Here we define $\pi \equiv n_1/n$. Throughout our derivation, other than $\mathrm{E}_p(\cdot)$ and $\mathrm{E}_q(\cdot)$, we also compute $\mathrm{E}(\cdot)$ that is with respect to this imaginary population; however, this imaginary population is only used as an intermediate tool to leverage information from two heterogeneous populations under label shift. Our final conclusion will only be made for the target population $\mathcal{Q}$.

For convenience, we introduce a binary indicator R in this stacked random sample, where $R = 1$ means the subject is from population $\mathcal{P}$ and $R = 0$ population $\mathcal{Q}$. Thus, the likelihood of one observation from the stacked random sample is

$$\{g(\mathbf{x}, y)p_Y(y)\}^r \left\{ \int g(\mathbf{x}, y)q_Y(y)dy \right\}^{1-r} \pi^r(1-\pi)^{1-r} \tag{5.1}$$

$$= \{g(\mathbf{x}, y)p_Y(y)\}^r \left\{ \int g(\mathbf{x}, y)\rho(y)p_Y(y)dy \right\}^{1-r} \pi^r(1-\pi)^{1-r}. \tag{5.2}$$

Although $g(\mathbf{x}, y)$ and $p_Y(y)$ can be identified from (5.1), unfortunately $q_Y(y)$ may not. Below is a simple example illustrating the possible nonidentifiability of $q_Y(y)$.

Example 5.2.1. *Consider a discrete Y with three supporting values $0, 1, 2$ and a discrete X with two supporting values $0, 1$. In both populations $\mathcal{P}$ and $\mathcal{Q}$, $g(x, y)$ is given as*

$$pr(X = 0 \mid Y = 0) = 1/5, pr(X = 0 \mid Y = 1) = 1/8, \text{ and } pr(X = 0 \mid Y = 2) = 2/3.$$

In population $\mathcal{P}$, the marginal distribution of Y, $p_Y(y)$, is given as

$$pr(Y = 0) = 5/16, pr(Y = 1) = 1/2, \text{ and } pr(Y = 2) = 3/16.$$

In population $\mathcal{Q}$, the marginal distribution of Y, $q_Y(y)$, is given as

$$pr(Y = 0) = \frac{5(25 - 416t)}{336}, pr(Y = 1) = \frac{32t - 1}{6}, \text{ and } pr(Y = 2) = \frac{89 + 96t}{112},$$

where $t \in (1/32, 25/336)$. Clearly this satisfies the label shift assumption, and the

marginal distribution of X in population $\mathcal{Q}$ is identifiable since $pr(X = 0) = 7/12$ is free of t; however, the marginal distribution of Y in population $\mathcal{Q}$, $q_Y(y)$, is not identifiable.

The following result demonstrates that the completeness condition on $p_{Y|\mathbf{X}}(y, \mathbf{x})$ would ensure the identifiability. Its proof is contained in Section D.1 of the Supplement.

Lemma 5.2.1. *If the conditional pdf/pmf $p_{Y|\mathbf{X}}(y, \mathbf{x})$ of population $\mathcal{P}$ satisfies the completeness condition in the sense that, for any function $h(Y)$ with finite mean, $\mathrm{E}_p\{h(Y) \mid \mathbf{X}\} = \int h(y)p_{Y|\mathbf{X}}(y, \mathbf{x})dy = 0$ implies $h(Y) = 0$ almost surely, then all the unknown components in (5.2), i.e., $g(\mathbf{x}, y)$, $p_Y(y)$ and $\rho(y)$, are identifiable. Subsequently, $q_Y(y)$ is also identifiable.*

The completeness condition in Lemma 5.2.1 is mild and has been widely assumed in instrumental variables, measurement error models, and econometrics; see, e.g, Newey & Powell (2003), d'Haultfoeuille (2011), Hu & Shiu (2018). Because the condition is imposed on population $\mathcal{P}$ and we have random observations $(\mathbf{X}_i, Y_i)$'s from population $\mathcal{P}$, it can be examined and verified in empirical studies. One can easily check that many commonly-used distributions such as exponential families satisfy the completeness condition. In particular, if the outcome Y is discrete with finitely many supporting values, Newey & Powell (2003) pointed out that the completeness condition only means that the covariate $\mathbf{X}$ has a support whose cardinality is no smaller than that of Y.

In this paper, we focus on estimating a characteristic of population $\mathcal{Q}$. For better clarity, we will present the results for $\theta = \mathrm{E}_q(Y) = \int yg(\mathbf{x}, y)q_Y(y)d\mathbf{x}dy$ in the main text, then generalize the results to $\boldsymbol{\theta}$ that satisfies $\mathrm{E}_q\{\mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})\} = \mathbf{0}$ in the Supplement. The main challenge in estimating θ is caused by the lack of knowledge and data on $q_Y(y)$, or equivalently, $\rho(y)$. Nevertheless, we will construct an estimator that bypasses the difficulty of assessing $\rho(y)$. We will show that we only need a working model of $\rho(y)$, denoted as $\rho^*(y)$, that can be flexible. Furthermore, we find that our procedure can also simultaneously avoid estimating $p_{Y|\mathbf{X}}(y, \mathbf{x})$, in that we can insert a possibly misspecified working model $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$. Thus, our procedure is flexible with respect to both $\rho(y)$ and $p_{Y|\mathbf{X}}(y, \mathbf{x})$—doubly flexible. This is a property different from the classic "double robustness" which means that one can only misspecify one of two models but not both. In contrast, here, we can misspecify both.

5.3 Proposed doubly flexible estimation for $\theta = \mathrm{E}_q(Y)$

If the density ratio function $\rho(y)$ were known, an intuitive estimator of $\theta = \mathrm{E}_q(Y)$ can be created by noticing the relation $\theta = \mathrm{E}_q(Y) = \mathrm{E}_p\{\rho(Y)Y\} = \mathrm{E}\{R\rho(Y)Y\}/\pi$; that is,

$$\breve{\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi}\rho(y_i)y_i. \tag{5.3}$$

We call this estimator shift-dependent since it requires the correct specification of $\rho(y)$. Clearly, if a working model $\rho^*(y)$ is adopted, the corresponding estimator $\breve{\theta}^*$ is likely biased.

5.3.1 General approach

The creation of an estimator that is not solely shift-dependent is possible. To motivate our proposed estimator, we first make some simple observations via balancing the samples from populations $\mathcal{P}$ and $\mathcal{Q}$. Recognizing the relation between $\mathrm{E}_p(\cdot)$, $\mathrm{E}_q(\cdot)$ and $\mathrm{E}(\cdot)$, the balancing of Y is

$$\mathrm{E}\left\{\frac{R}{\pi}\rho(Y)Y\right\} = \mathrm{E}_p\{\rho(Y)Y\} = \mathrm{E}_q(Y) = \mathrm{E}\left(\frac{1-R}{1-\pi}\theta\right).$$

Further, replacing the variable Y above by an arbitrary function of $\mathbf{X}$, we obtain another balancing function

$$\mathrm{E}\left\{\frac{R}{\pi}\rho(Y)b(\mathbf{X})\right\} = \mathrm{E}\left\{\frac{1-R}{1-\pi}b(\mathbf{X})\right\}.$$

Certainly, we also have

$$\mathrm{E}\left(\frac{R}{\pi}c\right) = \mathrm{E}\left(\frac{1-R}{1-\pi}c\right)$$

for any constant c. Combining the above three, we can obtain a family of mean zero functions

$$\frac{r}{\pi}\{\rho(y)y - b(\mathbf{x})\rho(y) + c\} + \frac{1-r}{1-\pi}\{b(\mathbf{x}) - \theta - c\} : \forall b(\mathbf{x}), \forall c. \tag{5.4}$$

Note that the model in (5.2) contains three unknown functions $p_Y(y)$, $g(\mathbf{x}, y)$ and

$\rho(y)$. For this model, in Section D.2 of the Supplement, we establish that

$$\mathcal{F} \equiv \left[\frac{r}{\pi}\{\rho(y)y - b(\mathbf{x})\rho(y) + c\} + \frac{1-r}{1-\pi}\{b(\mathbf{x}) - \theta - c\} : \mathrm{E}\{b(\mathbf{X}) \mid y\} = y, \forall c \right]$$

is the family of all influence functions (Bickel et al. 1993, Tsiatis 2006) for estimating θ. According to the definition of the influence function, $\mathcal{F}$ is sufficiently comprehensive since it can generate any regular asymptotically linear estimator of θ. The requirement $\mathrm{E}\{b(\mathbf{X}) \mid y\} = y$ in the definition of $\mathcal{F}$ is pivotal. Different from the mean zero function in (5.4), which critically relies on the correct specification of $\rho(y)$, the element in $\mathcal{F}$ preserves its zero mean even if $\rho(y)$ is misspecified as long as an appropriate $b(\mathbf{x})$ is chosen so that $\mathrm{E}\{b(\mathbf{X}) \mid y\} = y$. To further discover a wise choice of such a $b(\mathbf{x})$, we first derive a special element in $\mathcal{F}$, the efficient influence function $\phi_{\text{eff}}(\mathbf{x}, r, ry)$, that corresponds to the semiparametric efficiency bound and that provides guidance on constructing flexible estimators for θ.

Proposition 5.3.1. *The efficient influence function $\phi_{\text{eff}}(\mathbf{x}, r, ry)$ for θ is*

$$\begin{aligned} \phi_{\text{eff}}(\mathbf{x}, r, ry) &= \frac{r}{\pi}\rho(y)\left[y - \frac{\mathrm{E}_p\{a(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho(Y) \mid \mathbf{x}\}} \right] \\ &\quad + \frac{1-r}{1-\pi}\left[\frac{\mathrm{E}_p\{a(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho(Y) \mid \mathbf{x}\}} - \theta \right], \end{aligned}$$

where $a(y)$ satisfies

$$E\left[\frac{\mathrm{E}_p\{a(Y)\rho(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho(Y) \mid \mathbf{X}\}} \mid y \right] = y. \tag{5.5}$$

The detailed derivation of the efficient influence function in Proposition 5.3.1 is provided in Section D.3 of the Supplement. Clearly, the unique $b(\mathbf{x})$ that leads to the efficient influence function is

$$b(\mathbf{x}) \equiv \frac{\mathrm{E}_p\{a(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho(Y) \mid \mathbf{x}\}} = \frac{\mathrm{E}_q\{a(Y) \mid \mathbf{X}\}}{\mathrm{E}_q\{\rho(Y) \mid \mathbf{X}\} + \pi/(1-\pi)}.$$

In principle, if both $\rho(y)$ and $b(\mathbf{x})$ were known, we can estimate θ by solving the estimating equation

$\sum_{i=1}^{n} \phi_{\text{eff}}(\mathbf{x}_i, r_i, r_i y_i) = 0$, which leads to

$$\breve{\theta} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{r_i}{\pi} \rho(y_i)\{y_i - b(\mathbf{x}_i)\} + \frac{1 - r_i}{1 - \pi} b(\mathbf{x}_i) \right]. \tag{5.6}$$

However, the estimator $\breve{\theta}$ is impractical because of the following three obstacles. First, as we pointed out, $\rho(y)$ is almost infeasible to estimate based on the observed data. Second, $\mathrm{E}_p(\cdot \mid \mathbf{x})$ is unknown and needs to be estimated. Though various off-the-shelf machine learning or nonparametric regression methods are available, when the dimension of $\mathbf{x}$ is high, their performances are not always satisfactory and their computation can be expensive. The third obstacle lies in solving $a(y)$ from the integral equation (5.5), which requires $g(\mathbf{x}, y)$ to evaluate its left hand side. Estimating conditional density $g(\mathbf{x}, y)$ could be even more difficult than estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$, due to the curse of dimensionality.

Our proposed estimator will bypass the challenging task of estimating $\rho(y)$. Throughout the estimation procedure, only a working model $\rho^*(y)$ is needed, which can be arbitrarily misspecified hence is flexible. This turns out achievable through careful manipulation of other components of the efficient influence function. Our proposed estimator can also avoid estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$ even though we can do it if we decide to. This means that we can misspecify the conditional density model $p_{Y|\mathbf{X}}(y, \mathbf{x})$, encoded as $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x})$, while we also misspecify the density ratio $\rho(y)$. We call such an estimation procedure doubly flexible. To overcome the third obstacle, we recognize that $g(\mathbf{x}, y)$ only affects quantities of the form $\mathrm{E}(\cdot \mid y)$, which are one dimensional regression problems hence can be easily solved via the most basic nonparametric regression procedure such as the Nadaraya-Watson estimator.

In a nutshell, a unique feature of our work is the tolerance of both $\rho^*(y)$ and $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x})$, which can be simultaneously misspecified. We thus name the procedure doubly flexible.

5.3.2 Proposed doubly flexible estimator

Interestingly and critically, we discover that, even when both $\rho^*(y)$ and $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x})$ are misspecified, the corresponding estimator following the implementation of $\breve{\theta}$ in (5.6) is still consistent for θ. We summarize this result in Proposition 5.3.2 and give its proof in Section D.4 of the Supplement. Below, we use superscripts * and * to indicate that the corresponding quantities are calculated based on the working models $\rho^*(y)$ and $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x})$ respectively.

Proposition 5.3.2. *Define*

$$\widehat{\theta}_t = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\{y_i - b^{**}(\mathbf{x}_i)\} + \frac{1-r_i}{1-\pi}b^{**}(\mathbf{x}_i)\right],$$

where

$$b^{**}(\mathbf{x}) \equiv \frac{\mathrm{E}_p^\star\{a^{**}(Y)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}\}},$$

and $a^{**}(y)$ is a solution to

$$E\left[\frac{\mathrm{E}_p^\star\{a^{**}(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right] = y. \tag{5.7}$$

Then $\widehat{\theta}_t$ is a consistent estimator of θ.

In Proposition 5.3.2, the subscript $_t$ in $\widehat{\theta}_t$ indicates the conditional density $g(\mathbf{x}, y)$ in (5.7) is the truth. In reality, note that $g(\mathbf{x}, y)$ only involves in the evaluation of the conditional expectation $\mathrm{E}(\cdot \mid y)$ on the left hand side of (5.7). This is a one dimensional regression problem and can be easily estimated by many basic nonparametric regression procedures such as the Nadaraya-Watson estimator. Specifically, we approximate the integral equation (5.7) by

$$\begin{aligned} y &= \widehat{\mathrm{E}}\left[\frac{\mathrm{E}_p^\star\{a^{**}(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right] \qquad (5.8)\\ &= \sum_{i=1}^{n}\frac{\mathrm{E}_p^\star\{a^{**}(Y)\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}_i\}}\frac{r_i K_h(y-y_i)}{\sum_{j=1}^{n} r_j K_h(y-y_j)}\\ &= \int a^{**}(t)\rho^*(t)\sum_{i=1}^{n}\frac{p^\star_{Y|\mathbf{X}}(t,\mathbf{x}_i)}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}_i\}}\frac{r_i K_h(y-y_i)}{\sum_{j=1}^{n} r_j K_h(y-y_j)}dt, \end{aligned}$$

where $K_h(\cdot) \equiv K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth, with conditions imposed later in our theoretical investigation. (5.8) is a Fredholm integral equation of the first type, which is ill-posed. Numerical methods to provide stable and reliable solutions have been well studied in the literature (Hansen 1992). In our numerical implementations in Sections 5.5 and 5.6, we use Landweber's iterative method (Landweber 1951) that is well-known to produce a convergent solution. We provide those technical details in Section D.5 of the Supplement.

We summarize the complete estimation procedure in Algorithm 1.

Algorithm 1 Proposed Estimator $\widehat{\theta}$: Doubly Flexible in $\rho^*(y)$ and $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$

Input: data from population $\mathcal{P}$: $(y_i, \mathbf{x}_i, r_i = 1)$, $i = 1, \ldots, n_1$, data from population $\mathcal{Q}$: $(\mathbf{x}_j, r_j = 0)$, $j = n_1 + 1, \ldots, n$, and value $\pi = n_1/n$.

do

(a) adopt a working model for $\rho(y)$, denoted as $\rho^*(y)$;

(b) adopt a working model for $p_{Y|\mathbf{X}}(y,\mathbf{x})$, denoted as $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$ or $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$;

(c) compute $w_i = [\mathrm{E}^\star_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y) \mid \mathbf{x}_i\}]^{-1}$ for $i = 1, \ldots, n$;

(d) obtain $\widehat{a}^{**}(\cdot)$ by solving the integral equation (5.8);

(e) compute $\widehat{b}^{**}(\mathbf{x}_i) = w_i \mathrm{E}^\star_p\{\widehat{a}^{**}(Y)\rho^*(Y) \mid \mathbf{x}_i\}$ for $i = 1, \ldots, n$;

(f) obtain $\widehat{\theta}$ as

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\{y_i - \widehat{b}^{**}(\mathbf{x}_i)\} + \frac{1-r_i}{1-\pi}\widehat{b}^{**}(\mathbf{x}_i)\right]. \tag{5.9}$$

Output: $\widehat{\theta}$.

Remark 5.3.1. *In step (b) of Algorithm 1, one may adopt a completely specified $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$ or a partially specified model $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta})$ with an unknown parameter $\boldsymbol{\zeta}$. If the latter case, a natural strategy is to estimate $\boldsymbol{\zeta}$ first based on the observed samples from $\mathcal{P}$ via, say MLE, to obtain $\widehat{\boldsymbol{\zeta}}$, then use $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$ to replace the completely fixed $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$. In fact, we will show that the action of estimating $\boldsymbol{\zeta}$ has no consequence in terms of estimating θ. This is an important discovery, because this means one can always include a reasonably flexible model $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta})$ so that it has a good chance of approximating the true $p_{Y|\mathbf{X}}(y,\mathbf{x})$. If $p_{Y|\mathbf{X}}(y,\mathbf{x}) = p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta}_0)$ for certain $\boldsymbol{\zeta}_0$, then even though the additional parameter $\boldsymbol{\zeta}$ causes extra work, the reward is that θ can be estimated as efficiently as if we knew $p_{Y|\mathbf{X}}(y,\mathbf{x})$ completely. In all the subsequent steps, we replace $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$ by $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$ for its generality, bearing in mind that $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$ degenerates to $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$ when the parameter $\boldsymbol{\zeta}$ vanishes.*

We now study the theoretical properties of $\widehat{\theta}$ defined in (5.9). The main technical challenge is quantifying the gap between the solutions for the integral equations (5.7) and (5.8), encoded as $a^{**}(y)$ and $\widehat{a}^{**}(y)$ respectively. To facilitate the derivation, we define the linear operator

$$\mathcal{L}^{**}(a)(y) \equiv p_Y(y)\mathrm{E}\left[\frac{\mathrm{E}^\star_p\{a(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}^\star_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right] = \int a(t)u^{**}(t,y)dt,$$

where

$$u^{**}(t,y) \equiv p_Y(y)\int \frac{\rho^*(t)p^\star_{Y|\mathbf{X}}(t,\mathbf{x},\boldsymbol{\zeta})}{\mathrm{E}^\star_p\{\rho^{*2}(Y)\mid \mathbf{x}\}+\pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y)\mid \mathbf{x}\}}g(\mathbf{x},y)d\mathbf{x}.$$

Apparently, $a^{**}(y)$ satisfies

$$\mathcal{L}^{**}(a^{**})(y) = v(y), \text{ where } v(y)\equiv p_Y(y)y.$$

Similarly, $\widehat{a}^{**}(y)$ satisfies $\widehat{\mathcal{L}}^{**}(\widehat{a}^{**})(y)=\widehat{v}(y)$, where

$$\begin{aligned}\widehat{\mathcal{L}}^{**}(a)(y) &\equiv n_1^{-1}\sum_{i=1}^n r_i K_h(y-y_i)\frac{\mathrm{E}^\star_p\{a(Y)\rho^*(Y)\mid \mathbf{x}_i,\widehat{\boldsymbol{\zeta}}\}}{\mathrm{E}^\star_p\{\rho^{*2}(Y)\mid \mathbf{x}_i,\widehat{\boldsymbol{\zeta}}\}+\pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y)\mid \mathbf{x}_i,\widehat{\boldsymbol{\zeta}}\}}, \text{ and}\\ \widehat{v}(y) &\equiv n_1^{-1}\sum_{i=1}^n v_{i,h}(y) \equiv n_1^{-1}\sum_{i=1}^n r_i K_h(y-y_i)y.\end{aligned}$$

We first establish in Lemma 5.3.1 that given regularity conditions (A1)-(A4), the linear operator $\mathcal{L}^{**}$, as well as its inverse, is well behaved.

(A1) The working model $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$ or $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$ satisfies the completeness condition stated in Lemma 5.2.1.

(A2) $\rho^*(y)>\delta$ for all y on the support of $p_Y(y)$ where δ is a positive constant, and $\rho^*(y)$ is twice differentiable and its derivative is bounded.

(A3) The function $u^{**}(t,y)$ is bounded and has bounded derivatives with respect to t and y on its support. The function $a^{**}(y)$ in (5.7) is bounded.

(A4) The support sets of $g(\mathbf{x},y), p_Y(y), \rho^*(y)$ are compact.

Lemma 5.3.1. *Let $\|a\|_\infty \equiv \sup_y |a(y)|$. Under Conditions (A1)-(A4), the linear operator $\mathcal{L}^{**}: L^\infty(R)\to L^\infty(R)$ is invertible. In addition, there exist positive finite constants c_1, c_2 such that for all $a(y)\in L^\infty(R)$,*

(i) $c_1\|a\|_\infty \leqslant \|\mathcal{L}^{**}(a)\|_\infty \leqslant c_2\|a\|_\infty$,

(ii) $\|\mathcal{L}^{**-1}(a)\|_\infty \leqslant c_1^{-1}\|a\|_\infty$.

The proof of Lemma 5.3.1 is in Section D.6 of the Supplement. To analyze the asymptotic normality of the estimator $\widehat{\theta}$, we add two more regularity conditions on the kernel function and the bandwidth h.

(A5) The kernel function $K(\cdot) \geqslant 0$ is symmetric, bounded, and twice differentiable with bounded first derivative. It has support on $(-1, 1)$ and satisfies $\int_{-1}^{1} K(t)dt = 1$.

(A6) The bandwidth h satisfies $n_1(\log n_1)^{-4}h^2 \to \infty$ and $n^2 n_1^{-1} h^4 \to 0$.

Condition (A5) is standard for kernel functions. Note that we only need a one-dimensional kernel function $K(\cdot)$ in our estimation procedure. Condition (A6) specifies the requirement of the bandwidth h associated with kernel function $K(\cdot)$. In general we need both $h^{-1} = o\{n_1^{1/2}(\log n_1)^{-2}\}$ and $h = o(n^{-1/2}n_1^{1/4})$. If π is further assumed to be bounded away from zero, the second requirement becomes $h = o(n_1^{-1/4})$ and one can simply choose $h = n_1^{-1/3}$ to meet both requirements. We are now ready to present the asymptotic normality of the estimator $\widehat{\theta}$ below. Its proof is contained in Section D.7 of the Supplement.

Theorem 5.3.1. *Assume $\widehat{\boldsymbol{\zeta}}$ satisfies $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$ and $\mathrm{E}_p^\star\{\|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y, \mathbf{x}, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}\}$ is bounded, where $\mathbf{S}_{\boldsymbol{\zeta}}^\star(y, \mathbf{x}, \boldsymbol{\zeta}) \equiv \partial \log p_{Y|\mathbf{X}}^\star(y, \mathbf{x}, \boldsymbol{\zeta})/\partial\boldsymbol{\zeta}$. For any choice of $p_{Y|\mathbf{X}}^\star(y, \mathbf{x}, \boldsymbol{\zeta})$ and $\rho^*(y)$, under Conditions (A1)-(A6),*

$$\sqrt{n_1}(\widehat{\theta} - \theta) \to N(0, \sigma_\theta^2)$$

in distribution as $n_1 \to \infty$, where

$$\begin{aligned}\sigma_\theta^2 \;=\;& var\Big(\sqrt{\pi}\phi_{\mathrm{eff}}^{**\star}(\mathbf{X}, R, RY) \\ &+ \frac{R}{\sqrt{\pi}}\left[\frac{\mathrm{E}_p^\star\{a^{**\star}(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{X}\}} - Y\right]\{\rho^*(Y) - \rho(Y)\}\Big).\end{aligned}$$

In Theorem 5.3.1, the only requirement on $\widehat{\boldsymbol{\zeta}}$ is $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$. Thus, the asymptotic variance of $\sqrt{n_1}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})$ does not affect the result in Theorem 5.3.1 as long as $\widehat{\boldsymbol{\zeta}}$ is $\sqrt{n_1}$-consistent for $\boldsymbol{\zeta}$. This is easily achievable by constructing a standard MLE or moment based estimator for $\boldsymbol{\zeta}$ in the regression model of Y given $\mathbf{X}$, $p_{Y|\mathbf{X}}^\star(y, \mathbf{x}, \boldsymbol{\zeta})$, based on the n_1 observations from population $\mathcal{P}$.

In addition, it is clear that the estimator $\widehat{\theta}$ is $\sqrt{n_1}$-consistent, even if n goes to infinity much faster than n_1 does. The intuition is that when we only have n_1 complete observations in this problem, although a much larger n_0 can help us better understand the label shift mechanism, it cannot improve the convergence rate of $\widehat{\theta}$.

Finally, Theorem 5.3.1 also indicates that, when $\widehat{\boldsymbol{\zeta}}$ is a $\sqrt{n_1}$-consistent estimator for $\boldsymbol{\zeta}$ such that $p_{Y|\mathbf{X}}^\star(y, \mathbf{x}, \boldsymbol{\zeta}) = p_{Y|\mathbf{X}}(y, \mathbf{x})$, and $\rho^*(y)$ is correctly specified as $\rho(y)$, the corresponding estimator $\widehat{\theta}$ achieves the semiparametric efficiency bound and is the efficient

estimator. We state this result formally as Corollary 5.3.1. Since it is a special case of Theorem 5.3.1, its proof is omitted.

Corollary 5.3.1. *Assume* $\widehat{\boldsymbol{\zeta}}$ *satisfies* $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$ *and* $\mathrm{E}_p^\star\{\|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y, \mathbf{x}, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}\}$ *is bounded. If* $p_{Y|\mathbf{X}}^\star(y, \mathbf{x}, \boldsymbol{\zeta}) = p_{Y|\mathbf{X}}(y, \mathbf{x})$ *and* $\rho^*(y) = \rho(y)$*, under Conditions (A1)-(A6),*

$$\sqrt{n_1}(\widehat{\theta}_{\text{eff}} - \theta) \to N[0, var\{\sqrt{\pi}\phi_{\text{eff}}(\mathbf{X}, R, RY)\}]$$

in distribution as $n_1 \to \infty$*.*

5.4 Alternative singly flexible estimator

Because the assessment of $\mathrm{E}_p(\cdot \mid \mathbf{x})$ only relies on the observed data, instead of adopting an arbitrary known model $\mathrm{E}_p^\star(\cdot \mid \mathbf{x})$ or parametric model $\mathrm{E}_p^\star(\cdot \mid \mathbf{x}, \boldsymbol{\zeta})$, one might be willing to estimate $\mathrm{E}_p(\cdot \mid \mathbf{x})$ in a model free fashion and replace $\mathrm{E}_p^\star(\cdot \mid \mathbf{x})$ in the estimation procedure presented in Section 5.3.2 by a well-behaved estimator $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$. Here we consider a general estimator $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$ which has convergence rate faster than $n_1^{-1/4}$. This rate is achievable for many nonparametric regression or machine learning algorithms (Chernozhukov et al. 2018), see for example, Chen & White (1999) for a class of neural network models, Wager & Walther (2015) for a class of regression trees and random forests, and Bickel et al. (2009), Bühlmann & Van De Geer (2011), Belloni & Chernozhukov (2011, 2013) for a variety of sparse models. Meanwhile, we still do not aim to estimate $\rho(y)$ since we do not have the Y-data in population $\mathcal{Q}$. We denote the corresponding estimator $\widetilde{\theta}$ and call it singly flexible because of its flexibility in using a working model $\rho^*(y)$.

The idea behind the estimator $\widetilde{\theta}$ is similar to $\widehat{\theta}$, therefore we only emphasize the difference from Section 5.3.2. Similar to (5.7), we define $a^*(y)$ as the solution of

$$\mathrm{E}\left[\frac{\mathrm{E}_p\{a^*(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right] = y. \tag{5.10}$$

Equivalently, $a^*(y)$ satisfies $\mathcal{L}^*(a^*)(y) = v(y)$, where

$$\mathcal{L}^*(a)(y) \equiv p_Y(y)\mathrm{E}\left[\frac{\mathrm{E}_p\{a(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right] = \int a(t)u^*(t, y)dt,$$

and

$$u^*(t,y) \equiv p_Y(y)\int \frac{\rho^*(t)p_{Y|\mathbf{X}}(t,\mathbf{x})}{\mathrm{E}_p\{\rho^{*2}(Y)\mid\mathbf{x}\}+\pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y)\mid\mathbf{x}\}}g(\mathbf{x},y)d\mathbf{x}.$$

Using the estimator $\widehat{\mathrm{E}}_p(\cdot\mid\mathbf{x})$, we approximate the integral equation (5.10) as

$$\begin{aligned} y &= \widehat{\mathrm{E}}\left[\frac{\widehat{\mathrm{E}}_p\{a^*(Y)\rho^*(Y)\mid\mathbf{X}\}}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y)\mid\mathbf{X}\}+\pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y)\mid\mathbf{X}\}}\mid y\right] \\ &= \sum_{i=1}^n \frac{\widehat{\mathrm{E}}_p\{a^*(Y)\rho^*(Y)\mid\mathbf{x}_i\}}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y)\mid\mathbf{x}_i\}+\pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y)\mid\mathbf{x}_i\}}\frac{r_iK_h(y-y_i)}{\sum_{j=1}^n r_jK_h(y-y_j)}, \end{aligned} \quad (5.11)$$

and we write $\widehat{a}^*(y)$ as the solution to $\widehat{\mathcal{L}}^*(\widehat{a}^*)(y)=\widehat{v}(y)$, where

$$\widehat{\mathcal{L}}^*(a)(y) \equiv n_1^{-1}\sum_{i=1}^n r_iK_h(y-y_i)\frac{\widehat{\mathrm{E}}_p\{a(Y)\rho^*(Y)\mid\mathbf{x}_i\}}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y)\mid\mathbf{x}_i\}+\pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y)\mid\mathbf{x}_i\}}.$$

We summarize the algorithm for computing the estimator $\widetilde{\theta}$ below.

Algorithm 2 Alternative Estimator $\widetilde{\theta}$: Single Flexible in $\rho^*(y)$

Input: data from population $\mathcal{P}$: $(y_i,\mathbf{x}_i,r_i=1)$, $i=1,\ldots,n_1$, data from population $\mathcal{Q}$: $(\mathbf{x}_j,r_j=0)$, $j=n_1+1,\ldots,n$, and value $\pi=n_1/n$.

do

(a) adopt a working model for $\rho(y)$, denoted as $\rho^*(y)$;

(b) adopt a nonparametric or machine learning algorithm for estimating $\mathrm{E}_p(\cdot\mid\mathbf{x})$, denoted as $\widehat{\mathrm{E}}_p(\cdot\mid\mathbf{x})$;

(c) compute $\widehat{w}_i=[\widehat{\mathrm{E}}_p\{\rho^{*2}(Y)\mid\mathbf{x}_i\}+\pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y)\mid\mathbf{x}_i\}]^{-1}$ for $i=1,\ldots,n$;

(d) obtain $\widehat{a}^*(\cdot)$ by solving the integral equation (5.11);

(e) compute $\widehat{b}^*(\mathbf{x}_i)=\widehat{w}_i\widehat{\mathrm{E}}_p\{\widehat{a}^*(Y)\rho^*(Y)\mid\mathbf{x}_i\}$ for $i=1,\ldots,n$;

(f) obtain $\widetilde{\theta}$ as

$$\widetilde{\theta}=\frac{1}{n}\sum_{i=1}^n\left[\frac{r_i}{\pi}\rho^*(y_i)\{y_i-\widehat{b}^*(\mathbf{x}_i)\}+\frac{1-r_i}{1-\pi}\widehat{b}^*(\mathbf{x}_i)\right]. \quad (5.12)$$

Output: $\widetilde{\theta}$.

To develop the asymptotic normality of the estimator $\widetilde{\theta}$, instead of Condition (A3), we need

(A7) The function $u^*(t,y)$ is bounded and has bounded derivatives with respect to t and y on its support. The function $a^*(y)$ in (5.10) is bounded.

We present Theorem 5.4.1, with its proof contained in Section D.8 of the Supplement.

Theorem 5.4.1. *Assume $\widehat{E}_p$ satisfies $|\widehat{E}_p\{a(Y) \mid \mathbf{x}\} - \mathrm{E}_p\{a(Y) \mid \mathbf{x}\}| = o_p(n_1^{-1/4})$ for any bounded function $a(y)$. For any choice of $\rho^*(y)$, under Conditions (A2), (A4)-(A7),*

$$\sqrt{n_1}(\widetilde{\theta} - \theta) \to N(0, \sigma_\theta^2)$$

in distribution as $n_1 \to \infty$, where

$$\begin{aligned}\sigma_\theta^2 &= var\Big(\sqrt{\pi}\phi^*_{\text{eff}}(\mathbf{X}, R, RY) \\ &\quad + \frac{R}{\sqrt{\pi}}\left[\frac{\mathrm{E}_p\{a^*(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{X}\}} - Y\right]\{\rho^*(Y) - \rho(Y)\}\Big).\end{aligned}$$

It is direct from Theorem 5.4.1 that when the posited model $\rho^*(y)$ is correctly specified, the estimator $\widetilde{\theta}$ becomes the efficient estimator for θ. We point out this consequence as Corollary 5.4.1 below.

Corollary 5.4.1. *Assume $\widehat{E}_p$ satisfies $|\widehat{E}_p\{a(Y) \mid \mathbf{x}\} - \mathrm{E}_p\{a(Y) \mid \mathbf{x}\}| = o_p(n_1^{-1/4})$ for any bounded function $a(y)$. If $\rho^*(y) = \rho(y)$, under Conditions (A2), (A4)-(A7),*

$$\sqrt{n_1}(\widetilde{\theta}_{\text{eff}} - \theta) \to N[0, var\{\sqrt{\pi}\phi_{\text{eff}}(\mathbf{X}, R, RY)\}]$$

in distribution as $n_1 \to \infty$.

Last but not least, Sections 5.3 and 5.4 here only present the results for estimating $\theta = \mathrm{E}_q(Y)$. The whole story can be extended to a general parameter $\boldsymbol{\theta}$ such that $\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$, and the results are stated in Sections D.9 and D.11 of the Supplement. In our numerical studies in Sections 5.5 and 5.6 below, we analyze both $\mathrm{E}_q(Y)$ and the t-th quantile of population $\mathcal{Q}$, defined as

$$\tau_{q,t}(Y) = \inf\left[y : \mathrm{E}_q\{I(Y \leqslant y)\} \geqslant t\right],$$

where $0 < t < 1$. This corresponds to $\mathrm{E}_q\left[\eta_t\left\{Y - \tau_{q,t}(Y)\right\}\right] = 0$ where $\eta_t(r) = t - I(r < 0)$.

5.5 Simulation experiments

We conduct simulation studies to assess the finite sample performance of our proposed methods. We report the results for the mean $\mathrm{E}_q(Y)$ and the median $\tau_{q,0.5}(Y)$ of the outcome Y in population $\mathcal{Q}$.

We first generate a binary indicator $R_i, i = 1, \ldots, n$ from the Bernoulli distribution with probability 0.5, and record $n_1 = \sum_{i=1}^n r_i$, $\pi = n_1/n$. Then we generate Y_i from $N(0,1)$ if $R_i = 1$ and from $N(1,1)$ if $R_i = 0$. The $\mathbf{X} \mid Y$ distribution is generated from a 3-dimensional normal with mean $(-0.5, 0.5, 1)^{\mathrm{T}} Y_i$ and covariance $\mathbf{I}$, the identity matrix. This implies, $\mathrm{E}_q(Y) = 1$, $\tau_{q,0.5} = 1$ and the true density ratio model

$$\rho(y) = \exp(-0.5 + y).$$

One can derive that $p_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta})$ follows normal with mean $(1, \mathbf{x}^{\mathrm{T}})\boldsymbol{\beta}$ where

$$\boldsymbol{\beta} = (0, -0.2, 0.2, 0.4)^{\mathrm{T}}$$

and variance $\sigma^2 = 0.4$. Here we denote $\boldsymbol{\zeta} \equiv (\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$.

We use the following misspecified working models. We define

$$\rho^*(y) \equiv c^* \exp(-0.7 + 1.2y),$$

where $c^* \equiv \pi / \{n^{-1} \sum_{i=1}^n r_i \exp(-0.7 + 1.2 y_i)\}$ in order to satisify $\mathrm{E}\{R\rho^*(Y)\} = \pi$. For the working model $p^\star_{Y|\mathbf{X}}$, we define $\mathbf{x}^\star \equiv [x_1, \exp(x_2/2), x_3/\{1 + \exp(x_2)\} + 10]^{\mathrm{T}}$ and define $p^\star_{Y|\mathbf{X}}(y, \mathbf{x}^\star, \boldsymbol{\zeta}^\star)$ as the normal distribution with mean $(1, \mathbf{x}^{\star\mathrm{T}})\boldsymbol{\beta}^\star$ where $\boldsymbol{\beta}^\star = (-7.000, -0.223, 0.363, 0.664)^{\mathrm{T}}$ and variance $\sigma^{\star 2} = 0.449$. The parameter $\boldsymbol{\zeta}^\star \equiv (\boldsymbol{\beta}^{\star\mathrm{T}}, \sigma^{\star 2})^{\mathrm{T}}$ is obtained by minimizing the Kullback-Leibler distance $D_{kl}(p_{Y|\mathbf{X}} \| p^\star_{Y|\mathbf{X}})$.

We implement the following seven estimators:

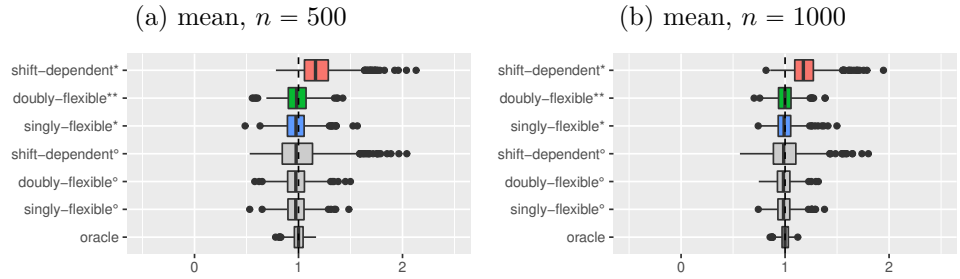
(1) `shift-dependent`*: $\breve{\theta}$ in (5.3) with $\rho^*(y)$;

(2) `doubly-flexible`**: $\widehat{\theta}$ in (5.9) with $\rho^*(y)$ and $p^\star_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta}^\star)$, theoretically analyzed in Theorem 5.3.1;

(3) `singly-flexible`*: $\widetilde{\theta}$ in (5.12) with $\rho^*(y)$, theoretically analyzed in Theorem 5.4.1;

(4) `shift-dependent`^{0}: $\breve{\theta}$ in (5.3) with correct $\rho(y)$;

Figure 5.1: Boxplots of estimates for mean in the simulation study. Dashed: the true estimand.



(5) `doubly-flexible`^{0}: $\widehat{\theta}_{\text{eff}}$ with correct $\rho(y)$ and $p_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta})$, theoretically analyzed in Corollary 5.3.1;

(6) `singly-flexible`^{0}: $\widetilde{\theta}_{\text{eff}}$ with correct $\rho(y)$, theoretically analyzed in Corollary 5.4.1;

(7) `oracle`: the $\sqrt{n_0}$-consistent estimator $\frac{1}{n}\sum_{i=1}^{n}\frac{1-r_i}{1-\pi}y_i$.

Note that the last four estimators (shown as "gray" in Figures 5.1 and 5.2) are unrealistic since they either use the unknown models $\rho(y)$ and $p_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta})$ or the Y-data in population $\mathcal{Q}$.

In implementing estimators `doubly-flexible`**, `singly-flexible`*, `doubly-flexible`^{0} and `singly-flexible`^{0}, we solve the integral equations (5.8) and (5.11) using the Nadaraya-Watson estimator for $\mathrm{E}(\cdot \mid y)$ with Gaussian kernel and bandwidth $h = n_1^{-1/3}$ that is discussed in Condition (A6). Numerically, the integrations are approximated by the Gauss-Legendre quadrature with 50 points on the interval $[-5, 5]$ and the integral equations are evaluated at $y_i, i = 1, \ldots, n_1$. In addition, for estimators `singly-flexible`* and `singly-flexible`^{0}, we estimate $\mathrm{E}_p(\cdot \mid \mathbf{x})$ using the Nadaraya-Watson estimator based on the product Gaussian kernel with bandwidth $2.5n_1^{-1/7}$, where the order comes from the optimal bandwidth $n_1^{-1/(4+d)}$ with d the dimensionality of covariate $\mathbf{X}$. See Section D.5 of the Supplement for technical details on the numerical implementation.

Based on 1000 simulation replicates, Figures 5.1 and 5.2 illustrate the boxplots of the estimates for the mean and the median, respectively. With the misspecified working model $\rho^*(y)$, the estimator `shift-dependent`* is biased; in contrast, the proposed estimators `doubly-flexible`** and `singly-flexible`* are both unbiased. When the correct model $\rho(y)$ is used, not surprisingly, all of the estimators `shift-dependent`^{0}, `doubly-flexible`^{0} and `singly-flexible`^{0} are unbiased. It is also clear that the two

Figure 5.2: Boxplots of estimates for median in the simulation study. Dashed: the true estimand.

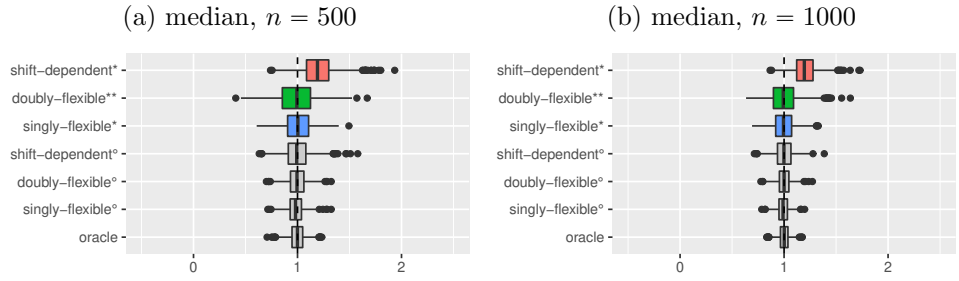


Table 5.1: Summary of mean estimation results in the simulation study.

n	Estimator	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x})$	MSE	Bias	SE	$\widehat{\text{SE}}$	CI
500	shift-dependent*	$\rho^*(y)$	-	.0699	.1840	.1899	.2791	1.000
	doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	.0173	-.0120	.1311	.1287	.943
	singly-flexible*	$\rho^*(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	.0162	-.0201	.1258	.1230	.941
	shift-dependent^0	$\rho(y)$	-	.0533	.0049	.2309	.2119	.899
	doubly-flexible^0	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	.0153	-.0231	.1214	.1212	.941
	singly-flexible^0	$\rho(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	.0138	-.0221	.1155	.1169	.939
	oracle	-	-	.0040	.0006	.0636	.0633	.943
1000	shift-dependent*	$\rho^*(y)$	-	.0561	.1906	.1406	.2077	.999
	doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	.0085	.0013	.0922	.0912	.955
	singly-flexible*	$\rho^*(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	.0081	-.0031	.0899	.0850	.952
	shift-dependent^0	$\rho(y)$	-	.0275	.0024	.1660	.1533	.927
	doubly-flexible^0	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	.0075	-.0125	.0856	.0861	.958
	singly-flexible^0	$\rho(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	.0069	-.0094	.0824	.0827	.955
	oracle	-	-	.0020	.0008	.0451	.0447	.945

proposed flexible estimators are always more efficient than the shift-dependent estimator no matter the correct model $\rho(y)$ is used or not.

To further demonstrate the efficiency comparison and the inference results, in Tables 5.1 and 5.2, we report the mean squared error (MSE), the empirical bias (Bias), the empirical standard error (SE), the average of estimated standard error ($\widehat{\text{SE}}$), and the empirical coverage at 95% confidence level (CI), for each of the estimators. The estimator `shift-dependent`* has an incorrect coverage (over-coverage in Table 5.1 and under-coverage in Table 5.2) because of its severe bias. This issue is not mitigated at all in Table 5.1 or becomes even worse in Table 5.2 when we increase the size of the stacked random sample from 500 to 1000. On the contrary, the estimators `doubly-flexible`**

Table 5.2: Summary of median estimation results in the simulation study.

n	Estimator	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x})$	MSE	Bias	SE	$\widehat{\text{SE}}$	CI
500	shift-dependent*	$\rho^*(y)$	-	.0695	.2025	.1687	.1547	.802
	doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	.0368	-.0072	.1918	.1764	.940
	singly-flexible*	$\rho^*(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	.0211	.0024	.1453	.1390	.941
	shift-dependent^0	$\rho(y)$	-	.0178	.0011	.1336	.1286	.947
	doubly-flexible^0	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	.0093	-.0033	.0964	.0950	.951
	singly-flexible^0	$\rho(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	.0071	-.0162	.0827	.0881	.956
	oracle	-	-	.0064	-.0020	.0799	.0821	.946
1000	shift-dependent*	$\rho^*(y)$	-	.0554	.2018	.1210	.1104	.560
	doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	.0229	-.0013	.1515	.1444	.943
	singly-flexible*	$\rho^*(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	.0128	-.0023	.1130	.1124	.954
	shift-dependent^0	$\rho(y)$	-	.0091	.0005	.0955	.0915	.934
	doubly-flexible^0	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	.0047	.0000	.0688	.0669	.954
	singly-flexible^0	$\rho(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	.0036	-.0098	.0594	.0625	.948
	oracle	-	-	.0031	.0004	.0558	.0578	.959

and `singly-flexible`* are correctly covered. Though there is no theoretical justification, `doubly-flexible`** is slightly less efficient than `singly-flexible`* in this setting. This indicates that the effort of correctly estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$ pays off in the sense of improving the estimation efficiency. With the correct $\rho(y)$ model used, each of the estimators `doubly-flexible`^{0} and `singly-flexible`^{0} is more efficient than its counterpart `doubly-flexible`** and `singly-flexible`*. The $\sqrt{n_0}$-consistent estimator `oracle` is the most efficient one in this simulation setting.

5.6 Data application

We now illustrate the numerical performance of our proposed method through analyzing the Medical Information Mart for Intensive Care III (MIMIC-III), an openly available electronic health records database, developed by the MIT Lab for Computational Physiology (Johnson et al. 2016). It comprises deidentified health-related records including demographics, vital signs, laboratory test and medications, for 46,520 patients who admitted to the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012.

The outcome of interest Y in our analysis is the sequential organ failure assessment (SOFA) score (Singer et al. 2016), used to track a patient's status during the stay in

an intensive care unit to determine the extent of a patient's organ function or rate of failure. The score is based on six different sub-scores, with one of each for the respiratory, cardiovascular, hepatic, coagulation, renal and neurological systems. The SOFA score ranges from 0 (best) to 24 (worst). We include 16 covariates from either chart events (6 variables, diastolic blood pressure, systolic blood pressure, blood glucose, respiratory rate per minute, and two measures from body temperature) or laboratory tests (10 variables, peripheral caillary oxygen saturation, two measures from each of hematocrit level, platelets count and red blood cell count, and three measures from blood urea nitrogen). We choose these covariates through assessing whether the absolute correlation with the outcome Y is greater than 0.2 and whether the missing rate is less than 1%. In our analysis, we only include the first admission if the patient was admitted to the intensive care unit more than once. We also exclude patients whose outcome Y is greater than or equal to 20, whose age is greater than or equal to 65, and who has missing values in any of the covariates. This results in a total of $n = 16,691$ records.

In our analysis, we define the population $\mathcal{P}$ as patients with private, goverment, and self-pay insurances ($R = 1, n_1 = 11,695$), and population $\mathcal{Q}$ as patients whose insurance type is either Medicaid or Medicare ($R = 0, n_0 = 4,996$). The label shift assumption that the conditional distribution of $\mathbf{X}$ given Y remains the same can be tested via the conditional independence of $\mathbf{X}$ and R given Y. In our analysis, we test the conditional independence between R and each of covariates by the invariant environment prediction test (Heinze-Deml et al. 2018) in R package `CondIndTests`, and the p-values range from 0.460 to 0.628. This indicates the label shift assumption is indeed sensible in our analysis. We first compute the sample mean (3.7409) and sample t-th quantiles (1,1,3,5,8 for $t = (10, 25, 50, 75, 90)\%$) of SOFA scores among patients whose insurance type is either Medicaid or Medicare. We regard these estimates as `oracle` in order to compare with our proposed methods.

To identify a reasonable working model $\rho^*(y)$, we model the data $Y + 0.001$ from population $\mathcal{P}$ as a parametric gamma distribution $f(y, \alpha, \beta) = \Gamma(\alpha)^{-1}\beta^{-\alpha}y^{\alpha-1}\exp(-y/\beta)$ with $\Gamma(\cdot)$ the Γ-function, the shape parameter $\alpha > 0$ and the scale parameter $\beta > 0$. We estimate the unknown parameters α and β as $\widehat{\alpha}$ and $\widehat{\beta}$ using the MLE. For the Y-data in population $\mathcal{Q}$, we assume $Y + 0.001$ follows a similar gamma distribution with shape parameter $\widehat{\alpha} + 1$ and scale parameter $\widehat{\beta}$. Hence, we use the working model

$$\rho^*(y) = \frac{f(y + 0.001, \widehat{\alpha} + 1, \widehat{\beta})}{f(y + 0.001, \widehat{\alpha}, \widehat{\beta})}.$$

Table 5.3: Mean estimation results in the data application.

Estimator	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x})$	Estimate	diff. with oracle	$\widehat{\text{SE}}$	CI
shift-dependent*	$\rho^*(y)$	-	4.0529	.3120	.0593	[3.9367, 4.1691]
doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	3.7579	.0170	.0803	[3.6005, 3.9153]
singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	3.7542	.0133	.0496	[3.6570, 3.8514]
oracle	-	-	3.7409	-	.0405	[3.6616, 3.8202]

Table 5.4: Quantile estimation results in the data application.

τ	Estimator	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x})$	Estimate	diff. with oracle	$\widehat{\text{SE}}$	CI
10%	shift-dependent*	$\rho^*(y)$	-	1.0000	.0000	.0102	[0.9801, 1.0199]
	doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.9998	-.0002	.0095	[0.9812, 1.0185]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	0.9998	-.0002	.0099	[0.9803, 1.0192]
	oracle	-	-	1.0000	-	.0218	[0.9572, 1.0428]
25%	shift-dependent*	$\rho^*(y)$	-	1.9995	.9995	.0249	[1.9508, 2.0483]
	doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	1.0002	.0002	.0276	[0.9460, 1.0543]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	1.0004	.0004	.0196	[0.9620, 1.0389]
	oracle	-	-	1.0000	-	.0315	[0.9383, 1.0617]
50%	shift-dependent*	$\rho^*(y)$	-	3.0002	.0002	.0408	[2.9203, 3.0801]
	doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	3.0001	.0001	.0333	[2.9348, 3.0654]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	3.0005	.0005	.0334	[2.9350, 3.0659]
	oracle	-	-	3.0000	-	.0550	[2.8923, 3.1077]
75%	shift-dependent*	$\rho^*(y)$	-	5.9999	.9999	.0742	[5.8544, 6.1454]
	doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	5.0001	.0001	.0759	[4.8512, 5.1489]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	5.0001	.0001	.0381	[4.9254, 5.0748]
	oracle	-	-	5.0000	-	.0591	[4.8842, 5.1158]
90%	shift-dependent*	$\rho^*(y)$	-	8.9999	.9999	.1380	[8.7295, 9.2703]
	doubly-flexible**	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	8.0004	.0004	.1003	[7.8039, 8.1969]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	8.0004	.0004	.0638	[7.8754, 8.1253]
	oracle	-	-	8.0000	-	.1093	[7.7858, 8.2142]

To implement the estimator `doubly-flexible`**, we impose a parametric model $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta})$ by regressing $Y + 0.001$ on $\mathbf{X}$ as a generalized linear model with gamma distribution, and estimate $\boldsymbol{\zeta}$ using the MLE. To implement the estimator `singly-flexible`*, we identify the first three principal components from the 16 covariates, and then estimate $\text{E}_p(\cdot \mid \mathbf{x})$ as a function of those three principal components using the Nadaraya-Watson estimator based on the product Gaussian kernel with bandwidth $0.5n_1^{-1/7}$. To solve the corresponding integral equations, similar to Section 5.5, we approximate $\text{E}(\cdot \mid y)$ by its Nadaraya-Watson estimator with the Gaussian kernel and bandwidth $h = n_1^{-1/3}$. In addition, for numerical implementation, the integration is approximated at 50 equally-spaced points on the interval $[0, 19]$, and the integral equations are evaluated at each supporting point of $\{y_i : i = 1, \ldots, n_1\}$. See Section D.5 of the Supplement for technical details on the implementation.

The results are summarized in Table 5.3 for the mean and in Table 5.4 for the quantiles. The estimator `shift-dependent`* that relies on a misspecified model $\rho^*(y)$ severely over-estimates the quantities compared to the `oracle` estimate, in most of the

scenarios including estimating the mean, 25%, 75% and 90% quantiles. As a consequence, the `oracle` estimate cannot be covered by the confidence intervals. In contrast, the proposed estimators `doubly-flexible`** and `singly-flexible`*, although also rely on $\rho^*(y)$, provide almost identical estimates as `oracle`. Accordingly, the confidence intervals from the proposed methods all cover the `oracle` estimate. In these scenarios, `singly-flexible`* is more efficient than `doubly-flexible`**, which echoes our findings in Section 5.5.

When estimating 10% and 50% quantiles, we find that `shift-dependent`* gives almost the same estimate as `oracle`. It might be plausible that the difference between $\rho^*(y)$ and the true $\rho(y)$ is minor for estimating these two quantities. Nevertheless, the proposed estimators `doubly-flexible`** and `singly-flexible`* are still more efficient than the estimator `shift-dependent`*.

Finally it is interesting to observe that, the estimator `oracle` is even less efficient than the proposed estimators `doubly-flexible`** and `singly-flexible`*, in estimating all of the quantiles. This is because `oracle` is $\sqrt{n_0}$-consistent whereas the two proposed estimators are both $\sqrt{n_1}$-consistent. In this application, n_1 is 2.34 times greater than n_0, which might result in the situation that the `oracle` estimate being less efficient. In Section 5.5, we also considered situations that n_1 is much larger than n_0 and similar phenomenon was observed as well, with detailed results omitted.

5.7 Discussion

In this paper, we estimate a characteristic of a target population $\mathcal{Q}$, via exploiting the data and information from a different but relevant population $\mathcal{P}$, under the label shift assumption. Different from most existing literatures, our proposal is devised to accommodate both classification and regression problems. We primarily propose the doubly flexible estimate, whose unique feature is to simultaneously allow both models to be misspecified thus is flexible: the density ratio model $\rho(y)$ that governs the label shift mechanism, and the conditional distribution model $p_{Y|\mathbf{X}}(y, \mathbf{x})$ of population $\mathcal{P}$. While the estimation of the latter can be done via off-the-shelf procedures sometimes, it often faces curse of dimensionality or computational challenges. Further, estimating $\rho(y)$ is even more difficult because the Y-data in population $\mathcal{Q}$ is not accessible in our procedure.

Appendix A
Supplement to Chapter 2

A.1 Categorical treatments

A.1.1 Asymptotic distribution of $\widehat{\theta}_k$

Let

$$\mathbf{f}_{ki}(\boldsymbol{\beta}) \equiv \left\{ \frac{I(A_i = k)}{\pi(k, \mathbf{X}_i, \boldsymbol{\beta})} - 1 \right\} \mathbf{B}(k, \mathbf{X}_i),$$

$\mathbf{f}_i(\boldsymbol{\beta}) \equiv \{\mathbf{f}_{1i}(\boldsymbol{\beta})^{\mathrm{T}}, \ldots, \mathbf{f}_{Ki}(\boldsymbol{\beta})^{\mathrm{T}}\}^{\mathrm{T}}$, $\mathbf{V}(\boldsymbol{\beta}) \equiv E\{\mathbf{f}_i(\boldsymbol{\beta})\mathbf{f}_i(\boldsymbol{\beta})^{\mathrm{T}}\}$, $\widehat{\mathbf{V}}(\boldsymbol{\beta}) \equiv n^{-1}\sum_{i=1}^n \mathbf{f}_i(\boldsymbol{\beta})\mathbf{f}_i(\boldsymbol{\beta})^{\mathrm{T}}$, $\mathbf{A}(\boldsymbol{\beta}) \equiv E\left\{\partial \mathbf{f}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{\mathrm{T}}\right\}$ and $\widehat{\mathbf{A}}(\boldsymbol{\beta}) \equiv n^{-1}\sum_{i=1}^n \partial \mathbf{f}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{\mathrm{T}}$.

Lemma A.1.1. *Under regularity conditions A0, A1, A2 and A3, the GMM estimator $\widehat{\boldsymbol{\beta}}$ obtained by minimizing $\{\sum_{i=1}^n \mathbf{f}_i(\boldsymbol{\beta})\}\widehat{\mathbf{V}}(\boldsymbol{\beta})^{-1}\{\sum_{i=1}^n \mathbf{f}_i(\boldsymbol{\beta})\}$, is such that*

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = -\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\{n^{-1/2}\sum_{i=1}^n \mathbf{f}_i(\boldsymbol{\beta}^*)\} + O_p(n^{-1/2}).$$

When (2.1) holds $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$.

Proof. The GMM estimator $\widehat{\boldsymbol{\beta}}$ is obtained by minimizing $\{\sum_{i=1}^n \mathbf{f}_i(\boldsymbol{\beta})\}\widehat{\mathbf{V}}(\boldsymbol{\beta})^{-1}\{\sum_{i=1}^n \mathbf{f}_i(\boldsymbol{\beta})\}$. This entails

$$\begin{aligned}
\mathbf{0} &= \widehat{\mathbf{A}}(\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})^{-1}\{n^{-1/2}\sum_{i=1}^n \mathbf{f}_i(\widehat{\boldsymbol{\beta}})\} + \frac{n^{1/2}}{2}[\{\frac{1}{n}\sum_{i=1}^n \mathbf{f}_i(\boldsymbol{\beta})\}\frac{\partial\{\widehat{\mathbf{V}}(\boldsymbol{\beta})^{-1}\}}{\partial \beta_k}\{\frac{1}{n}\sum_{i=1}^n \mathbf{f}_i(\boldsymbol{\beta})\}]_{k=1}^p \\
&= \widehat{\mathbf{A}}(\boldsymbol{\beta})^{\mathrm{T}}\widehat{\mathbf{V}}(\boldsymbol{\beta})^{-1}\{n^{-1/2}\sum_{i=1}^n \mathbf{f}_i(\widehat{\boldsymbol{\beta}})\} + O_p(n^{-1/2})
\end{aligned}$$

$$= \mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\{n^{-1/2}\sum_{i=1}^{n}\mathbf{f}_i(\boldsymbol{\beta}^*)\} + \mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + O_p(n^{-1/2}),$$

hence

$$n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) = -\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\{n^{-1/2}\sum_{i=1}^{n}\mathbf{f}_i(\boldsymbol{\beta}^*)\} + O_p(n^{-1/2}).$$

□

Proof of Theorem 2.2.1. Using Lemma A.1.1 we can write

$$\begin{aligned}
n^{-1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) &= n^{-1/2}\sum_{i=1}^{n}\mathbf{g}_i(\widehat{\boldsymbol{\beta}}) \\
&= n^{-1/2}\sum_{i=1}^{n}\{\mathbf{g}_i(\widehat{\boldsymbol{\beta}}) - \mathbf{g}_i(\boldsymbol{\beta}^*)\} + n^{-1/2}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta}^*) \\
&= -\mathbf{B}(\boldsymbol{\beta}^*)\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\{n^{-1/2}\sum_{i=1}^{n}\mathbf{f}_i(\boldsymbol{\beta}^*)\} \\
&\quad +n^{-1/2}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta}^*) + O_p(n^{-1/2}).
\end{aligned}$$

When either (2.1) and/or (2.2) hold, we already know that $E\{\mathbf{g}_i(\boldsymbol{\beta}^*)\} = \mathbf{0}$. Thus, under regularity conditions, $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ has asymptotic normal distribution with mean zero and variance

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathrm{var}\left[-\mathbf{B}(\boldsymbol{\beta}^*)\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{f}_i(\boldsymbol{\beta}^*) + \mathbf{g}_i(\boldsymbol{\beta}^*)\right] \\
&= \mathbf{B}(\boldsymbol{\beta}^*)\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{B}(\boldsymbol{\beta}^*)^{\mathrm{T}} + \mathbf{C}(\boldsymbol{\beta}^*) \\
&\quad -\mathbf{B}(\boldsymbol{\beta}^*)\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{D}(\boldsymbol{\beta}^*) \\
&\quad -\mathbf{D}(\boldsymbol{\beta}^*)^{\mathrm{T}}[\mathbf{B}(\boldsymbol{\beta}^*)\{\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}\mathbf{A}(\boldsymbol{\beta}^*)\}^{-1}\mathbf{A}(\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}^*)^{-1}]^{\mathrm{T}},
\end{aligned}$$

where $\mathbf{C}(\boldsymbol{\beta}^*) \equiv E\{\mathbf{g}_i(\boldsymbol{\beta}^*)^{\otimes 2}\}$ and $\mathbf{D}(\boldsymbol{\beta}^*) \equiv E\{\mathbf{f}_i(\boldsymbol{\beta}^*)\mathbf{g}_i(\boldsymbol{\beta}^*)^{\mathrm{T}}\}$. □

Proof of Corollary 2.2.1. When all models are correctly specified, i.e. (2.1-2.2) hold, we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. Then

$$\mathbf{A}_k(\boldsymbol{\beta}_0) = E\left\{-\frac{\mathbf{B}(k,\mathbf{X}_i)\pi'_{\boldsymbol{\beta}}(k,\mathbf{X}_i,\boldsymbol{\beta}_0)^{\mathrm{T}}}{\pi(k,\mathbf{X}_i,\boldsymbol{\beta}_0)}\right\},$$

$$
\begin{aligned}
\mathbf{B}_k(\boldsymbol{\beta}_0) &= E\left\{-\frac{m(k,\mathbf{X}_i)\pi'_{\boldsymbol{\beta}}(k,\mathbf{X}_i,\boldsymbol{\beta}_0)^{\mathrm{T}}}{\pi(k,\mathbf{X}_i,\boldsymbol{\beta}_0)}\right\},\\
\mathbf{V}_{kl}(\boldsymbol{\beta}_0) &= E\left[\left\{\frac{I(k=l)}{\pi(k,\mathbf{X}_i,\boldsymbol{\beta}_0)}-1\right\}\mathbf{B}(k,\mathbf{X}_i)\mathbf{B}(l,\mathbf{X}_i)^{\mathrm{T}}\right],\\
\mathbf{C}_{kl}(\boldsymbol{\beta}_0) &= E\left\{I(k=l)\frac{m(k,\mathbf{X}_i)^2+v(k,\mathbf{X}_i)}{\pi(k,\mathbf{X}_i,\boldsymbol{\beta}_0)}-m(k,\mathbf{X}_i)m(l,\mathbf{X}_i)\right\}\\
&\quad +E\left([m(k,\mathbf{X}_i)-E\{m(k,\mathbf{X}_i)\}][m(l,\mathbf{X}_i)-E\{m(l,\mathbf{X}_i)\}]\right),\\
\mathbf{D}_{kl}(\boldsymbol{\beta}_0) &= E\left[\left\{\frac{I(k=l)}{\pi(k,\mathbf{X}_i,\boldsymbol{\beta}_0)}-1\right\}\mathbf{B}(k,\mathbf{X}_i)m(l,\mathbf{X}_i)\right],
\end{aligned}
$$

and $\mathbf{A}(\boldsymbol{\beta}_0)=\{\mathbf{A}_1(\boldsymbol{\beta}_0)^{\mathrm{T}},\ldots,\mathbf{A}_K(\boldsymbol{\beta}_0)^{\mathrm{T}}\}^{\mathrm{T}}$, $\mathbf{B}(\boldsymbol{\beta}_0)=\{\mathbf{B}_1(\boldsymbol{\beta}_0)^{\mathrm{T}},\ldots,\mathbf{B}_K(\boldsymbol{\beta}_0)^{\mathrm{T}}\}^{\mathrm{T}}$, $\mathbf{V}(\boldsymbol{\beta}_0)=\{\mathbf{V}_{kl}(\boldsymbol{\beta}_0)\}_{k,l=1}^K$, $\mathbf{C}(\boldsymbol{\beta}_0)=\{\mathbf{C}_{kl}(\boldsymbol{\beta}_0)\}_{k,l=1}^K$, $\mathbf{D}(\boldsymbol{\beta}_0)=\{\mathbf{D}_{kl}(\boldsymbol{\beta}_0)\}_{k,l=1}^K$.

Note that $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{A}_k(\boldsymbol{\beta}_0)=\mathbf{B}_k(\boldsymbol{\beta}_0)$ and $\mathbf{V}_{kl}(\boldsymbol{\beta}_0)\boldsymbol{\alpha}=\mathbf{D}_{kl}(\boldsymbol{\beta}_0)$, so $(\mathbf{I}_{K+1}\otimes\boldsymbol{\alpha}^{\mathrm{T}})\mathbf{A}(\boldsymbol{\beta}_0)=\mathbf{B}(\boldsymbol{\beta}_0)$ and $\mathbf{V}(\boldsymbol{\beta}_0)(\mathbf{1}_{K,K}\otimes\boldsymbol{\alpha})=\mathbf{D}(\boldsymbol{\beta}^*)$. Thus,

$$
\boldsymbol{\Sigma}=\mathbf{C}(\boldsymbol{\beta}_0)-\mathbf{B}(\boldsymbol{\beta}_0)\{\mathbf{A}(\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}_0)^{-1}\mathbf{A}(\boldsymbol{\beta}_0)\}^{-1}\mathbf{B}(\boldsymbol{\beta}_0)^{\mathrm{T}}.
$$

□

Proof of Corollary 2.2.2. Here we have set the dimension of $\mathbf{f}_i(\boldsymbol{\beta})$ to be the same as the dimension of $\boldsymbol{\beta}$, hence we can solve $\sum\mathbf{f}_i(\boldsymbol{\beta})=\mathbf{0}$ directly. As a consequence, we can write

$$
\begin{aligned}
\mathbf{0} &= n^{-1/2}\sum_{i=1}^n\mathbf{f}_i(\widehat{\boldsymbol{\beta}})+O_p(n^{-1/2})\\
&= n^{-1/2}\sum_{i=1}^n\mathbf{f}_i(\boldsymbol{\beta}^*)+\mathbf{A}(\boldsymbol{\beta}^*)n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})+O_p(n^{-1/2}),
\end{aligned}
$$

hence

$$
n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})=-\mathbf{A}(\boldsymbol{\beta}^*)^{-1}\{n^{-1/2}\sum_{i=1}^n\mathbf{f}_i(\boldsymbol{\beta}^*)\}+O_p(n^{-1/2}).
$$

This leads to

$$
\begin{aligned}
n^{-1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) &= n^{-1/2}\sum_{i=1}^n\mathbf{g}_i(\widehat{\boldsymbol{\beta}})\\
&= n^{-1/2}\sum_{i=1}^n\{\mathbf{g}_i(\widehat{\boldsymbol{\beta}})-\mathbf{g}_i(\boldsymbol{\beta}_0)\}+n^{-1/2}\sum_{i=1}^n\mathbf{g}_i(\boldsymbol{\beta}_0)
\end{aligned}
$$

$$
\begin{aligned}
&= -\mathbf{B}(\boldsymbol{\beta}_0)\mathbf{A}(\boldsymbol{\beta}_0)^{-1}\{n^{-1/2}\sum_{i=1}^{n}\mathbf{f}_i(\boldsymbol{\beta}_0)\} + n^{-1/2}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta}_0) + O_p(n^{-1/2}) \\
&= -(\mathbf{I}_{K+1}\otimes\boldsymbol{\alpha}^{\mathrm{T}})\{n^{-1/2}\sum_{i=1}^{n}\mathbf{f}_i(\boldsymbol{\beta}_0)\} + n^{-1/2}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta}_0) + O_p(n^{-1/2}).
\end{aligned}
$$

Thus, $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ has asymptotic normal distribution with mean zero and variance

$$
\begin{aligned}
\boldsymbol{\Sigma} &= \mathrm{var}\left\{\mathbf{g}_i(\boldsymbol{\beta}_0) - (\mathbf{I}_{K+1}\otimes\boldsymbol{\alpha}^{\mathrm{T}})\mathbf{f}_i(\boldsymbol{\beta}_0)\right\} \\
&= \mathrm{var}\left(\begin{bmatrix} \frac{I(A=0)Y}{\pi(0,\mathbf{X})} - E\{m(0,\mathbf{X})\} \\ \vdots \\ \frac{I(A=K)Y}{\pi(K,\mathbf{X})} - E\{m(K,\mathbf{X})\} \end{bmatrix} - (\mathbf{I}_{K+1}\otimes\boldsymbol{\alpha}^{\mathrm{T}}) \begin{bmatrix} \left\{\frac{I(A=0)}{\pi(0,\mathbf{X})}-1\right\}\mathbf{B}(0,\mathbf{X}) \\ \vdots \\ \left\{\frac{I(A=k)}{\pi(K,\mathbf{X})}-1\right\}\mathbf{B}(K,\mathbf{X}) \end{bmatrix}\right) \\
&= \mathrm{var}\left(\begin{bmatrix} \frac{I(A=0)Y}{\pi(0,\mathbf{X})} - E\{m(0,\mathbf{X})\} \\ \vdots \\ \frac{I(A=K)Y}{\pi(K,\mathbf{X})} - E\{m(K,\mathbf{X})\} \end{bmatrix} - \begin{bmatrix} \left\{\frac{I(A=0)}{\pi(0,\mathbf{X})}-1\right\}m(0,\mathbf{X}) \\ \vdots \\ \left\{\frac{I(A=k)}{\pi(K,\mathbf{X})}-1\right\}m(K,\mathbf{X}) \end{bmatrix}\right) \\
&= \mathrm{var}\begin{Bmatrix} \frac{I(A=0)\{Y-m(0,\mathbf{X})\}}{\pi(0,\mathbf{X})} + m(0,\mathbf{X}) - E\{m(0,\mathbf{X})\} \\ \frac{I(A=1)\{Y-m(1,\mathbf{X})\}}{\pi(1,\mathbf{X})} + m(1,\mathbf{X}) - E\{m(1,\mathbf{X})\} \\ \vdots \\ \frac{I(A=K)\{Y-m(K,X)\}}{\pi(K,\mathbf{X})} + m(K,\mathbf{X}) - E\{m(K,\mathbf{X})\} \end{Bmatrix},
\end{aligned}
$$

i.e., the (k,l) entry of $\boldsymbol{\Sigma}$ is

$$
\boldsymbol{\Sigma}_{kl} = I(k=l)E\left\{\frac{v(k,\mathbf{X})}{\pi(k,\mathbf{X})}\right\} + E([m(k,\mathbf{X}) - E\{m(k,\mathbf{X})\}][m(l,\mathbf{X}) - E\{m(l,\mathbf{X})\}]).
$$

Compared to the semiparametric efficiency bound obtained in Section A.1.2 below, we see that the estimator is asymptotically efficient. □

A.1.2 Semiparametric efficiency bound

The original model can be written in general as

$$
f_{\mathbf{X},A,Y}(\mathbf{x},a,y) = f_{\mathbf{X}}(\mathbf{x})\prod_{k=0}^{K}[\pi(k,\mathbf{x})f_{\epsilon|(A,\mathbf{X})}\{y-m(k,\mathbf{x}),k,\mathbf{x}\}]^{I(a=k)}, \tag{A.1}
$$

where $\pi(k,\mathbf{x})$ satisfies $0<\pi(k,\mathbf{x})<1$, $\sum_{k=0}^{K}\pi(k,\mathbf{x})=1$ and $f_{\epsilon|(A,\mathbf{X})}\{y-m(k,\mathbf{x}),k,\mathbf{x}\}$ satisfies $\int f_{\epsilon|(A,\mathbf{X})}(\epsilon,k,\mathbf{x})d\epsilon = 1$ and $\int \epsilon f_{\epsilon|(A,\mathbf{X})}(\epsilon,k,\mathbf{x})d\epsilon = 0$ for all $k=0,\ldots,K$. The

parameter of interest is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^{\mathrm{T}}$, where $\theta_k = E\{m(k, \mathbf{X})\}$. Here, we sometimes write $\epsilon = y - m(a, \mathbf{x})$ for convenience. Consider an arbitrary parametric submodel

$$f_{\mathbf{X},A,Y}(\mathbf{x}, a, y, \boldsymbol{\delta}) = f_{\mathbf{X}}(\mathbf{x}, \boldsymbol{\zeta}) \prod_{k=0}^{K} [\pi(k, \mathbf{x}, \boldsymbol{\beta}) f_{\epsilon|(A,\mathbf{X})}\{y - m(k, \mathbf{x}, \boldsymbol{\alpha}), k, \mathbf{x}, \boldsymbol{\gamma}\}]^{I(a=k)},$$

where $\boldsymbol{\delta} = (\boldsymbol{\zeta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. We get the score function $\mathbf{S}_{\boldsymbol{\delta}} = (\mathbf{S}_{\boldsymbol{\zeta}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\alpha}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, where

$$\begin{aligned}
\mathbf{S}_{\boldsymbol{\zeta}} &= \frac{\partial f_{\mathbf{X}}(\mathbf{x}, \boldsymbol{\zeta}) / \partial \boldsymbol{\zeta}}{f_{\mathbf{X}}(\mathbf{x}, \boldsymbol{\zeta})}, \\
\mathbf{S}_{\boldsymbol{\beta}} &= \sum_{k=0}^{K} \left\{ I(A = k) \frac{\partial \pi(k, \mathbf{x}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}}{\pi(k, \mathbf{x}, \boldsymbol{\beta})} \right\}, \\
\mathbf{S}_{\boldsymbol{\alpha}} &= \sum_{k=0}^{K} I(A = k) \left[-\frac{\partial m(k, \mathbf{x}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \frac{\partial f_{\epsilon|(A,\mathbf{X})}\{y - m(k, \mathbf{x}, \boldsymbol{\alpha}), k, \mathbf{x}, \boldsymbol{\gamma}\} / \partial \{y - m(k, \mathbf{x}, \boldsymbol{\alpha})\}}{f_{\epsilon|(A,\mathbf{X})}\{y - m(k, \mathbf{x}, \boldsymbol{\alpha}), k, \mathbf{x}, \boldsymbol{\gamma}\}} \right], \\
\mathbf{S}_{\boldsymbol{\gamma}} &= \sum_{k=0}^{K} \left[I(A = k) \frac{\partial f_{\epsilon|(A,\mathbf{X})}\{y - m(k, \mathbf{x}, \boldsymbol{\alpha}), k, \mathbf{x}, \boldsymbol{\gamma}\} / \partial \boldsymbol{\gamma}}{f_{\epsilon|(A,\mathbf{X})}\{y - m(k, \mathbf{x}, \boldsymbol{\alpha}), k, \mathbf{x}, \boldsymbol{\gamma}\}} \right].
\end{aligned}$$

The tangent space of (A.1) is $\mathcal{T} = \mathcal{T}_{\boldsymbol{\zeta}} + \mathcal{T}_{\boldsymbol{\beta}} + \mathcal{T}_{\boldsymbol{\alpha}} + \mathcal{T}_{\boldsymbol{\gamma}}$, where

$$\begin{aligned}
\mathcal{T}_{\boldsymbol{\zeta}} &= [\mathbf{a}(\mathbf{X}) : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}], \\
\mathcal{T}_{\boldsymbol{\beta}} &= [\mathbf{a}(A, \mathbf{X}) : \sum_{k=0}^{K} \mathbf{a}(k, \mathbf{x}) \pi(k, \mathbf{x}) = \mathbf{0}], \\
\mathcal{T}_{\boldsymbol{\alpha}} &= \left[\mathbf{a}(A, \mathbf{X}) \frac{f'_{\epsilon|(A,\mathbf{X})}\{Y - m(A, \mathbf{X}), A, \mathbf{X}\}}{f_{\epsilon|(A,\mathbf{X})}\{Y - m(A, \mathbf{X}), A, \mathbf{X}\}} : \forall \mathbf{a}(A, \mathbf{X}) \right], \\
\mathcal{T}_{\boldsymbol{\gamma}} &= [\mathbf{a}(\epsilon, A, \mathbf{X}) : E\{\mathbf{a}(\epsilon, A, \mathbf{X}) \mid A, \mathbf{X}\} = \mathbf{0}, E\{\epsilon \mathbf{a}(\epsilon, A, \mathbf{X}) \mid A, \mathbf{X}\} = \mathbf{0}].
\end{aligned}$$

The parameter of interest in the submodel is

$$\boldsymbol{\theta}(\boldsymbol{\zeta}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = [E\{m(0, \mathbf{X}, \boldsymbol{\alpha})\}, \ldots, E\{m(K, \mathbf{X}, \boldsymbol{\alpha})\}]^{\mathrm{T}},$$

where

$$E\{m(k, \mathbf{X}, \boldsymbol{\alpha})\} = \int m(k, \mathbf{x}, \boldsymbol{\alpha}) f_{\mathbf{X}}(\mathbf{x}, \boldsymbol{\zeta}) d\mu(\mathbf{x}).$$

Thus,

$$\begin{aligned}
\frac{\partial \boldsymbol{\theta}(\boldsymbol{\zeta},\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} &= \left[E\left\{\frac{\partial m(0,\mathbf{X},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right\},\ldots,E\left\{\frac{\partial m(K,\mathbf{X},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right\}\right]^{\mathrm{T}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0},\\
\frac{\partial \boldsymbol{\theta}(\boldsymbol{\zeta},\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} &= \left[E\left\{m(0,\mathbf{X})\mathbf{S}_{\boldsymbol{\zeta}}\right\},\ldots,E\left\{m(K,\mathbf{X})\mathbf{S}_{\boldsymbol{\zeta}}\right\}\right]^{\mathrm{T}}\Big|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0},
\end{aligned}$$

while $\partial \boldsymbol{\theta}(\boldsymbol{\zeta},\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial \boldsymbol{\beta}^{\mathrm{T}} = \mathbf{0}$ and $\partial \boldsymbol{\theta}(\boldsymbol{\zeta},\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^{\mathrm{T}} = \mathbf{0}$.

Now consider

$$\boldsymbol{\phi} = \left[I(A=0)\frac{Y-m(0,\mathbf{X})}{\pi(0,\mathbf{X})} + m(0,\mathbf{X}_i),\ldots,I(A=K)\frac{Y-m(K,\mathbf{X})}{\pi(K,\mathbf{X})} + m(K,\mathbf{X}_i)\right]^{\mathrm{T}}.$$

Denote $\phi_k = I(A=k)\frac{Y-m(k,\mathbf{X})}{\pi(k,\mathbf{X})} + m(k,\mathbf{X}_i)$. We can easily verify that

$$\begin{aligned}
& E(\phi_k \mathbf{S}_{\boldsymbol{\beta}})\\
=\ & E\left[\left\{I(A=k)\frac{Y-m(k,\mathbf{X})}{\pi(k,\mathbf{X})} + m(k,\mathbf{X}_i)\right\}\left\{\sum_{l=0}^{K} I(A=l)\frac{\partial \pi(l,\mathbf{x},\boldsymbol{\beta})/\partial \boldsymbol{\beta}}{\pi(l,\mathbf{x},\boldsymbol{\beta})}\right\}\right]\\
=\ & E\left[\left\{I(A=k)\frac{Y-m(k,\mathbf{X})}{\pi(k,\mathbf{X})}\frac{\partial \pi(k,\mathbf{x},\boldsymbol{\beta})/\partial \boldsymbol{\beta}}{\pi(k,\mathbf{x},\boldsymbol{\beta})}\right\}\right]\\
& + E\left[m(k,\mathbf{X}_i)\left\{\sum_{l=0}^{K} I(A=l)\frac{\partial \pi(l,\mathbf{x},\boldsymbol{\beta})/\partial \boldsymbol{\beta}}{\pi(l,\mathbf{x},\boldsymbol{\beta})}\right\}\right]\\
=\ & E\left[\{Y^k - m(k,\mathbf{X})\}\frac{\partial \pi(k,\mathbf{x},\boldsymbol{\beta})/\partial \boldsymbol{\beta}}{\pi(k,\mathbf{x},\boldsymbol{\beta})}\right] + E\left[m(k,\mathbf{X}_i)\left\{\frac{\partial \sum_{l=0}^{K}\pi(l,\mathbf{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}\right]\\
=\ & \mathbf{0},
\end{aligned}$$

and

$$\begin{aligned}
& E(\phi_k \mathbf{S}_{\boldsymbol{\gamma}})\\
=\ & E\left(\left\{I(A=k)\frac{Y-m(k,\mathbf{X})}{\pi(k,\mathbf{X})} + m(k,\mathbf{X})\right\}\right.\\
& \left.\times\left[\sum_{l=0}^{K} I(A=l)\frac{\partial f_{\epsilon|(A,\mathbf{X})}\{Y-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}/\partial \boldsymbol{\gamma}}{f_{\epsilon|(A,\mathbf{X})}\{Y-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}}\right]\right)\\
=\ & E\left(I(A=k)\frac{Y-m(k,\mathbf{X})}{\pi(k,\mathbf{X})}\left[\sum_{l=0}^{K} I(A=l)\frac{\partial f_{\epsilon|(A,\mathbf{X})}\{Y-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}/\partial \boldsymbol{\gamma}}{f_{\epsilon|(A,\mathbf{X})}\{Y-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}}\right]\right)\\
& + E\left(m(k,\mathbf{X})\left[\sum_{l=0}^{K} I(A=l)\frac{\partial f_{\epsilon|(A,\mathbf{X})}\{Y-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}/\partial \boldsymbol{\gamma}}{f_{\epsilon|(A,\mathbf{X})}\{Y-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}}\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&= E\left\{\frac{\partial}{\partial\boldsymbol{\gamma}}\int \epsilon f_{\epsilon|(A,\mathbf{X})}(\epsilon,k,\mathbf{X},\boldsymbol{\gamma})d\epsilon\right\} + E\left[m(k,\mathbf{X})\left\{\sum_{l=0}^{K}\pi(l,\mathbf{X})\frac{\partial}{\partial\boldsymbol{\gamma}}\int f_{\epsilon|(A,\mathbf{X})}(\epsilon,l,\mathbf{X},\boldsymbol{\gamma})d\epsilon\right\}\right] \\
&= \mathbf{0}.
\end{aligned}
$$

Hence $E(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}) = \mathbf{0}$ and $E(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}}) = \mathbf{0}$. Further,

$$
\begin{aligned}
E(\phi_k\mathbf{S}_{\boldsymbol{\zeta}}) &= E\left[\left\{I(A=k)\frac{Y-m(k,\mathbf{X})}{\pi(k,\mathbf{X})} + m(k,\mathbf{X}_i)\right\}\frac{\partial f_{\mathbf{X}}(\mathbf{x},\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}}{f_{\mathbf{X}}(\mathbf{x},\boldsymbol{\zeta})}\right] \\
&= \mathbf{0} + E\left\{m(k,\mathbf{X}_i)\frac{\partial f_{\mathbf{X}}(\mathbf{x},\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}}{f_{\mathbf{X}}(\mathbf{x},\boldsymbol{\zeta})}\right\} \\
&= E\{m(k,\mathbf{X})\mathbf{S}_{\boldsymbol{\zeta}}(\mathbf{X},\boldsymbol{\zeta})\},
\end{aligned}
$$

and

$$
\begin{aligned}
&E(\phi_k\mathbf{S}_{\boldsymbol{\alpha}}) \\
&= E\left(\left\{I(A=k)\frac{Y-m(k,\mathbf{X})}{\pi(k,\mathbf{X})} + m(k,\mathbf{X})\right\}\right. \\
&\quad\left.\times\left[\sum_{l=0}^{K} I(A=l)\frac{-\partial m(l,\mathbf{X},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\frac{\partial f_{\epsilon|(A,\mathbf{X})}\{Y-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}/\partial\{Y-m(l,\mathbf{X},\boldsymbol{\alpha})\}}{f_{\epsilon|(A,\mathbf{X})}\{Y-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}}\right]\right) \\
&= E\left[\{Y^k-m(k,\mathbf{X})\}\frac{-\partial m(k,\mathbf{X},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right. \\
&\quad\left.\times\frac{\partial f_{\epsilon|(A,\mathbf{X})}\{Y^k-m(k,\mathbf{X},\boldsymbol{\alpha}),k,\mathbf{X},\boldsymbol{\gamma}\}/\partial\{Y^k-m(k,\mathbf{X},\boldsymbol{\alpha})\}}{f_{\epsilon|(A,\mathbf{X})}\{Y^k-m(k,\mathbf{X},\boldsymbol{\alpha}),k,\mathbf{X},\boldsymbol{\gamma}\}}\right] \\
&\quad+E\left(m(k,\mathbf{X})\left[\sum_{l=0}^{K}\pi(l,\mathbf{X})\frac{-\partial m(l,\mathbf{X},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right.\right. \\
&\quad\left.\left.\times\frac{\partial f_{\epsilon|(A,\mathbf{X})}\{Y^l-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}/\partial\{Y^l-m(l,\mathbf{X},\boldsymbol{\alpha})\}}{f_{\epsilon|(A,\mathbf{X})}\{Y-m(l,\mathbf{X},\boldsymbol{\alpha}),l,\mathbf{X},\boldsymbol{\gamma}\}}\right]\right) \\
&= E\left\{\frac{-\partial m(k,\mathbf{X},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\int\epsilon\frac{\partial f_{\epsilon|(A,\mathbf{X})}(\epsilon,k,\mathbf{X},\boldsymbol{\gamma})}{\partial\epsilon}d\epsilon\right\} \\
&\quad+E\left[m(k,\mathbf{X})\left\{\sum_{l=0}^{K}\pi(l,\mathbf{X})\frac{-\partial m(l,\mathbf{X},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\int\frac{\partial f_{\epsilon|(A,\mathbf{X})}(\epsilon,l,\mathbf{X},\boldsymbol{\gamma})}{\partial\epsilon}d\epsilon\right\}\right] \\
&= E\left\{\frac{\partial m(k,\mathbf{X},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right\},
\end{aligned}
$$

where $\boldsymbol{\alpha},\boldsymbol{\zeta}$ are evaluated at the true value $\boldsymbol{\alpha}_0,\boldsymbol{\zeta}_0$. Therefore,

$$
E(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\zeta}}^{\mathrm{T}}) = [E\{m(0,\mathbf{X})\mathbf{S}_{\boldsymbol{\zeta}}(\mathbf{X},\boldsymbol{\zeta})\},\ldots,E\{m(K,\mathbf{X})\mathbf{S}_{\boldsymbol{\zeta}}(\mathbf{X},\boldsymbol{\zeta})\}]^{\mathrm{T}} = \partial\boldsymbol{\theta}(\boldsymbol{\zeta},\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial\boldsymbol{\zeta}^{\mathrm{T}}.
$$

and

$$E(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\alpha}}^{\mathrm{T}}) = [E\{\partial m(0,\mathbf{x},\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\},\ldots,\partial E\{m(K,\mathbf{x},\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\}]^{\mathrm{T}} = \partial\boldsymbol{\theta}(\boldsymbol{\zeta},\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial\boldsymbol{\alpha}^{\mathrm{T}}.$$

Thus, $\boldsymbol{\phi}$ satisfies $E(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\delta}}^{\mathrm{T}}) = \partial\boldsymbol{\theta}(\boldsymbol{\delta})/\partial\boldsymbol{\delta}^{\mathrm{T}}$. Because the submodel is arbitrary, $\boldsymbol{\phi}$ is an influence function of $\boldsymbol{\theta}$. We now try to obtain $\Pi(\boldsymbol{\phi} \mid \mathcal{T})$ so we can obtain the efficient influence function. Further, we decompose $\boldsymbol{\phi}$ as $\boldsymbol{\phi} = (\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2 + \boldsymbol{\phi}_3 + \mathbf{c})$, where

$$\begin{aligned}
\boldsymbol{\phi}_1 &= \begin{pmatrix} \frac{I(A=0)}{\pi(0,\mathbf{X})}\left[Y - m(0,\mathbf{X}) + v(0,\mathbf{X})\frac{f'_{\epsilon|(A,\mathbf{X})}\{Y-m(0,\mathbf{X}),0,\mathbf{X}\}}{f_{\epsilon|(A,\mathbf{X})}\{Y-m(0,\mathbf{X}),0,\mathbf{X}\}}\right] \\ \vdots \\ \frac{I(A=K)}{\pi(K,\mathbf{X})}\left[Y - m(K,\mathbf{X}) + v(K,\mathbf{X})\frac{f'_{\epsilon|(A,\mathbf{X})}\{Y-m(K,\mathbf{X}),K,\mathbf{X}\}}{f_{\epsilon|(A,\mathbf{X})}\{Y-m(K,\mathbf{X}),K,\mathbf{X}\}}\right] \end{pmatrix}, \\
\boldsymbol{\phi}_2 &= -\begin{bmatrix} \frac{I(A=0)}{\pi(0,\mathbf{X})}v(0,\mathbf{X})\frac{f'_{\epsilon|(A,\mathbf{X})}\{Y-m(0,\mathbf{X}),0,\mathbf{X}\}}{f_{\epsilon|(A,\mathbf{X})}\{Y-m(0,\mathbf{X}),0,\mathbf{X}\}} \\ \vdots \\ \frac{I(A=K)}{\pi(K,\mathbf{X})}v(K,\mathbf{X})\frac{f'_{\epsilon|(A,\mathbf{X})}\{Y-m(K,\mathbf{X}),K,\mathbf{X}\}}{f_{\epsilon|(A,\mathbf{X})}\{Y-m(K,\mathbf{X}),K,\mathbf{X}\}} \end{bmatrix}, \\
\boldsymbol{\phi}_3 &= \begin{bmatrix} m(0,\mathbf{X}) - E\{m(0,\mathbf{X})\} \\ \vdots \\ m(K,\mathbf{X}) - E\{m(K,\mathbf{X})\} \end{bmatrix}
\end{aligned}$$

and $\mathbf{c} = [E\{m(0,\mathbf{X})\},\ldots,E\{m(K,\mathbf{X})\}]^{\mathrm{T}}$, where $v(k,\mathbf{X}) \equiv \mathrm{var}(Y^k \mid \mathbf{X}, A = k)$. We can verify that $\boldsymbol{\phi}_1 \in \mathcal{T}_{\boldsymbol{\gamma}}$, $\boldsymbol{\phi}_2 \in \mathcal{T}_{\boldsymbol{\alpha}}$, and $\boldsymbol{\phi}_3 \in \mathcal{T}_{\boldsymbol{\zeta}}$, while $\mathbf{c}$ is a constant. Then $\boldsymbol{\phi} - \mathbf{c}$ is the efficient influence function. Thus, the efficient variance is $\boldsymbol{\Sigma}_{\text{eff}} = \mathrm{var}(\boldsymbol{\phi})$, where the (k,l) entry of $\boldsymbol{\Sigma}_{\text{eff}}$ is

$$\boldsymbol{\Sigma}_{\text{eff,k,l}} = I(k=l)E\{v(k,\mathbf{X})/\pi(k,\mathbf{X})\} + E([m(k,\mathbf{X}) - E\{m(k,\mathbf{X})\}][m(l,\mathbf{X}) - E\{m(k,\mathbf{X})\}]).$$

When $K = 1$, this agrees with the special case corresponding to the binary treatments (Hahn 1998), and when $K > 1$, with earlier results (Cattaneo 2010).

A.2 Continuous treatments

We prove all results under a general weight function $w(A_j)$, where $w(A_j) = \sum_{i=1}^{n} K_l(A_i - A_j)$ in the main paper.

A.2.1 Convergence rate of $\widehat{\boldsymbol{\beta}}$

Proof of Lemma 2.3.1. From (2.14), $\boldsymbol{\beta}^*$ satisfies

$$\begin{aligned}
\mathbf{0} &= E_j\left(\left[E_i\left\{\frac{K_l(A_i-A_j)\pi'_{\boldsymbol{\beta}}(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}{\pi^2(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}\mathbf{B}^{\mathrm{T}}(A_j,\mathbf{X}_i)\right\}\right]\right.\\
&\quad\left.\times w(A_j)E_i\left[\left\{\frac{K_l(A_i-A_j)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right)\\
&= E_j\left(\left[E_i\left\{\frac{\pi_0(A_j,\mathbf{X}_i)\pi'_{\boldsymbol{\beta}}(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}{\pi^2(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}\mathbf{B}^{\mathrm{T}}(A_j,\mathbf{X}_i)\right\}\right]\right.\\
&\quad\left.\times w(A_j)E_i\left[\left\{\frac{\pi_0(A_j,\mathbf{X}_i)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right)+O(l^2),\\
&= E_j\left(\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\right.\\
&\quad\left.\times E_i\left[\left\{\frac{\pi_0(A_j,\mathbf{X}_i)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right)+O(l^2)\\
&= E_j\left(\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)E_i\left[\left\{\frac{\pi_0(A_j,\mathbf{X}_i)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right)+O(l^2),
\end{aligned}$$

where

$$\mathbf{U}(a_j,\boldsymbol{\beta}^*)\equiv E\left\{\frac{\pi_0(a_j,\mathbf{X})\pi'_{\boldsymbol{\beta}}(a_j,\mathbf{X},\boldsymbol{\beta}^*)}{\pi^2(a_j,\mathbf{X},\boldsymbol{\beta}^*)}\mathbf{B}(a_j,\mathbf{X})^{\mathrm{T}}\right\}.$$

We now investigate the convergence rate of $\widehat{\boldsymbol{\beta}}$ from (2.12). We note that

$$\begin{aligned}
\mathbf{0} &= \frac{1}{n}\sum_{j=1}^{n}\\
&\quad\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{K_l(A_i-A_j)\pi'_{\boldsymbol{\beta}}(A_j,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}{\pi^2(A_j,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\mathbf{B}^{\mathrm{T}}(A_j,\mathbf{X}_i)\right\}\right]\\
&\quad\times w(A_j)\left(\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[\left\{\frac{K_l(A_i-A_j)}{\pi(A_j,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right)\\
&= \frac{1}{n}\sum_{j=1}^{n}\\
&\quad\left[E\left\{\frac{K_l(A_i-a_j)\pi'_{\boldsymbol{\beta}}(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}{\pi^2(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}\mathbf{B}(a_j,\mathbf{X}_i)^{\mathrm{T}}\right\}+o_p(1)\right]\\
&\quad\times w(A_j)\left(\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[\left\{\frac{K_l(A_i-A_j)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right)
\end{aligned}$$

$$
\begin{aligned}
& -\frac{1}{n}\sum_{j=1}^{n} \\
& \left(\left[E\left\{\frac{K_l(A_i-a_j)\pi'_{\boldsymbol{\beta}}(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}{\pi^2(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}\mathbf{B}(a_j,\mathbf{X}_i)^{\mathrm{T}}\right\}\right]^{\otimes 2}\right. \\
& \left. w(A_j)+o_p(1)\right) n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) \\
= \;& \frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j) \\
& \times\left\{\frac{K_l(A_i-A_j)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i) \\
& -E\left[\{\mathbf{U}(A_j,\boldsymbol{\beta}^*)\}^{\otimes 2}w(A_j)\right] n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)+o_p(1).
\end{aligned}
$$

We have

$$
\begin{aligned}
& \frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\left\{\frac{K_l(A_i-A_j)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i) \\
= \;& \frac{1}{n^{1/2}}\sum_{j=1}^{n}\mathbf{U}(a_j,\boldsymbol{\beta}^*)w(a_j)E_i\left[\left\{\frac{K_l(A_i-a_j)}{\pi(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(a_j,\mathbf{X}_i)\right] \\
& +\frac{1}{n^{1/2}}\sum_{i=1}^{n}E_j\left[\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\left\{\frac{K_l(a_i-A_j)}{\pi(A_j,\mathbf{x}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{x}_i)\right] \\
& -n^{1/2}E_{ij}\left[\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\left\{\frac{K_l(A_i-A_j)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]+o_p(1) \\
= \;& \frac{1}{n^{1/2}}\sum_{j=1}^{n}\mathbf{U}(a_j,\boldsymbol{\beta}^*)w(a_j)E_i\left\{\frac{\pi_0(a_j,\mathbf{X}_i)}{\pi(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}\mathbf{B}(a_j,\mathbf{X}_i)-\mathbf{B}(a_j,\mathbf{X}_i)\right\} \\
& +\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[f_A(a_i)\mathbf{U}(a_i,\boldsymbol{\beta}^*)w(a_i)\frac{\mathbf{B}(a_i,\mathbf{x}_i)}{\pi(a_i,\mathbf{x}_i,\boldsymbol{\beta}^*)}-E_j\left\{\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\mathbf{B}(A_j,\mathbf{x}_i)\right\}\right] \\
& -n^{1/2}E_j\left[\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)E_i\left\{\frac{\pi_0(A_j,\mathbf{X}_i)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}\mathbf{B}(A_j,\mathbf{X}_i)-\mathbf{B}(A_j,\mathbf{X}_i)\right\}\right] \\
& +o_p(1)+O_p(n^{1/2}l^2).
\end{aligned}
$$

Thus, when $nl^4\to 0$, we get

$$
\begin{aligned}
& E\left\{\mathbf{U}(A_j,\boldsymbol{\beta}^*)^{\otimes 2}w(A_j)\right\} n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) \\
= \;& \frac{1}{n^{1/2}}\sum_{j=1}^{n}\mathbf{U}(a_j,\boldsymbol{\beta}^*)w(a_j)E_i\left[\left\{\frac{\pi_0(a_j,\mathbf{X}_i)}{\pi(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(a_j,\mathbf{X}_i)\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[f_A(a_i)\mathbf{U}(a_i,\boldsymbol{\beta}^*)w(a_i)\frac{\mathbf{B}(a_i,\mathbf{x}_i)}{\pi(a_i,\mathbf{x}_i,\boldsymbol{\beta}^*)}-E_j\{\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\mathbf{B}(A_j,\mathbf{x}_i)\}\right]\\
&-n^{1/2}E_j\left(\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)E_i\left[\left\{\frac{\pi_0(A_j,\mathbf{X}_i)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right)\\
&+o_p(1)+O_p(n^{1/2}l^2)\\
=\;&\frac{1}{n^{1/2}}\sum_{i=1}^{n}\mathbf{U}(a_i,\boldsymbol{\beta}^*)w(a_i)E_k\left[\left\{\frac{\pi_0(a_i,\mathbf{X}_k)}{\pi(a_i,\mathbf{X}_k,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(a_i,\mathbf{X}_k)\right]\\
&+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[f_A(a_i)\mathbf{U}(a_i,\boldsymbol{\beta}^*)w(a_i)\frac{\mathbf{B}(a_i,\mathbf{x}_i)}{\pi(a_i,\mathbf{x}_i,\boldsymbol{\beta}^*)}-E_j\left\{\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\mathbf{B}(A_j,\mathbf{x}_i)\right\}\right]\\
&+o_p(1).
\end{aligned}
$$

Obviously,

$$
E_i\left(\mathbf{U}(A_i,\boldsymbol{\beta}^*)w(A_i)E_k\left[\left\{\frac{\pi_0(A_i,\mathbf{X}_k)}{\pi(A_i,\mathbf{X}_k,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_i,\mathbf{X}_k)\right]\right)=O(l^2)
$$

due to the definition of $\boldsymbol{\beta}^*$. Further, we can verify that

$$
\begin{aligned}
&E_i\left[f_A(A_i)\mathbf{U}(A_i,\boldsymbol{\beta}^*)w(A_i)\frac{\mathbf{B}(A_i,\mathbf{X}_i)}{\pi(A_i,\mathbf{X}_i,\boldsymbol{\beta}^*)}-E_j\left\{\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\mathbf{B}(A_j,\mathbf{X}_i)\right\}\right]\\
=\;&E_{i,j}\left\{\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\frac{\pi_0(A_j,\mathbf{X}_i)\mathbf{B}(A_j,\mathbf{X}_i)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right\}-E_{i,j}\left\{\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)\mathbf{B}(A_j,\mathbf{X}_i)\right\}\\
=\;&E_j\left(\mathbf{U}(A_j,\boldsymbol{\beta}^*)w(A_j)E_i\left[\left\{\frac{\pi_0(A_j,\mathbf{X}_i)}{\pi(A_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}\mathbf{B}(A_j,\mathbf{X}_i)\right]\right)\\
=\;&O(l^2)
\end{aligned}
$$

also due to the definition of $\boldsymbol{\beta}^*$. Thus, as long as $nl^4\to 0$, $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*=O_p(n^{-1/2})$. □

A.2.2 Robustness, asymptotic bias, and variance

Proof of Theorem 2.3.1. When model (2.9) holds, we can easily check that the expectation of the left hand side of (2.11) at the true parameter value $\boldsymbol{\beta}_0$ and any function $m(a,\mathbf{x})=\mathbf{B}(a,\mathbf{x})^{\mathrm{T}}\boldsymbol{\gamma}$ satisfies

$$
\begin{aligned}
&E\left[\left\{\frac{K_h(A_i-a)}{\pi_0(a,\mathbf{X}_i)}-1\right\}m(a,\mathbf{X}_i)\right]\\
=\;&E\left[\left\{\frac{E\{K_h(A_i-a)\mid\mathbf{X}_i\}}{\pi_0(a,\mathbf{X}_i)}-1\right\}m(a,\mathbf{X}_i)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\left\{\frac{\int K_h(A_i-a)\pi_0(A_i,\mathbf{X}_i)dA_i}{\pi_0(a,\mathbf{X}_i)}-1\right\}m(a,\mathbf{X}_i)\right]\\
&= E\left[\left\{\frac{\int K(t)\pi_0(a+ht,\mathbf{X}_i)dt}{\pi_0(a,\mathbf{X}_i)}-1\right\}m(a,\mathbf{X}_i)\right]\\
&= E\left[\left\{\frac{\int K(t)\pi_0(a,\mathbf{X}_i)dt}{\pi_0(a,\mathbf{X}_i)}-1\right\}m(a,\mathbf{X}_i)\right]+O(h^2)\\
&= O(h^2).
\end{aligned}
$$

Thus, because the nonparametric estimation convergence rate is slower than $O_p(n^{-1/2})$, by Lemma 2.3.1 we can fix $\boldsymbol{\beta}$ at $\boldsymbol{\beta}_0$ in the following analysis, and the first order bias and variance property of $\widehat{\theta}(a)$ will not be affected.

Hence, for (2.13), we have

$$
\begin{aligned}
E\{\widehat{\theta}(a)\} &= E\left\{\frac{K_h(A_i-a)Y_i}{\pi_0(a,\mathbf{X}_i)}\right\}+O(n^{-1/2})\\
&= E\left\{\frac{K_h(A_i-a)Y_i(A_i)}{\pi_0(a,\mathbf{X}_i)}\right\}+O(n^{-1/2})\\
&= E\left\{\frac{K_h(A_i-a)m(A_i,\mathbf{X}_i)}{\pi_0(a,\mathbf{X}_i)}\right\}+O(n^{-1/2})\\
&= E\left[m(a,\mathbf{X}_i)+\frac{\partial^2\{\pi_0(a,\mathbf{X}_i)m(a,\mathbf{X}_i)\}}{\pi_0(a,\mathbf{X}_i)\partial a^2}\frac{h^2}{2}\int t^2K(t)dt\right]\\
&\quad +O(h^4+n^{-1/2})\\
&= \theta(a)+E\left[\frac{\partial^2\{\pi_0(a,\mathbf{X}_i)m(a,\mathbf{X}_i)\}}{\pi_0(a,\mathbf{X}_i)\partial a^2}\right]\frac{h^2}{2}\int t^2K(t)dt+O(h^4+n^{-1/2}).
\end{aligned}
$$

The variance is calculated as

$$
\mathrm{var}\{\widehat{\theta}(a)\}=\mathrm{var}\left[n^{-1}\sum_{i=1}^{n}\left\{\frac{K_h(A_i-a)Y_i}{\pi_0(a,\mathbf{X}_i)}\right\}+O_p(n^{-1/2})\right].
$$

Now, recall that the variance of $Y_i(A_i)$ conditional on $\mathbf{X}_i, A_i$ is denoted $\sigma^2(A_i,\mathbf{X}_i)$, then

$$
\begin{aligned}
&E\left[\left\{\frac{K_h(A_i-a)Y_i}{\pi_0(a,\mathbf{X}_i)}\right\}^2\right]\\
&= E\left[\left\{\frac{K_h(A_i-a)}{\pi_0(a,\mathbf{X}_i)}\right\}^2\{m^2(A_i,\mathbf{X}_i)+\sigma^2(A_i,\mathbf{X}_i)\}\right]\\
&= \frac{\int K^2(t)dt}{h}E\left\{\frac{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)}{\pi_0(a,\mathbf{X}_i)}\right\}+O(h).
\end{aligned}
$$

Thus,

$$\begin{aligned}\text{var}\{\widehat{\theta}(a)\} &= \text{var}\left[n^{-1}\sum_{i=1}^{n}\left\{\frac{K_h(A_i-a)Y_i}{\pi_0(a,\mathbf{X}_i)}\right\}+O_p(n^{-1/2})\right]\\ &= \frac{\int K^2(t)dt}{nh}E\left\{\frac{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)}{\pi_0(a,\mathbf{X}_i)}\right\}\\ &\quad +O(n^{-1}h+n^{-1}+n^{-1}h^{-1/2}).\end{aligned}$$

The asymptotic normality is shown in Section A.2.3 below. □

Proof of Theorem 2.3.2. When model (2.10) is correct, then $\widehat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}^*$ at root-n rate (Lemma 2.3.1). Thus,

$$\begin{aligned}E\{\widehat{\theta}(a)\} &= E\left\{\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right\}+O(n^{-1/2})\\ &= E\left\{\frac{K_h(A_i-a)Y_i(A_i)}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right\}+O(n^{-1/2})\\ &= E\left[\frac{K_h(A_i-a)m(A_i,\mathbf{X}_i)}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right]+O(n^{-1/2})\\ &= E\left[\frac{\pi_0(a,\mathbf{X}_i)m(a,\mathbf{X}_i)}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right]+\frac{\int t^2K(t)dt}{2}h^2\\ &\quad \times E\left[\frac{\partial^2\{m(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)\partial a^2}\right]+O(n^{-1/2})\\ &= E\left[\left\{\frac{\pi_0(a,\mathbf{X}_i)}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}-1\right\}m(a,\mathbf{X}_i)\right]+E\{m(a,\mathbf{X}_i)\}\\ &\quad +\frac{\int t^2K(t)dt}{2}h^2E\left[\frac{\partial^2\{m(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)\partial a^2}\right]+O(n^{-1/2})\\ &= E\{m(a,\mathbf{X}_i)\}+\frac{\int t^2K(t)dt}{2}h^2E\left[\frac{\partial^2\{m(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)\partial a^2}\right]+O(n^{-1/2}).\end{aligned}$$

The variance is calculated as

$$\text{var}\{\widehat{\theta}(a)\}=\text{var}\left[n^{-1}\sum_{i=1}^{n}\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}+O_p(n^{-1/2})\right].$$

Then

$$E\left[\left\{\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right\}^2\right]$$

$$
\begin{aligned}
&= E\left[\left\{\frac{K_h(A_i-a)}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right\}^2\{m^2(A_i,\mathbf{X}_i)+\sigma^2(A_i,\mathbf{X}_i)\}\right]\\
&= \frac{\int K^2(t)dt}{h}E\left[\frac{\pi_0(a,\mathbf{X}_i)\{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\}}{\pi^2(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right]+O(h).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathrm{var}\{\widehat{\theta}(a)\} &= \mathrm{var}\left\{n^{-1}\sum_{i=1}^n\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}+O_p(n^{-1/2})\right\}\\
&= \frac{\int K^2(t)dt}{nh}E\left[\frac{\pi_0(a,\mathbf{X}_i)\{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\}}{\pi^2(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right]\\
&\quad +O(n^{-1}h+n^{-1}+n^{-1}h^{-1/2}).
\end{aligned}
$$

The asymptotic normality is shown in Section A.2.3 below. □

Proof of Theorem 2.3.3.

$$
\begin{aligned}
&\mathrm{cov}\left\{n^{-1}\sum_{i=1}^n\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})},n^{-1}\sum_{i=1}^n\frac{K_h(A_i-b)Y_i}{\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}\\
=\ & E\left[\left\{n^{-1}\sum_{i=1}^n\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}\left\{n^{-1}\sum_{i=1}^n\frac{K_h(A_i-b)Y_i}{\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}\right]\\
&-E\left\{n^{-1}\sum_{i=1}^n\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}E\left\{n^{-1}\sum_{i=1}^n\frac{K_h(A_i-b)Y_i}{\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}\\
=\ & n^{-2}\sum_{i=1}^n E\left\{\frac{K_h(A_i-a)K_h(A_i-b)Y_i^2}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}+n^{-2}\sum_{i\neq j,i.j=1}^n E\left\{\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\frac{K_h(A_j-b)Y_j}{\pi(b,\mathbf{X}_j,\widehat{\boldsymbol{\beta}})}\right\}\\
&-E\left\{\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}E\left\{\frac{K_h(A_i-b)Y_i}{\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}\\
=\ & n^{-1}E\left\{\frac{K_h(A_i-a)K_h(A_i-b)Y_i^2}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}-n^{-1}E\left\{\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}E\left\{\frac{K_h(A_i-b)Y_i}{\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}\\
=\ & n^{-1}E\left\{\frac{K_h(A_i-a)K_h(A_i-b)Y_i^2}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}-n^{-1}E\{\widehat{\theta}(a)\}E\{\widehat{\theta}(b)\}\\
=\ & n^{-1}E\left\{\frac{K_h(A_i-a)K_h(A_i-b)Y_i^2}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\}-n^{-1}\{\theta(a)\theta(b)+O(h^2)\}.
\end{aligned}
$$

When a and b are sufficiently close, so that $c \equiv (a-b)/h \in (-2,1)$, we have

$$
\begin{aligned}
& E\left\{\frac{K_h(A_i-a)K_h(A_i-b)Y_i^2}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\} \\
= \; & E\left\{\frac{K_h(A_i-a)K_h(A_i-b)\{m^2(A_i,\mathbf{X}_i)+\sigma^2(A_i,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\right\} \\
= \; & h^{-1}E\int_0^1 \frac{K(t)K(t+c)\{m^2(a+ht,\mathbf{X}_i)+\sigma^2(a+ht,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\pi_0(a+ht,\mathbf{X}_i)dt \\
= \; & h^{-1}E\int_0^1 \frac{K(t)K(t+c)\{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}\pi_0(a,\mathbf{X}_i)dt \\
& +E\int_0^1 K(t)K(t+c)\left\{2m(a,\mathbf{X}_i)m'_a(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i)+m^2(a,\mathbf{X}_i)\pi'_{0a}(a,\mathbf{X}_i)\right. \\
& \left.+2\sigma(a,\mathbf{X}_i)\sigma'_a(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\pi'_{0a}(a,\mathbf{X}_i)\right\} t/\{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\pi(b,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})\}dt \\
& +O(h) \\
= \; & h^{-1}E\int_0^1 \frac{K(t)K(t+c)\{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)\pi(b,\mathbf{X}_i,\boldsymbol{\beta}^*)}\pi_0(a,\mathbf{X}_i)dt \\
& +E\int_0^1 K(t)K(t+c)\left\{2m(a,\mathbf{X}_i)m'_a(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i)+m^2(a,\mathbf{X}_i)\pi'_{0a}(a,\mathbf{X}_i)\right. \\
& \left.+2\sigma(a,\mathbf{X}_i)\sigma'_a(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\pi'_{0a}(a,\mathbf{X}_i)\right\} t/\{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)\pi(b,\mathbf{X}_i,\boldsymbol{\beta}^*)\}dt \\
& +O(h+h^{-1}n^{-1/2}).
\end{aligned}
$$

Note that when $c \notin (-2,1)$, $K(t)K(t+c) = 0$ for all $t \notin [-1,1]$ hence the above expression still holds. Thus, we obtain

$$
\begin{aligned}
& \mathrm{cov}\{\widehat{\theta}(a),\widehat{\theta}(b)\} \\
= \; & (nh)^{-1}E\int_0^1 \frac{K(t)K(t+c)\{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)\pi(b,\mathbf{X}_i,\boldsymbol{\beta}^*)}\pi_0(a,\mathbf{X}_i)dt \\
& +n^{-1}E\int_0^1 K(t)K(t+c)\left\{2m(a,\mathbf{X}_i)m'_a(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i)+m^2(a,\mathbf{X}_i)\pi'_{0a}(a,\mathbf{X}_i)\right. \\
& \left.+2\sigma(a,\mathbf{X}_i)\sigma'_a(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\pi'_{0a}(a,\mathbf{X}_i)\right\} t/\{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)\pi(b,\mathbf{X}_i,\boldsymbol{\beta}^*)\}dt \\
& -n^{-1}\theta(a)\theta(b)+O(n^{-1}h+h^{-1}n^{-3/2}).
\end{aligned}
$$

The asymptotic normality is shown in Section A.2.3 below. □

A.2.3 Asymptotic distribution of $\widehat{\theta}(a)$

Proof of asymptotic normality, Theorems 2.3.1-2.3.3. When (2.9) is correct, define

$$\text{bias}\{\widehat{\theta}(a)\} = \frac{h^2}{2} E\left[\frac{\partial^2\{\pi_0(a,\mathbf{X}_i)m(a,\mathbf{X}_i)\}}{\pi_0(a,\mathbf{X}_i)\partial a^2}\right]\int t^2K(t)dt.$$

On the other hand, when (2.10) is correct, define

$$\text{bias}\{\widehat{\theta}(a)\} \;=\; \frac{h^2}{2} E\left[\frac{\partial^2\{\pi_0(a,\mathbf{X}_i)m(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)\partial a^2}\right]\int t^2K(t)dt.$$

Regardless (2.9) or (2.10) is correct, define

$$\text{var}(\widehat{\theta}) = \frac{\int K^2(t)dt}{nh} E\left[\frac{\pi_0(a,\mathbf{X}_i)\{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\}}{\pi^2(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right].$$

Note that when (2.9) is correct, it degenerates to

$$\text{var}\{\widehat{\theta}(a)\} = \frac{\int K^2(t)dt}{nh} E\left\{\frac{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)}{\pi_0(a,\mathbf{X}_i)}\right\}.$$

Then

$$\begin{aligned}
&\left[\widehat{\theta}(a)-\theta(a)-\text{bias}\{\widehat{\theta}(a)\}\right]\\
=\;& n^{-1}\sum_{i=1}^{n}\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)} - \theta(a) - \text{bias}\{\widehat{\theta}(a)\} + O_p(n^{-1/2})\\
=\;& n^{-1}\sum_{i=1}^{n}\left[\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)} - E\left\{\frac{K_h(A_i-a)Y_i}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right\}\right] + O_p(h^4+n^{-1/2}).
\end{aligned}$$

Thus, when $n\to\infty$, following the variance result, we get that

$$\sqrt{nh}\left[\widehat{\theta}(a)-\theta(a)-\text{bias}\{\widehat{\theta}(a)\}\right]$$

converges to a normal distribution with mean zero and variance $nh\text{var}\{\widehat{\theta}(a)\}$.

Consider an arbitrary linear combination $\sum_{j=1}^{J} c_j\widehat{\theta}(a_j)$. Then

$$\left[\sum_{j=1}^{J} c_j\widehat{\theta}(a_j) - \sum_{j=1}^{J} c_j\theta(a_j) - \text{bias}\left\{\sum_{j=1}^{J} c_j\widehat{\theta}(a_j)\right\}\right]$$

$$
\begin{aligned}
&= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J} c_j \frac{K_h(A_i-a_j)Y_i}{\pi(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)} - \sum_{j=1}^{J} c_j\theta(a_j) - \sum_{j=1}^{J} c_j \text{bias}\{\widehat{\theta}(a)\} + O_p(n^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J} c_j \left[\frac{K_h(A_i-a_j)Y_i}{\pi(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)} - E\left\{\frac{K_h(A_i-a_j)Y_i}{\pi(a_j,\mathbf{X}_i,\boldsymbol{\beta}^*)}\right\}\right] + O_p(h^4+n^{-1/2})
\end{aligned}
$$

converges to a normal distribution with mean zero. To compute its variance, we compute $\text{cov}\{\widehat{\theta}(a),\widehat{\theta}(b)\}$ for arbitrary a, b below.

Let $\text{cov}\{\widehat{\theta}(a),\widehat{\theta}(b)\}$ be given as the leading term in (2.15). Note that when (2.9) is correct, it degenerates to

$$
\begin{aligned}
&\text{cov}\{\widehat{\theta}(a),\widehat{\theta}(b)\} \\
= \;& (nh)^{-1}E\int_0^1 \frac{K(t)K(t+c)\{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\}}{\pi_0(b,\mathbf{X}_i)}dt \\
& + n^{-1}E\int_0^1 K(t)K(t+c)\left\{2m(a,\mathbf{X}_i)m'_a(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i) + m^2(a,\mathbf{X}_i)\pi'_{0a}(a,\mathbf{X}_i)\right. \\
& \left. + 2\sigma(a,\mathbf{X}_i)\sigma'_a(a,\mathbf{X}_i)\pi_0(a,\mathbf{X}_i) + \sigma^2(a,\mathbf{X}_i)\pi'_{0a}(a,\mathbf{X}_i)\right\} t/\{\pi_0(a,\mathbf{X}_i)\pi_0(b,\mathbf{X}_i)\}dt \\
& - n^{-1}\theta(a)\theta(b).
\end{aligned}
$$

Here $c=(a-b)/h$. Then the above analysis leads to that $\widehat{\theta}(a)-\theta(a)$ is asymptotically a Gaussian process with mean given by $\text{bias}\{\widehat{\theta}(a)\}$ and variance-covariance function given in $\text{cov}\{\widehat{\theta}(a),\widehat{\theta}(b)\}$. □

A.2.4 Variance estimation

By Theorem 2.3.2,

$$
\begin{aligned}
&\text{var}\{\widehat{\theta}(a)\} \\
= \;& (nh)^{-1}E\int_0^1 \frac{K(t)^2\{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)^2}\pi_0(a,\mathbf{X}_i)dt + O(n^{-1}+n^{-1}h+n^{-3/2}h^{-1}) \\
= \;& \frac{\int_0^1 K(t)^2dt}{nh}E\left[\frac{\pi_0(a,\mathbf{X}_i)\{m^2(a,\mathbf{X}_i)+\sigma^2(a,\mathbf{X}_i)\}}{\pi(a,\mathbf{X}_i,\boldsymbol{\beta}^*)^2}\right] + O(n^{-1}+n^{-1}h+n^{-3/2}h^{-1}).
\end{aligned}
$$

Thus, an estimator of this variance is obtained as

$$
\widehat{\text{var}}\{\widehat{\theta}(a)\} = \frac{\int_0^1 K(t)^2dt}{nh}n^{-1}\sum_{i=1}^{n}\left[\frac{K_h(A_i-a)\{\widehat{m}^2(A_i,\mathbf{X}_i)+\{Y_i-\widehat{m}(A_i,\mathbf{X}_i)\}^2\}}{\pi(a,\mathbf{X}_i,\widehat{\boldsymbol{\beta}})^2}\right].
$$

Let $\widehat{m}(A_i, \mathbf{X}_i) = Y_i$. Then the above estimator becomes

$$\widehat{\text{var}}\{\widehat{\theta}(a)\} = \frac{\int_0^1 K(t)^2 dt}{nh} n^{-1} \sum_{i=1}^{n} \frac{K_h(A_i - a) Y_i^2}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})^2}.$$

Its expectation is

$$\begin{aligned}
& \frac{\int_0^1 K(t)^2 dt}{nh} E\left\{ n^{-1} \sum_{i=1}^{n} \frac{K_h(A_i - a) Y_i^2}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})^2} \right\} \\
= & \frac{\int_0^1 K(t)^2 dt}{nh} E\left\{ n^{-1} \sum_{i=1}^{n} \frac{K_h(A_i - a) Y_i^2}{\pi(a, \mathbf{X}_i, \boldsymbol{\beta}^*)^2} + O(n^{-1/2}) \right\} \\
= & \frac{\int_0^1 K(t)^2 dt}{nh} E\left\{ \frac{K_h(A_i - a) Y_i^2}{\pi(a, \mathbf{X}_i, \boldsymbol{\beta}^*)^2} \right\} + O(n^{-3/2} h^{-1}) \\
= & \frac{\int_0^1 K(t)^2 dt}{nh} E\left[\frac{K_h(A_i - a) \left\{ m^2(A_i, \mathbf{X}_i) + \sigma^2(A_i, \mathbf{X}_i) \right\}}{\pi(a, \mathbf{X}_i, \boldsymbol{\beta}^*)^2} \right] + O(n^{-3/2} h^{-1}) \\
= & \frac{\int_0^1 K(t)^2 dt}{nh} \left(E\left[\frac{\pi_0(a, \mathbf{X}_i) \left\{ m^2(a, \mathbf{X}_i) + \sigma^2(a, \mathbf{X}_i) \right\}}{\pi(a, \mathbf{X}_i, \boldsymbol{\beta}^*)^2} \right] \right. \\
& \left. + \frac{\int t^2 K(t) dt}{2} h^2 \times E\left[\frac{\partial^2}{\partial a^2} \frac{\pi_0(a, \mathbf{X}_i) \left\{ m^2(a, \mathbf{X}_i) + \sigma^2(a, \mathbf{X}_i) \right\}}{\pi(a, \mathbf{X}_i, \boldsymbol{\beta}^*)^2} \right] \right) + O(n^{-3/2} h^{-1}) \\
= & \frac{\int_0^1 K(t)^2 dt}{nh} E\left[\frac{\pi_0(a, \mathbf{X}_i) \left\{ m^2(a, \mathbf{X}_i) + \sigma^2(a, \mathbf{X}_i) \right\}}{\pi(a, \mathbf{X}_i, \boldsymbol{\beta}^*)^2} \right] + O(n^{-1} h + n^{-3/2} h^{-1}).
\end{aligned}$$

Following Remark 2.3.1, an alternative variance estimator is:

$$\widehat{\text{var}}\{\widehat{\theta}(a)\} = \frac{\int_0^1 K(t)^2 dt}{nh} \left\{ \sum_{i=1}^{n} \frac{K_h(A_i - a)}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})} \right\}^{-1} \sum_{i=1}^{n} \frac{K_h(A_i - a) Y_i^2}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})^2}.$$

Similarly, we can show from Theorem 2.3.3 that the covariance can be estimated by

$$\widehat{\text{cov}}\{\widehat{\theta}(a), \widehat{\theta}(b)\} = \frac{\int_0^1 K(t) K(t + c) dt}{nh} \left\{ \sum_{i=1}^{n} \frac{K_h(A_i - a)}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})} \right\}^{-1} \sum_{i=1}^{n} \frac{K_h(A_i - a) Y_i^2}{\pi(a, \mathbf{X}_i, \widehat{\boldsymbol{\beta}}) \pi(b, \mathbf{X}_i, \widehat{\boldsymbol{\beta}})},$$

where $c \equiv (a - b)/h$.

Appendix B
Supplement to Chapter 3

B.1 Derivation of the efficiency bound of marginal effect estimation

Consider an arbitrary parametric submodel

$$f_{\mathbf{X},Y}(y,\mathbf{x},\boldsymbol{\theta}) = f_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha})\frac{\exp\left\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}+c(y,\boldsymbol{\gamma})\right\}}{\int\exp\left\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}+c(y,\boldsymbol{\gamma})\right\}d\mu(y)}, \tag{B.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. We can verify that the score functions associated with an arbitrary $\boldsymbol{\theta}$ are $\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x}) = \mathbf{a}(\mathbf{x})$, where $\mathbf{a}(\mathbf{x})$ can be any function that satisfies $E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}$, $\mathbf{S}_{\boldsymbol{\beta}}(y,\mathbf{x}) = \mathbf{x}\{y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}$, and $\mathbf{S}_{\boldsymbol{\gamma}}(y,\mathbf{x}) = \mathbf{a}(y) - E\{\mathbf{a}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}$, where $\mathbf{a}(y)$ can be any function. We can verify that $\boldsymbol{\xi} = \boldsymbol{\beta}E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}$ and

$$\begin{aligned}
\frac{\partial\boldsymbol{\xi}}{\partial\boldsymbol{\alpha}^{\mathrm{T}}} &= \boldsymbol{\beta}E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{a}^{\mathrm{T}}(\mathbf{X})\},\\
\frac{\partial\boldsymbol{\xi}}{\partial\boldsymbol{\beta}^{\mathrm{T}}} &= E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} + \boldsymbol{\beta}E[\mathbf{X}^{\mathrm{T}}\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}^3],\\
\frac{\partial\boldsymbol{\xi}}{\partial\boldsymbol{\gamma}^{\mathrm{T}}} &= \boldsymbol{\beta}E\left([\mathbf{a}(Y) - E\{\mathbf{a}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}}\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}^2\right).
\end{aligned}$$

Thus, following Bickel et al. (1998) and Tsiatis (2006), a possible influence function is

$$\begin{aligned}
&\boldsymbol{\phi}(y,\mathbf{x})\\
=\ &\boldsymbol{\beta}v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) - \boldsymbol{\beta}E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} + \boldsymbol{\beta}\{y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}^2\\
&-\boldsymbol{\beta}E[\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}^2 \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}] + \mathbf{B}E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}[\mathbf{b}(y,\mathbf{x}) - E\{\mathbf{b}(Y,\mathbf{x}) \mid \mathbf{x}\}],
\end{aligned}$$

where

$$\mathbf{B} = \left(E[\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{b}(Y, \mathbf{X})\mathbf{X}^{\mathrm{T}}]\right)^{-1}, \tag{B.2}$$

and $\mathbf{b}(y, \mathbf{x})$ is such that

$$E\{\mathbf{b}(y, \mathbf{X}) \mid y\} = E[E\{\mathbf{b}(Y, \mathbf{X}) \mid \mathbf{X}\} \mid y]. \tag{B.3}$$

We can verify that $\partial\boldsymbol{\xi}/\partial\boldsymbol{\theta}^{\mathrm{T}} = E(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\theta}}^{\mathrm{T}})$ where $\mathbf{S}_{\boldsymbol{\theta}} = (\mathbf{S}_{\boldsymbol{\alpha}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$. Now, the tangent space is $\mathcal{T} = \mathcal{T}_{\boldsymbol{\alpha}} \oplus (\mathcal{T}_{\boldsymbol{\beta}} + \mathcal{T}_{\boldsymbol{\gamma}})$ where

$$\begin{aligned}
\mathcal{T}_{\boldsymbol{\alpha}} &= [\mathbf{a}(\mathbf{x}) : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}], \\
\mathcal{T}_{\boldsymbol{\beta}} &= [\mathbf{M}\mathbf{x}\{y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} : \forall \mathbf{M}], \\
\mathcal{T}_{\boldsymbol{\gamma}} &= [\mathbf{a}(y) - E\{\mathbf{a}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\} : \forall \mathbf{a}(y)].
\end{aligned}$$

Let

$$\begin{aligned}
\mathbf{M} = {} & E\left(\mathbf{B}E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{b}(Y, \mathbf{X})\mathbf{X}^{\mathrm{T}}\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} - 2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}}Yv(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right. \\
& \left. -\mathbf{a}(Y)\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right)\left[E\{\mathbf{X}\mathbf{X}^{\mathrm{T}}v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\right]^{-1},
\end{aligned} \tag{B.4}$$

where $\mathbf{a}(y)$ is such that

$$\begin{aligned}
& E[E\{\mathbf{a}(Y) \mid \mathbf{X}\} \mid y] - \mathbf{a}(y) \\
= {} & 2\boldsymbol{\beta}E[yE(Y \mid \mathbf{X}) - E\{YE(Y \mid \mathbf{X}) \mid \mathbf{X}\} \mid y] + \mathbf{M}E[\mathbf{X}\{y - E(Y \mid \mathbf{X})\} \mid y].
\end{aligned} \tag{B.5}$$

Then the efficient influence function is

$$\begin{aligned}
\boldsymbol{\phi}_{\mathrm{eff}}(y, \mathbf{x}) = {} & \boldsymbol{\beta}v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) - \boldsymbol{\beta}E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} + \boldsymbol{\beta}\{y^2 - E(Y^2 \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} \\
& + \mathbf{M}\mathbf{x}\{y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} + \mathbf{a}(y) - E\{\mathbf{a}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\},
\end{aligned}$$

where $\mathbf{a}(y)$ satisfies (B.5), $\mathbf{M}$ is given in (B.4), $\mathbf{b}(y, \mathbf{x})$ satisfies (B.3) and $\mathbf{B}$ is given in (B.2).

Note that once we have $\mathbf{a}(y)$, we can let $\mathbf{b}(y, \mathbf{x}) = 2\boldsymbol{\beta}yE(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) + \mathbf{M}\mathbf{x}y + \mathbf{a}(y)$ and it satisfies (B.3). Using this specific $\mathbf{b}(y, \mathbf{x})$ function, we obtain that $\mathbf{M}$ and $\mathbf{B}$ need to satisfy

$$\mathbf{B} = \left(E[\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\{2\boldsymbol{\beta}YE(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \mathbf{M}\mathbf{X}Y + \mathbf{a}(Y)\}\mathbf{X}^{\mathrm{T}}]\right)^{-1}$$

and

$$\mathbf{M}E\{\mathbf{X}\mathbf{X}^{\mathrm{T}}v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} = E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\left[2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}}Yv(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \mathbf{a}(Y)\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right].$$

This leads to

$$\begin{aligned}\mathbf{M} &= \left(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\left[2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}}Yv(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \mathbf{a}(Y)\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right) \\ &\quad \times \left[E\{\mathbf{X}\mathbf{X}^{\mathrm{T}}v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\right]^{-1}, \qquad \text{(B.6)}\end{aligned}$$

where $\mathbf{a}(y)$ satisfies (B.5).

Gathering the above derivations and results, we obtain the summary description of the efficient influence function as

$$\boldsymbol{\phi}_{\text{eff}}(y, \mathbf{x}) = \boldsymbol{\beta}v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) - \boldsymbol{\beta}E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} + \boldsymbol{\beta}y^2 + \mathbf{M}\mathbf{x}y + \mathbf{a}(y) - E\{\boldsymbol{\beta}Y^2 + \mathbf{M}\mathbf{x}Y + \mathbf{a}(Y) \mid \mathbf{x}\},$$

where $\mathbf{a}(y)$ satisfies (B.5) and $\mathbf{M}$ is given in (B.6). Obviously, the variance of the efficient influence function, i.e. $\text{var}\{\boldsymbol{\phi}_{\text{eff}}(Y, \mathbf{X})\}$, is the efficiency bound in estimating $\boldsymbol{\xi}$.

B.2 Derivation of the efficiency bound of marginal quantile effect estimation

Recall that for an arbitrary parametric submodel in (B.1), we have already derived the corresponding score functions $\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x})$, $\mathbf{S}_{\boldsymbol{\beta}}(y, \mathbf{x})$ and $\mathbf{S}_{\boldsymbol{\gamma}}(y, \mathbf{x})$. Now for notational brevity, let $\nu \equiv \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, $q(\nu) \equiv Q_\tau(Y \mid \nu)$, $\epsilon \equiv \tau - I\{Y < q(\nu)\}$, and $\epsilon'_\nu = -\delta\{q(\nu) - Y\}q'(\nu)$. Write $f(y, \nu) \equiv f_{Y|\mathbf{X}}(y, \nu)$. Using the quantile definition, we further have

$$\begin{aligned}E(\epsilon \mid \nu) &= 0, \\ E(\epsilon Y \mid \nu) &= f\{q(\nu), \nu\}q'(\nu), \\ \frac{E(\epsilon Y^2 \mid \nu)}{f\{q(\nu), \nu\}} &= 2q(\nu)q'(\nu) + q'(\nu)^2[\nu + c'\{q(\nu)\}] + q''(\nu),\end{aligned}$$

which can be verified based on

$$\begin{aligned}\tau &= E[I\{Y \leqslant q(\nu)\} \mid \nu], \\ q'(\nu) &= \frac{E\left([\tau - I\{Y < q(\nu)\}]Y \mid \nu\right)}{f\{q(\nu), \nu\}},\end{aligned}$$

$$q''(\nu) = \frac{E\left([\tau - I\{Y < q(\nu)\}]Y^2 \mid \nu\right)}{f\{q(\nu), \nu\}} - q(\nu)q'(\nu) - q'(\nu)[q(\nu) + q'(\nu)\nu + q'(\nu)c'\{q(\nu)\}].$$

We also have

$$\begin{aligned}
\frac{\partial q(\nu)}{\partial \boldsymbol{\gamma}} &= \frac{E\left\{\epsilon \mathbf{a}(Y) \mid \nu\right\}}{f\{q(\nu), \nu\}}, \\
\frac{\partial q'(v)}{\partial \boldsymbol{\gamma}} &= \frac{E\left\{\epsilon Y \mathbf{a}(Y) \mid \nu\right\}}{f\{q(\nu), \nu\}} - \mathbf{a}\{q(\nu)\}q'(\nu) - \frac{\partial q(\nu)}{\partial \boldsymbol{\gamma}}\left[q(\nu) + q'(\nu)\nu + q'(\nu)c'\{q(\nu)\}\right].
\end{aligned}$$

Note that $\boldsymbol{\eta}_\tau = \boldsymbol{\beta} E\{q'(\nu)\}$. We can verify that

$$\begin{aligned}
\frac{\partial \boldsymbol{\eta}_\tau}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} &= \boldsymbol{\beta} E\{q'(\nu)\mathbf{a}^{\mathrm{T}}(\mathbf{X})\}, \\
\frac{\partial \boldsymbol{\eta}_\tau}{\partial \boldsymbol{\beta}^{\mathrm{T}}} &= E\{q'(\nu)\}\mathbf{I} + \boldsymbol{\beta} E[r(Y, \nu)\mathbf{X}^{\mathrm{T}}\{Y - E(Y \mid \nu)\}], \\
\frac{\partial \boldsymbol{\eta}_\tau}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} &= \boldsymbol{\beta} E\left(r(Y, \nu)[\mathbf{a}^{\mathrm{T}}(Y) - E\{\mathbf{a}^{\mathrm{T}}(Y) \mid \mathbf{X}\}]\right),
\end{aligned}$$

where

$$r(Y, \nu) \equiv \frac{\epsilon Y + \epsilon'_\nu - \epsilon[q(\nu) + q'(\nu)\nu + q'(\nu)c'\{q(\nu)\}]}{f\{q(\nu), \nu\}}. \tag{B.7}$$

Hence, a possible influence function is

$$\begin{aligned}
\boldsymbol{\phi}(y, \mathbf{x}) &= \boldsymbol{\beta} q'(\nu) - \boldsymbol{\beta} E\{q'(\nu)\} + \boldsymbol{\beta}[r(Y, \nu) - E\{r(Y, \nu) \mid \nu\}] \\
&\quad + E\{q'(\nu)\}\mathbf{B}[\mathbf{b}(y, \mathbf{x}) - E\{\mathbf{b}(Y, \mathbf{x}) \mid \mathbf{x}\}],
\end{aligned}$$

where

$$\mathbf{B} \equiv \left(E[\mathbf{b}(Y, \mathbf{X})\mathbf{X}^{\mathrm{T}}\{Y - E(Y \mid \nu)\}]\right)^{-1}, \tag{B.8}$$

and $\mathbf{b}(y, \mathbf{x})$ is such that

$$E\{\mathbf{b}(y, \mathbf{X}) \mid y\} = E[E\{\mathbf{b}(Y, \mathbf{X}) \mid \mathbf{X}\} \mid y]. \tag{B.9}$$

We can verify that $\partial \boldsymbol{\eta}_\tau / \partial \boldsymbol{\theta}^{\mathrm{T}} = E(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\theta}}^{\mathrm{T}})$. Note that we have derived the tangent space $\mathcal{T}$ in Section 3.2.1. Let $\boldsymbol{\rho}(y, \mathbf{x}) \equiv E\{q'(\nu)\}\mathbf{B}\mathbf{b}(y, \mathbf{x}) + \boldsymbol{\beta} r(y, \nu) - \mathbf{a}(y)$, $v(\nu) \equiv E(Y^2 \mid$

$\nu) - \{E(Y \mid \nu)\}^2$, and

$$\mathbf{M}_1 \equiv E\left[\boldsymbol{\rho}(Y, \mathbf{X})\mathbf{X}^{\mathrm{T}}\{Y - E(Y \mid \nu)\}\right]\left[E\{\mathbf{X}\mathbf{X}^{\mathrm{T}}v(\nu)\}\right]^{-1}, \tag{B.10}$$

where $\mathbf{a}(y)$ is such that

$$E[E\{\mathbf{a}(Y) \mid \mathbf{X}\} \mid y] - \mathbf{a}(y) = -\boldsymbol{\beta}E\{r(y, \nu)|y\} + \mathbf{M}_1 E[\mathbf{X}\{y - E(Y|\mathbf{X})\}|y]. \tag{B.11}$$

Then the efficient influence function is

$$\boldsymbol{\phi}_{\text{eff}}(y, \mathbf{x}) = \boldsymbol{\beta}q'(\nu) - \boldsymbol{\beta}E\{q'(\nu)\} + \mathbf{M}_1\mathbf{x}y + \mathbf{a}(y) - E\{\mathbf{M}_1\mathbf{x}Y + \mathbf{a}(Y)|\mathbf{x}\}, \tag{B.12}$$

where $\mathbf{a}(y)$ satisfies (B.11), $\mathbf{M}_1$ is given in (B.10), $\mathbf{b}(y, \mathbf{x})$ satisfies (B.9) and $\mathbf{B}$ is given in (B.8). Note that once we have $\mathbf{a}(y)$, we can let $\mathbf{b}(y, \mathbf{x}) = \boldsymbol{\beta}r(y, \nu) - \mathbf{M}_1\mathbf{x}y - \mathbf{a}(y)$ and it satisfies (B.9). Using this specific $\mathbf{b}(y, \mathbf{x})$ function, we obtain that $\mathbf{M}_1$ and $\mathbf{B}$ need to satisfy

$$\mathbf{B} = \left(E[\{\boldsymbol{\beta}r(Y, \nu) - \mathbf{M}_1\mathbf{X}Y - \mathbf{a}(Y)\}\mathbf{X}^{\mathrm{T}}\{Y - E(Y \mid \nu)\}]\right)^{-1}$$

and

$$\mathbf{M}_1 E\{\mathbf{X}\mathbf{X}^{\mathrm{T}}v(\nu)\} = E\{q'(\nu)\}\mathbf{I} + \boldsymbol{\beta}E\{\mathbf{X}^{\mathrm{T}}q''(\nu)\} - E[\mathbf{a}(Y)\mathbf{X}^{\mathrm{T}}\{Y - E(Y|\mathbf{X})\}].$$

This leads to

$$\mathbf{M}_1 = (E\{q'(\nu)\}\mathbf{I} + \boldsymbol{\beta}E\{\mathbf{X}^{\mathrm{T}}q''(\nu)\} - E[\mathbf{a}(Y)\mathbf{X}^{\mathrm{T}}\{Y - E(Y|\mathbf{X})\}])[E\{\mathbf{X}\mathbf{X}^{\mathrm{T}}v(\nu)\}]^{-1}, \tag{B.13}$$

where $\mathbf{a}(y)$ satisfies (B.11).

Thus, in summary, the efficient influence function for estimating $\boldsymbol{\eta}_\tau$ is given in (B.12), where $\mathbf{a}(y)$ satisfies (B.11), $r(Y, \nu)$ is given in (B.7), and $\mathbf{M}_1$ is given in (B.13).

B.3 Preliminaries and lemmas

To ease theoretical analysis, we consider MLE under an identification constraint different from our original problem (3.5). We define $(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}} = \arg\max_{(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\zeta}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Theta}} l(\boldsymbol{\beta}, \boldsymbol{\zeta})$ where

$\boldsymbol{\Theta} = \{(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\zeta}^{\mathrm{T}})^{\mathrm{T}} : \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\zeta} \in \boldsymbol{\zeta}\}, \boldsymbol{\zeta} = \{\boldsymbol{\zeta} \in \mathbb{R}^{m+1} : \mathbf{1}^{\mathrm{T}}\boldsymbol{\zeta} = 0, \|\boldsymbol{\zeta}\|_\infty \leqslant C_{\boldsymbol{\zeta}}\},$

$$l(\boldsymbol{\beta}, \boldsymbol{\zeta}) \equiv \sum_{i=1}^{n} \left[y_i \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i + \mathbf{B}_+(y_i)^{\mathrm{T}} \boldsymbol{\zeta} - \log \int \exp\left\{ y \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i + \mathbf{B}_+(y)^{\mathrm{T}} \boldsymbol{\zeta} \right\} d\mu(y) \right],$$

and $\mathbf{B}_+(\cdot) \equiv \{B_0(\cdot), \mathbf{B}(\cdot)^{\mathrm{T}}\}^{\mathrm{T}} = \mathbf{e}_1 + \mathbf{T}\mathbf{B}(\cdot)$ where $\mathbf{T} \equiv (-\mathbf{1}_m, \mathbf{I}_m)^{\mathrm{T}}$ because $\mathbf{1}^{\mathrm{T}}\mathbf{B}_+(\cdot) = 1$. We also define $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}) \equiv \arg\max_{\boldsymbol{\zeta} \in \boldsymbol{\zeta}} l(\boldsymbol{\beta}, \boldsymbol{\zeta})$ for any given $\boldsymbol{\beta}$. Note that $\mathbf{B}_+(\cdot)^{\mathrm{T}}\boldsymbol{\zeta} = \zeta_0 + \mathbf{B}(\cdot)^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\boldsymbol{\zeta})$. Then since $f_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}, c)$ defined in (3.4) is identical for $c = c_1$ and $c = c_2$ if $c_1(\cdot) = a + c_2(\cdot)$ for some constant a, we have $\boldsymbol{\gamma} = \mathbf{T}^{\mathrm{T}}\boldsymbol{\zeta}$.

Now, we introduce some notations. We let $\mathbf{S}^*(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) \equiv (\mathbf{S}^{*\mathrm{T}}_{\boldsymbol{\beta}}, \mathbf{S}^{*\mathrm{T}}_{\boldsymbol{\zeta}})^{\mathrm{T}}(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta})$ and $\mathbf{I}^*(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) \equiv \{(\mathbf{I}^*_{\boldsymbol{\beta}}, \mathbf{I}^*_{\boldsymbol{\beta}\boldsymbol{\zeta}})^{\mathrm{T}}, (\mathbf{I}^*_{\boldsymbol{\zeta}\boldsymbol{\beta}}, \mathbf{I}^*_{\boldsymbol{\zeta}})^{\mathrm{T}}\}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta})$, where

$$\begin{aligned}
\mathbf{S}^*_{\boldsymbol{\beta}}(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) &\equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \mathbf{x}\{y - E^*(Y \mid \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta})\}, \\
\mathbf{S}^*_{\boldsymbol{\zeta}}(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) &\equiv \frac{\partial}{\partial \boldsymbol{\zeta}} \log f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \mathbf{B}_+(y) - E^*\{\mathbf{B}_+(Y) \mid \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}\}, \\
\mathbf{I}^*_{\boldsymbol{\beta}}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) &\equiv -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \log f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \mathbf{x}\mathbf{x}^{\mathrm{T}} \mathrm{var}^*(Y \mid \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}), \\
\mathbf{I}^*_{\boldsymbol{\beta}\boldsymbol{\zeta}}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) &\equiv -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\zeta}^{\mathrm{T}}} \log f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \mathbf{x}\mathrm{cov}^*\{Y, \mathbf{B}_+(Y) \mid \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}\} \equiv \mathbf{I}^{*\mathrm{T}}_{\boldsymbol{\zeta}\boldsymbol{\beta}}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}), \\
\mathbf{I}^*_{\boldsymbol{\zeta}}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) &\equiv -\frac{\partial^2}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^{\mathrm{T}}} \log f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \mathrm{var}^*\{\mathbf{B}_+(Y) \mid \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}\}.
\end{aligned}$$

Similarly, we define $\mathbf{S}(y, \mathbf{x}) \equiv (\mathbf{S}^{\mathrm{T}}_{\boldsymbol{\beta}}, \mathbf{S}^{\mathrm{T}}_{\boldsymbol{\zeta}})^{\mathrm{T}}(y, \mathbf{x})$ and $\mathbf{I}(\mathbf{x}) \equiv \{(\mathbf{I}_{\boldsymbol{\beta}}, \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\zeta}})^{\mathrm{T}}, (\mathbf{I}_{\boldsymbol{\zeta}\boldsymbol{\beta}}, \mathbf{I}_{\boldsymbol{\zeta}})^{\mathrm{T}}\}(\mathbf{x})$, where $\mathbf{S}_{\boldsymbol{\beta}}(y, \mathbf{x}) \equiv \mathbf{x}\{y - E(Y \mid \mathbf{x})\}, \mathbf{S}_{\boldsymbol{\zeta}}(y, \mathbf{x}) \equiv \mathbf{B}_+(y) - E\{\mathbf{B}_+(Y) \mid \mathbf{x}\}, \mathbf{I}_{\boldsymbol{\beta}}(\mathbf{x}) \equiv \mathbf{x}\mathbf{x}^{\mathrm{T}}\mathrm{var}(Y \mid \mathbf{x}), \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\zeta}}(\mathbf{x}) \equiv \mathbf{x}\mathrm{cov}\{Y, \mathbf{B}_+(Y) \mid \mathbf{x}\} \equiv \mathbf{I}^{\mathrm{T}}_{\boldsymbol{\zeta}\boldsymbol{\beta}}(\mathbf{x}), \mathbf{I}_{\boldsymbol{\zeta}}(\mathbf{x}) \equiv \mathrm{var}\{\mathbf{B}_+(Y) \mid \mathbf{x}\}$. Note that $\mathbf{S}_{\boldsymbol{\zeta}}(y, \mathbf{x}) = \mathbf{T}[\mathbf{B}(y) - E\{\mathbf{B}(Y) \mid \mathbf{x}\}]$, $\mathbf{I}_{\boldsymbol{\zeta}\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{T}\mathrm{cov}\{\mathbf{B}(Y), Y \mid \mathbf{x}\}\mathbf{x}^{\mathrm{T}}$, and $\mathbf{I}_{\boldsymbol{\zeta}}(\mathbf{x}) = \mathbf{T}\mathrm{var}\{\mathbf{B}(Y) \mid \mathbf{x}\}\mathbf{T}^{\mathrm{T}}$.

Lemma B.3.1. *(i)* $\mathbf{S}^*(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}), \mathbf{S}(y, \mathbf{x}) \in \boldsymbol{\Theta}$,

(ii) $\mathbf{I}^*(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta})\boldsymbol{\theta}, \mathbf{I}(\mathbf{x})\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *for any* $\boldsymbol{\theta} \in \mathbb{R}^{p+m+1}$.

Proof. Since $\mathbf{1}^{\mathrm{T}}\mathbf{B}_+(y) = 1$, we immediately have $\mathbf{1}^{\mathrm{T}}\mathbf{S}^*_{\boldsymbol{\zeta}}(y, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = 0$, $\mathbf{1}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\zeta}}(y, \mathbf{x}) = 0$, $\mathbf{1}^{\mathrm{T}}\mathbf{I}^*_{\boldsymbol{\zeta}\boldsymbol{\beta}}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \mathbf{0}^{\mathrm{T}}$, $\mathbf{1}^{\mathrm{T}}\mathbf{I}^*_{\boldsymbol{\zeta}}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \mathbf{0}^{\mathrm{T}}$, $\mathbf{1}^{\mathrm{T}}\mathbf{I}_{\boldsymbol{\zeta}\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{0}^{\mathrm{T}}$, $\mathbf{1}^{\mathrm{T}}\mathbf{I}_{\boldsymbol{\zeta}}(\mathbf{x}) = \mathbf{0}^{\mathrm{T}}$, which imply the results. □

In addition, we let $\boldsymbol{\Omega} \equiv E\{\mathbf{I}(\mathbf{X})\}$, $\boldsymbol{\Omega}_{11} \equiv E\{\mathbf{I}_{\boldsymbol{\beta}}(\mathbf{X})\}$, $\boldsymbol{\Omega}_{12} \equiv E\{\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\zeta}}(\mathbf{X})\}$, $\boldsymbol{\Omega}_{21} \equiv \boldsymbol{\Omega}^{\mathrm{T}}_{12}$, $\boldsymbol{\Omega}_{22} \equiv E\{\mathbf{I}_{\boldsymbol{\zeta}}(\mathbf{X})\}$, $\widehat{\boldsymbol{\Omega}}^*(\boldsymbol{\beta}, \boldsymbol{\zeta}) \equiv n^{-1}\sum_{i=1}^n \mathbf{I}^*(\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\zeta})$, $\widehat{\boldsymbol{\Omega}}^*_{11}(\boldsymbol{\beta}, \boldsymbol{\zeta}) \equiv n^{-1}\sum_{i=1}^n \mathbf{I}^*_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\zeta})$, $\widehat{\boldsymbol{\Omega}}^*_{12}(\boldsymbol{\beta}, \boldsymbol{\zeta}) \equiv n^{-1}\sum_{i=1}^n \mathbf{I}^*_{\boldsymbol{\beta}\boldsymbol{\zeta}}(\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\zeta})$, $\widehat{\boldsymbol{\Omega}}^*_{21}(\boldsymbol{\beta}, \boldsymbol{\zeta}) \equiv \widehat{\boldsymbol{\Omega}}^*_{12}(\boldsymbol{\beta}, \boldsymbol{\zeta})^{\mathrm{T}}$, $\widehat{\boldsymbol{\Omega}}^*_{22}(\boldsymbol{\beta}, \boldsymbol{\zeta}) \equiv n^{-1}\sum_{i=1}^n \mathbf{I}^*_{\boldsymbol{\zeta}}(\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\zeta})$.

Note that $\mathbf{\Omega}_{11} = \mathbf{\Sigma}_{11}$, $\mathbf{\Omega}_{21} = \mathbf{T}\mathbf{\Sigma}_{21}$, and $\mathbf{\Omega}_{22} = \mathbf{T}\mathbf{\Sigma}_{22}\mathbf{T}^{\mathrm{T}}$. Also, Condition (C6) is equivalent to Condition (C6').

(C6') $\mathbf{\Omega}$ is positive semi-definite and $\mathbf{\Omega}$ has only one eigenvector $(\mathbf{0}_p^{\mathrm{T}}, \mathbf{1}_{m+1}^{\mathrm{T}})^{\mathrm{T}}$ which corresponds to eigenvalue 0, because of the property of the B-spline bases $\mathbf{1}^{\mathrm{T}}\mathbf{B}_+(y) = 1$.

Also, let $\mathbf{\Omega}_{22}^-$ be the Moore-Penrose inverse of $\mathbf{\Omega}_{22}$ and

$$\mathbf{\Omega}^- \equiv \{(\mathbf{\Omega}_{\boldsymbol{\beta}}, -\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^-)^{\mathrm{T}}, (-\mathbf{\Omega}_{22}^-\mathbf{\Omega}_{21}\mathbf{\Omega}_{\boldsymbol{\beta}}, \mathbf{\Omega}_{22}^- + \mathbf{\Omega}_{22}^-\mathbf{\Omega}_{21}\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^-)^{\mathrm{T}}\}$$

where $\mathbf{\Omega}_{\boldsymbol{\beta}} \equiv (\mathbf{\Omega}_{11} - \mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^-\mathbf{\Omega}_{21})^{-1}$. Similarly we define $\widehat{\mathbf{\Omega}}_{22}^{*-}(\boldsymbol{\beta}, \boldsymbol{\zeta})$, $\widehat{\mathbf{\Omega}}^{*-}(\boldsymbol{\beta}, \boldsymbol{\zeta})$ and $\widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}}^{*}(\boldsymbol{\beta}, \boldsymbol{\zeta})$.

Lemma B.3.2. *We have the following results.*

(i) Under Condition (C6'), $\mathbf{\Omega}_{\boldsymbol{\beta}}$ and $\widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}}^{*}(\boldsymbol{\beta}, \boldsymbol{\zeta})$ are positive definite.

(ii) When $\mathbf{A} = \mathbf{\Omega}_{22}$, $\widehat{\mathbf{\Omega}}_{22}^{*}(\boldsymbol{\beta}, \boldsymbol{\zeta})$, $\mathbf{\Omega}$, or $\widehat{\mathbf{\Omega}}^{*}(\boldsymbol{\beta}, \boldsymbol{\zeta})$, we have $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$, $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$, and $\mathbf{A}^-\mathbf{A} = \mathbf{A}\mathbf{A}^-$.

Proof. Note that $\mathbf{\Omega}_{22}$ has rank m. Let $\mathbf{\Omega}_{22} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}$ be the singular value decomposition (svd) where $\mathbf{\Lambda}$ is nonsingular. Then $\mathbf{\Omega}_{22}^- = \mathbf{P}\mathbf{\Lambda}^{-1}\mathbf{P}^{\mathrm{T}}$ and $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}_m$. We can verify that $\mathbf{\Omega}_{22}^-\mathbf{\Omega}_{22} = \mathbf{\Omega}_{22}\mathbf{\Omega}_{22}^- = \mathbf{P}\mathbf{P}^{\mathrm{T}}$ and $\mathbf{\Omega}_{22}^-$ satisfies the properties stated in the lemma.

In addition, note that $\mathbf{\Omega}_{21} = \mathbf{\Omega}_{22}\mathbf{\Omega}_{22}^-\mathbf{\Omega}_{21}$, since $\mathbf{\Omega}_{21}\mathbf{e}_j \in \boldsymbol{\zeta}$, $\mathbf{\Omega}_{22}$ has only one eigenvector $\mathbf{1}_{m+1}$ associated with eigenvalue 0, and $\mathbf{1}_{m+1}$ is orthogonal to $\boldsymbol{\zeta}$. Then simple calculation will show $\mathbf{\Omega}^-\mathbf{\Omega} = \mathbf{\Omega}\mathbf{\Omega}^- = \{(\mathbf{I}, \mathbf{0})^{\mathrm{T}}, (\mathbf{0}, \mathbf{\Omega}_{22}^-\mathbf{\Omega}_{22})^{\mathrm{T}}\}$ and this leads to the properties of $\mathbf{\Omega}^-$ stated in the lemma. In addition, we can decompose $\mathbf{\Omega} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ where $\mathbf{U} \equiv \{(\mathbf{I}, \mathbf{0})^{\mathrm{T}}, (\mathbf{\Omega}_{22}^-\mathbf{\Omega}_{21}, \mathbf{I})^{\mathrm{T}}\}$ and $\mathbf{D} \equiv \mathrm{diag}(\mathbf{\Omega}_{11} - \mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^-\mathbf{\Omega}_{21}, \mathbf{\Omega}_{22})$. Since $\mathbf{U}$ is invertible, $\mathrm{rank}(\mathbf{\Omega}) = \mathrm{rank}(\mathbf{\Omega}_{11} - \mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^-\mathbf{\Omega}_{21}) + \mathrm{rank}(\mathbf{\Omega}_{22})$. Noting that $\mathrm{rank}(\mathbf{\Omega}_{22}) = m$ and $\mathrm{rank}(\mathbf{\Omega}) = p + m$ by Condition (C6'), we get $\mathbf{\Omega}_{11} - \mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^-\mathbf{\Omega}_{21}$ to have full rank. Further, $\mathbf{\Omega}_{11} - \mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^-\mathbf{\Omega}_{21}$ is positive definite since it is the Schur complement of $\mathbf{\Omega}_{22}^-$ in the positive semi-definite matrix $\mathbf{\Omega}$. Therefore, $\mathbf{\Omega}_{\boldsymbol{\beta}}$ is positive definite.

We omit the proof for the properties of $\widehat{\mathbf{\Omega}}_{22}^{*-}(\boldsymbol{\beta}, \boldsymbol{\zeta})$, $\widehat{\mathbf{\Omega}}^{*-}(\boldsymbol{\beta}, \boldsymbol{\zeta})$, and $\widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}}^{*}(\boldsymbol{\beta}, \boldsymbol{\zeta})$ since they can be shown similarly. □

Lemma B.3.3. *We have (i) $\mathbf{T}^{\mathrm{T}}\mathbf{\Omega}_{22}^-\mathbf{T} = \mathbf{\Sigma}_{22}^{-1}$, (ii) $\mathbf{\Omega}_{\boldsymbol{\beta}} = \mathbf{\Sigma}_{\boldsymbol{\beta}}$, and (iii)*

$$\mathit{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\mathbf{\Omega}^- \mathit{diag}(\mathbf{I}_p, \mathbf{T}) = \mathbf{\Sigma}^{-1}.$$

Proof. Let $\boldsymbol{\Omega}_{22} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}}$ be the singular value decomposition where $\boldsymbol{\Lambda}$ is nonsingular and $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}_m$. Since $\boldsymbol{\Omega}_{22}$ has a single zero eigenvalue with the corresponding eigenvector $\mathbf{1}_{m+1}$, the columns of $\mathbf{P}$ are orthogonal to $\mathbf{1}_{m+1}$, i.e. $\mathbf{1}_{m+1}^{\mathrm{T}}\mathbf{P} = \mathbf{0}_m$. First, we have $\mathbf{T} = (-\mathbf{1}_m, \mathbf{I}_m)^{\mathrm{T}} = \mathbf{P}\mathbf{Q}$ for some $\mathbf{Q}$ of full rank since $\mathbf{T}$ has rank m and the columns of $\mathbf{T}$ is orthogonal to the vector $\mathbf{1}_{m+1}$ as well, i.e., $\mathbf{1}_{m+1}^{\mathrm{T}}\mathbf{T} = \mathbf{0}_m$. Then $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}} = \boldsymbol{\Omega}_{22} = \mathbf{T}\boldsymbol{\Sigma}_{22}\mathbf{T}^{\mathrm{T}} = \mathbf{P}\mathbf{Q}\boldsymbol{\Sigma}_{22}(\mathbf{P}\mathbf{Q})^{\mathrm{T}}$. Pre-multiplying $\mathbf{P}^{\mathrm{T}}$ and post-multiplying $\mathbf{P}$ on both sides give $\boldsymbol{\Lambda} = \mathbf{Q}\boldsymbol{\Sigma}_{22}\mathbf{Q}^{\mathrm{T}}$. This leads to $\mathbf{T}^{\mathrm{T}}\boldsymbol{\Omega}_{22}^{-}\mathbf{T} = (\mathbf{P}\mathbf{Q})^{\mathrm{T}}\{\mathbf{P}(\mathbf{Q}\boldsymbol{\Sigma}_{22}\mathbf{Q}^{\mathrm{T}})^{-1}\mathbf{P}^{\mathrm{T}}\}(\mathbf{P}\mathbf{Q}) = \boldsymbol{\Sigma}_{22}^{-1}$. In addition, $\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\Omega}_{21} = \boldsymbol{\Sigma}_{12}\mathbf{T}^{\mathrm{T}}\boldsymbol{\Omega}_{22}^{-}\mathbf{T}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, which leads to $\boldsymbol{\Omega}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. Finally, using the above results,

$$\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\boldsymbol{\Omega}^{-}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}) = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} & -\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{22}^{-1} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix} = \boldsymbol{\Sigma}^{-1}.$$

□

Lemma B.3.4. *Under Conditions (C3)-(C4),* $\|\mathbf{B}_+(\cdot)^{\mathrm{T}}\boldsymbol{\zeta}\|_p \asymp N^{-1/p}\|\boldsymbol{\zeta}\|_p$ *for* $1 \leqslant p \leqslant \infty$.

Proof. We recall Lemma 1 in the supplement of Jiang et al. (2018), which is a direct result from Theorem 5.4.2 on page 145 of DeVore & Lorentz (1993). For each spline $\sum_{k=0}^{m}\zeta_k B_k(y)$ and $1 \leqslant p \leqslant \infty$, there exists a constant $C_r > 0$ such that $C_r\|\boldsymbol{\zeta}'\|_p \leqslant \left\|\mathbf{B}_+(\cdot)^{\mathrm{T}}\boldsymbol{\zeta}\right\|_p \leqslant \|\boldsymbol{\zeta}'\|_p$, where $\boldsymbol{\zeta}' \equiv [\zeta_k\{(t_k - t_{k-r})/r\}^{1/p}, k = 0, \ldots, m]^{\mathrm{T}}$. This implies $\|\mathbf{B}_+(\cdot)^{\mathrm{T}}\boldsymbol{\zeta}\|_p \asymp \|\boldsymbol{\zeta}'\|_p$ where

$$\|\boldsymbol{\zeta}'\|_p = \left\{\sum_{k=0}^{m}\left|u_k\left(\frac{t_k - t_{k-r}}{r}\right)^{1/p}\right|^p\right\}^{1/p} \asymp N^{-1/p}\left(\sum_{k=0}^{m}|u_k|^p\right)^{1/p} = N^{-1/p}\|\boldsymbol{\zeta}\|_p$$

by Conditions (C3) and (C4). □

Lemma B.3.5. *Under Conditions (C1), (C3), and (C4), we have the following results.*

(i) $\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{I}_{\boldsymbol{\zeta}}(\mathbf{x})\boldsymbol{\zeta} \asymp N^{-1}\|\boldsymbol{\zeta}\|_2^2$ *and* $\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\zeta} \asymp N\|\boldsymbol{\zeta}\|_2^2$ *for* $\boldsymbol{\zeta} \in \boldsymbol{\zeta}$. *Similarly,* $\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{I}_{\boldsymbol{\zeta}}^{*}(\mathbf{x}, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\boldsymbol{\zeta} \asymp N^{-1}\|\boldsymbol{\zeta}\|_2^2$ *and* $\boldsymbol{\zeta}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\boldsymbol{\zeta} \asymp N\|\boldsymbol{\zeta}\|_2^2$ *for* $\boldsymbol{\zeta} \in \boldsymbol{\zeta}$ *and fixed* $\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*$.

(ii) Let the singular value decomposition of $\boldsymbol{\Omega}_{22}$ *be* $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}}$, *then* $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}_m$, $\mathbf{P}^{\mathrm{T}}\mathbf{1}_{m+1} = \mathbf{0}_m$, *the columns of* $\mathbf{P}$ *span* $\boldsymbol{\zeta}$, *and all diagonal elements of* $\boldsymbol{\Lambda}$ *are of order* N^{-1}. *Similarly, let* $\mathbf{P}_n\boldsymbol{\Lambda}_n\mathbf{P}_n^{\mathrm{T}}$ *be the singular value decomposition of* $\widehat{\boldsymbol{\Omega}}_{22}^{*}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)$, *then* $\mathbf{P}_n^{\mathrm{T}}\mathbf{P}_n = \mathbf{I}_m$, $\mathbf{P}_n^{\mathrm{T}}\mathbf{1}_{m+1} = \mathbf{0}_m$, *the columns of* $\mathbf{P}_n$ *span* $\boldsymbol{\zeta}$, *and all diagonal elements of* $\boldsymbol{\Lambda}_n$ *are of order* N^{-1}.

Proof. Let $\boldsymbol{\zeta} \in \boldsymbol{\zeta}$. Note that $\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{I}_{\boldsymbol{\zeta}}(\mathbf{x})\boldsymbol{\zeta} = \mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}\}$. We first have

$$\mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}\} \leqslant E[\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta}\}^2 \mid \mathbf{x}] \leqslant \|\mathbf{B}_+(\cdot)^{\mathrm{T}}\boldsymbol{\zeta}\|_2^2 \|f_{Y|\mathbf{X}}(\cdot, \mathbf{x})\|_\infty \asymp N^{-1}\|\boldsymbol{\zeta}\|_2^2$$

by Condition (C1) and Lemma B.3.4. In addition, since $\mathbf{1}^{\mathrm{T}}\mathbf{B}_+(y) = 1$,

$$\begin{aligned}
\mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}\} &= \int [\mathbf{B}_+(y)^{\mathrm{T}}\boldsymbol{\zeta} - E\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}\}]^2 f_{Y|\mathbf{X}}(y, \mathbf{x}) dy \\
&\geqslant \|\mathbf{B}_+(\cdot)^{\mathrm{T}}\boldsymbol{\zeta} - \mathbf{B}_+(\cdot)^{\mathrm{T}}\mathbf{1}E\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}\}\|_2^2 \inf_{y \in [0,1]} f_{Y|\mathbf{X}}(y, \mathbf{x}) \\
&\asymp N^{-1}\|\boldsymbol{\zeta} - \mathbf{1}E\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}\}\|_2^2 \\
&\geqslant N^{-1}\|\boldsymbol{\zeta}\|_2^2.
\end{aligned}$$

The third argument is by Condition (C1) and Lemma B.3.4, and the last inequality holds since $\mathbf{1}^{\mathrm{T}}\boldsymbol{\zeta} = 0$. Therefore, we get $\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{I}_{\boldsymbol{\zeta}}(\mathbf{x})\boldsymbol{\zeta} \asymp N^{-1}\|\boldsymbol{\zeta}\|_2^2$ for all $\boldsymbol{\zeta} \in \boldsymbol{\zeta}$. Further, we have $\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\Omega}_{22}\boldsymbol{\zeta} = \boldsymbol{\zeta}^{\mathrm{T}}E\{\mathbf{I}_{\boldsymbol{\zeta}}(\mathbf{X})\}\boldsymbol{\zeta} \asymp N^{-1}\|\boldsymbol{\zeta}\|_2^2$ by Condition (C1), and $\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\zeta} \asymp N\|\boldsymbol{\zeta}\|_2^2$ by the definition of $\boldsymbol{\Omega}_{22}^{-}$. Now, let the singular value decomposition of $\boldsymbol{\Omega}_{22}$ be $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}}$, then $\boldsymbol{\Omega}_{22}\mathbf{1}_{m+1} = \mathbf{0}_m$ and $\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\Omega}_{22}\boldsymbol{\zeta} \asymp N^{-1}\|\boldsymbol{\zeta}\|_2^2$ for $\boldsymbol{\zeta} \in \boldsymbol{\zeta}$ imply the second property.

We omit the proof of the results for $\mathbf{I}_{\boldsymbol{\zeta}}^*(\mathbf{x}, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)$ and $\widehat{\boldsymbol{\Omega}}_{22}^*(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)$ since it can be shown similarly. □

B.4 Proof of Proposition 3.3.1

We first show that $\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$. Since $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)$ maximizes $l(\boldsymbol{\beta}_0, \boldsymbol{\zeta})$, it satisfies $\sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}^*\{y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)\} = \mathbf{0}$. Then by the Taylor expansion, with $\boldsymbol{\zeta}^* \equiv \alpha\boldsymbol{\zeta}_0 + (1-\alpha)\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)$ for some $\alpha \in (0, 1)$,

$$\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0 = \widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) n^{-1} \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}^*(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0). \tag{B.14}$$

Note that $\|\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) n^{-1} \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}^*(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\|_2 \asymp n^{-1}N\|\sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}^*(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\|_2$ by Lemmas B.3.1 and B.3.5. In addition, $\|\sum_{i=1}^n \{\mathbf{S}_{\boldsymbol{\zeta}}^*(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\}\|_2 = o(n^{1/2})$ since

$$\begin{aligned}
\left\|\mathbf{S}_{\boldsymbol{\zeta}}^*(y, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \mathbf{S}_{\boldsymbol{\zeta}}(y, \mathbf{x})\right\|_2 &= \|E^*\{\mathbf{B}_+(Y) \mid \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0\} - E\{\mathbf{B}_+(Y) \mid \mathbf{x}\}\|_2 \\
&\leqslant \left\| \int \mathbf{B}_+(y) |f_{Y|\mathbf{X}}^*(y, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - f_{Y|\mathbf{X}}(y, \mathbf{x})| dy \right\|_2
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \int \|\mathbf{B}_+(y)\|_2 \, |f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - f_{Y|\mathbf{X}}(y, \mathbf{x})| dy \\
&\leqslant \int \|\mathbf{B}_+(y)\|_1 \, |f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - f_{Y|\mathbf{X}}(y, \mathbf{x})| dy \\
&= O(N^{-q}) = o(n^{-1/2} N^{-1/2}) \qquad \text{(B.15)}
\end{aligned}
$$

uniformly in $(y, \mathbf{x})$ by Conditions (C1), (C3), and (C5). Here we used Jensen's inequality in the second inequality since $\phi(\cdot) = \|\cdot\|_2$ is convex, and we used $\|\mathbf{B}_+(y)\|_2 \leqslant \|\mathbf{B}_+(y)\|_1 = 1$ for any y in the third inequality. Further, $\| \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\|_2 = O_p(n^{1/2})$ since

$$
\begin{aligned}
E\left\{ \left\| \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}(Y_i, \mathbf{X}_i) \right\|_2^2 \right\} &= nE\{\|\mathbf{S}_{\boldsymbol{\zeta}}(Y, \mathbf{X})\|_2^2\} \\
&= nE\left[\sum_{k=0}^m \mathrm{var}\{B_k(Y) \mid \mathbf{X}\} \right] \\
&\leqslant nE\left\{ \sum_{k=0}^m B_k(Y)^2 \right\} \\
&\leqslant n. \qquad \text{(B.16)}
\end{aligned}
$$

Therefore, we get the result by the triangle inequality.

We now show $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0 = \boldsymbol{\Omega}_{22}^- n^{-1} \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) + \mathbf{r}_1$ where $\mathbf{r}_1$ satisfies $\|\mathbf{r}_1\|_2 = o_p(n^{-1/2}N)$. (B.14) leads to $\mathbf{r}_1 = \mathbf{r}_{11} + \mathbf{r}_{12}$ where $\mathbf{r}_{11} \equiv \{\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{22}^-\} n^{-1} \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)$ and $\mathbf{r}_{12} \equiv \widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) n^{-1} \sum_{i=1}^n \{\mathbf{S}^*_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\}$. First, note that $\|\mathbf{r}_{12}\|_2 \asymp n^{-1}N \| \sum_{i=1}^n \{\mathbf{S}^*_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\}\|_2 = o(n^{-1/2}N^{1/2})$ by Lemmas B.3.1, B.3.5 and (B.15). In addition, $\|\boldsymbol{\Omega}_{22}^- n^{-1} \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\|_2 \asymp n^{-1}N \| \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\|_2 = O_p(n^{-1/2}N)$ by Lemmas B.3.1, B.3.5 and (B.16). Therefore, it suffices to show

$$
\|\mathbf{r}_{11}\|_2 = o_p\left(\left\| \boldsymbol{\Omega}_{22}^- n^{-1} \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) \right\|_2 \right).
$$

We will show this via proving

$$
\boldsymbol{\zeta}^{\mathrm{T}} \{\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{22}^-\} \boldsymbol{\zeta} = o_p(\boldsymbol{\zeta}^{\mathrm{T}} \boldsymbol{\Omega}_{22}^- \boldsymbol{\zeta}) = o_p(N \|\boldsymbol{\zeta}\|_2^2) \qquad \text{(B.17)}
$$

for $\boldsymbol{\zeta} \in \boldsymbol{\zeta}$ by Lemma B.3.5. Note that

$$
\boldsymbol{\zeta}^{\mathrm{T}} \{\boldsymbol{\Omega}_{22} - \widehat{\boldsymbol{\Omega}}_{22}^*(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*)\} \boldsymbol{\zeta} = E[\mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}} \boldsymbol{\zeta} \mid \mathbf{X}\}] - n^{-1} \sum_{i=1}^n \mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}} \boldsymbol{\zeta} \mid \mathbf{x}_i\}
$$

$$+n^{-1}\sum_{i=1}^{n}[\mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}_i\} - \mathrm{var}^*\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*\}].$$

We first have $w(\mathbf{x}) \equiv \mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}\} \leqslant CN^{-1}\|\boldsymbol{\zeta}\|_2^2$ for some constant $C > 0$ by Lemma B.3.5 and Condition (C1), and further $\mathrm{var}\{w(\mathbf{X})\} \leqslant C^2N^{-2}\|\boldsymbol{\zeta}\|_2^4$. Then by Bernstein's inequality,

$$\mathrm{pr}\left[\left|n^{-1}\sum_{i=1}^{n} w(\mathbf{x}_i) - E\{w(\mathbf{X})\}\right| \geqslant \epsilon\right] \leqslant 2\exp\left(-\frac{n^2\epsilon^2/2}{CnN^{-1}\|\boldsymbol{\zeta}\|_2^2\epsilon/3 + C^2nN^{-2}\|\boldsymbol{\zeta}\|_2^4}\right),$$

which approaches to zero when $\epsilon = C_\epsilon n^{-1/2}N^{-1}\|\boldsymbol{\zeta}\|_2^2$ for a sufficiently large C_ϵ, and this implies $E[\mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{X}\}] - n^{-1}\sum_{i=1}^{n}\mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}_i\} = O_p(n^{-1/2}N^{-1}\|\boldsymbol{\zeta}\|_2^2) = o_p(N^{-1}\|\boldsymbol{\zeta}\|_2^2)$. In addition, since $\|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N) = o_p(1)$ by construction, we have $\mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}\} - \mathrm{var}^*\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*\} = o_p[\mathrm{var}\{\mathbf{B}_+(Y)^{\mathrm{T}}\boldsymbol{\zeta} \mid \mathbf{x}\}] = o_p(N^{-1}\|\boldsymbol{\zeta}\|_2^2)$ uniformly in $\mathbf{x}$ by Conditions (C1), (C5) and Lemma B.3.5. Therefore, we have $\boldsymbol{\zeta}^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}_{22}^*(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{22}\}\boldsymbol{\zeta} = o_p(N^{-1}\|\boldsymbol{\zeta}\|_2^2)$. In addition, $\boldsymbol{\Omega}_{22}^-\boldsymbol{\zeta}, \widehat{\boldsymbol{\Omega}}_{22}^{*-}\boldsymbol{\zeta} \in \boldsymbol{\zeta}$ by Lemmas B.3.1 and B.3.2, and $\|\boldsymbol{\Omega}_{22}^-\boldsymbol{\zeta}\|_2, \|\widehat{\boldsymbol{\Omega}}_{22}^{*-}\boldsymbol{\zeta}\|_2 \asymp N\|\boldsymbol{\zeta}\|_2$ by Lemma B.3.5. Now, since $\boldsymbol{\Omega}_{22}\boldsymbol{\Omega}_{22}^-\boldsymbol{\zeta} = \boldsymbol{\zeta}$ and $\boldsymbol{\Omega}_{22}^*\boldsymbol{\Omega}_{22}^{*-}\boldsymbol{\zeta} = \boldsymbol{\zeta}$ by Lemmas B.3.1 and B.3.2, $\boldsymbol{\zeta}^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{22}^-\}\boldsymbol{\zeta} = \boldsymbol{\zeta}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*)\{\boldsymbol{\Omega}_{22} - \widehat{\boldsymbol{\Omega}}_{22}^*(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*)\}\boldsymbol{\Omega}_{22}^-\boldsymbol{\zeta} = o_p(N\|\boldsymbol{\zeta}\|_2^2)$, which shows (B.17). Therefore, we get $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0 = \boldsymbol{\Omega}_{22}^- n^{-1}\sum_{i=1}^{n}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) + \mathbf{r}_1$ where $\|\mathbf{r}_1\|_2 = o_p(n^{-1/2}N)$. In addition, we can write $\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{22}^- = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\mathrm{T}}$ for $\mathbf{Q}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$, $\mathbf{Q}^{\mathrm{T}}\mathbf{1} = \mathbf{0}$, and all diagonal elements of $\mathbf{D}$ are of order $o_p(N)$ by (B.17) and the fact that $\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*)\mathbf{1} = \boldsymbol{\Omega}_{22}^-\mathbf{1} = \mathbf{0}$ based on Lemma B.3.5. Then for $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2 \in \boldsymbol{\zeta}$, $\boldsymbol{\zeta}_1^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{22}^-\}\boldsymbol{\zeta}_2 = o_p(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2) = o_p(|\boldsymbol{\zeta}_1^{\mathrm{T}}\boldsymbol{\Omega}_{22}^-\boldsymbol{\zeta}_2|)$ by Lemma B.3.5. Further, using $\mathbf{1}^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{22}^-\}\mathbf{1} = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}_{22}^-\mathbf{1} = 0$, for $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2 \in \mathbb{R}^{m+1}$, we get

$$\boldsymbol{\zeta}_1^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{22}^-\}\boldsymbol{\zeta}_2 = o_p(|\boldsymbol{\zeta}_1^{\mathrm{T}}\boldsymbol{\Omega}_{22}^-\boldsymbol{\zeta}_2|). \tag{B.18}$$

Finally, we show $\|\widehat{c}(\cdot, \boldsymbol{\beta}_0) - c(\cdot)\|_\infty = O_p\{n^{-1/2}(N\log N)^{1/2}\}$. Note that $\|\widehat{c}(\cdot, \boldsymbol{\beta}_0) - c(\cdot)\|_\infty \leqslant \|\mathbf{B}_+(\cdot)^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}\|_\infty + \|\mathbf{B}_+(\cdot)^{\mathrm{T}}\boldsymbol{\zeta}_0 - c(\cdot)\|_\infty \asymp \|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_\infty + O(N^{-q})$ by Lemma B.3.4 and Condition (C5), and $N^{-q} = o\{n^{-1/2}(N\log N)^{1/2}\}$ by Condition (C3), hence it suffices to show that $\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_\infty = O_p\{n^{-1/2}(N\log N)^{1/2}\}$. Let $\boldsymbol{\zeta} \in \boldsymbol{\zeta}$, then we have $\boldsymbol{\zeta}^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{22}^-\}\sum_{i=1}^{n}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) = o_p\{|\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\Omega}_{22}^-\sum_{i=1}^{n}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)|\}$ by (B.18). In addition, we have

$$\sum_{i=1}^{n}\boldsymbol{\zeta}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{22}^{*-}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*)\{\mathbf{S}_{\boldsymbol{\zeta}}^*(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\}$$

$$
\begin{aligned}
&\asymp N\|\boldsymbol{\zeta}\|_2 \left\| \sum_{i=1}^{n} \{\mathbf{S}^*_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\} \right\|_2 \\
&= o(n^{1/2} N^{1/2} \|\boldsymbol{\zeta}\|_2)
\end{aligned}
$$

by Lemmas B.3.1, B.3.5 and (B.15), and

$$
\sum_{i=1}^{n} \boldsymbol{\zeta}^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) \asymp_p n^{1/2} N^{1/2} \|\boldsymbol{\zeta}\|_2 \tag{B.19}
$$

because $E\{\boldsymbol{\zeta}^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{S}_{\boldsymbol{\zeta}}(Y, \mathbf{X})\} = 0$ and $\mathrm{var}\{\boldsymbol{\zeta}^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{S}_{\boldsymbol{\zeta}}(Y, \mathbf{X})\} = \boldsymbol{\zeta}^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \boldsymbol{\zeta} \asymp N\|\boldsymbol{\zeta}\|_2^2$ by Lemma B.3.5. Then (B.14) leads to

$$
\begin{aligned}
\boldsymbol{\zeta}^{\mathrm{T}} \{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\} &= n^{-1} \sum_{i=1}^{n} \Big[\boldsymbol{\zeta}^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) + \boldsymbol{\zeta}^{\mathrm{T}} \{\widehat{\boldsymbol{\Omega}}^{*-}_{22}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}^{-}_{22}\} \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) \\
&\qquad + \boldsymbol{\zeta}^{\mathrm{T}} \widehat{\boldsymbol{\Omega}}^{*-}_{22}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) \{\mathbf{S}^*_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\} \Big] \\
&= n^{-1} \sum_{i=1}^{n} \boldsymbol{\zeta}^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) + o_p(n^{-1/2} N^{1/2} \|\boldsymbol{\zeta}\|_2) \qquad \text{(B.20)} \\
&\asymp_p n^{-1/2} N^{1/2} \|\boldsymbol{\zeta}\|_2. \qquad \text{(B.21)}
\end{aligned}
$$

Now, note that $\mathbf{e}_k = \mathbf{e}_k - (m+1)^{-1}\mathbf{1} + (m+1)^{-1}\mathbf{1}$, $\mathbf{e}_k - (m+1)^{-1}\mathbf{1} \in \boldsymbol{\zeta}$, and $\|\mathbf{e}_k - (m+1)^{-1}\mathbf{1}\|_2 \asymp 1$ by Condition (C3). Then we have $\mathbf{e}_k^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{S}_{\boldsymbol{\zeta}}(y, \mathbf{x}) \leqslant C_1 N$ for some constant $C_1 > 0$ uniformly in $(y, \mathbf{x})$ by Condition (C1) and Lemma B.3.5, since $\|\mathbf{S}_{\boldsymbol{\zeta}}(y, \mathbf{x})\|_2 \leqslant \|\mathbf{B}_+(y)\|_2 + E\{\|\mathbf{B}_+(Y)\|_2 | \mathbf{x}\} \leqslant \|\mathbf{B}_+(y)\|_1 + E\{\|\mathbf{B}_+(Y)\|_1 | \mathbf{x}\} = 2$. In addition, we have $\mathrm{var}\{\mathbf{e}_k^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{S}_{\boldsymbol{\zeta}}(Y, \mathbf{X})\} = \mathbf{e}_k^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{e}_k \leqslant C_2 N$ for some constant $C_2 > 0$ by Lemmas B.3.2 and B.3.5. Hence, by Bernstein's inequality, for $\epsilon = C_\epsilon n^{-1/2} (N \log N)^{1/2}$ where C_ϵ is a sufficiently large constant, we get

$$
\begin{aligned}
&\mathrm{pr}\left\{ \|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_\infty \geqslant \epsilon \right\} \\
\leqslant\ & \sum_{k=1}^{m+1} \mathrm{pr}\left[|\mathbf{e}_k^{\mathrm{T}} \{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}| \geqslant \epsilon \right] \\
=\ & \sum_{k=1}^{m+1} \mathrm{pr}\left\{ \left| n^{-1} \sum_{i=1}^{n} \mathbf{e}_k^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) + o_p(n^{-1/2} N^{1/2}) \right| \geqslant \epsilon \right\} \\
\leqslant\ & \sum_{k=1}^{m+1} \mathrm{pr}\left\{ \left| n^{-1} \sum_{i=1}^{n} \mathbf{e}_k^{\mathrm{T}} \boldsymbol{\Omega}^{-}_{22} \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) \right| \geqslant \epsilon/2 \right\}
\end{aligned}
$$

$$\leqslant \quad 2(m+1)\exp\left(-\frac{n^2\epsilon^2/8}{C_1 nN\epsilon/6 + C_2 nN}\right) \to 0$$

when $n \to 0$ by Condition (C3). This implies $\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_\infty = O_p\{n^{-1/2}(N\log N)^{1/2}\}$.

Now we show the results in Proposition 3.3.1. We first have $\|\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0\|_2 = \|\mathbf{T}^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}\|_2 = (\sum_{k=2}^{m+1}[-\mathbf{e}_1^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\} + \mathbf{e}_k^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}]^2)^{1/2} \leqslant \|\mathbf{e}_1^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}\mathbf{1}_m\|_2 + \|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_2 \asymp \sqrt{N}|\mathbf{e}_1^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}| + \|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$ because $\mathbf{e}_1^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\} \asymp_p n^{-1/2}N^{1/2}$ by (B.21) and $\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$. In addition, by Lemma B.3.3,

$$\begin{aligned}
\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0 &= \mathbf{T}^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\} \\
&= \mathbf{T}^{\mathrm{T}}\{\boldsymbol{\Omega}_{22}^{-}n^{-1}\sum_{i=1}^{n}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) + \mathbf{r}_1\} \\
&= \boldsymbol{\Sigma}_{22}^{-1}n^{-1}\sum_{i=1}^{n}[\mathbf{B}(y_i) - E\{\mathbf{B}(Y) \mid \mathbf{x}_i\}] + \mathbf{T}^{\mathrm{T}}\mathbf{r}_1,
\end{aligned}$$

where $\|\mathbf{T}^{\mathrm{T}}\mathbf{r}_1\|_2 = (\sum_{k=2}^{m+1}[-\mathbf{e}_1^{\mathrm{T}}\mathbf{r}_1 + \mathbf{e}_k^{\mathrm{T}}\mathbf{r}_1]^2)^{1/2} \leqslant \|\mathbf{e}_1^{\mathrm{T}}\mathbf{r}_1\mathbf{1}_m\|_2 + \|\mathbf{r}_1\|_2 \asymp \sqrt{N}|\mathbf{e}_1^{\mathrm{T}}\mathbf{r}_1| + \|\mathbf{r}_1\|_2 = o_p(n^{-1/2}N)$, because $\|\mathbf{r}_1\|_2 = o_p(n^{-1/2}N)$ and $\mathbf{e}_1^{\mathrm{T}}\mathbf{r}_1 = o_p(n^{-1/2}N^{1/2})$ by (B.20). Lastly, to show $\|\widehat{c}(\cdot, \boldsymbol{\beta}_0) - c(\cdot)\|_\infty = O_p\{n^{-1/2}(N\log N)^{1/2}\}$, since $\|\widehat{c}(\cdot, \boldsymbol{\beta}_0) - c(\cdot)\|_\infty \leqslant \|\mathbf{B}(\cdot)^{\mathrm{T}}\{\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0\}\|_\infty + \|\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}_0 - c(\cdot)\|_\infty \asymp \|\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0\|_\infty + O(N^{-q})$ by Lemma B.3.4 and Condition (C5), and $N^{-q} = o\{n^{-1/2}(N\log N)^{1/2}\}$ by Condition (C3), it suffices to show $\|\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0\|_\infty = O_p\{n^{-1/2}(N\log N)^{1/2}\}$. We have $\|\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0\|_\infty = \|\mathbf{T}^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}\|_\infty = \max_{k=2,\ldots,m+1}|-\mathbf{e}_1^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\} + \mathbf{e}_k^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}| \leqslant |\mathbf{e}_1^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}| + \|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_\infty \leqslant 2\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_\infty = O_p\{n^{-1/2}(N\log N)^{1/2}\}$. This completes the proof. □

B.5 Proof of Proposition 3.3.2

We first show $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$. Since $\sum_{i=1}^{n}\mathbf{S}_{\boldsymbol{\zeta}}^{*}\{y_i, \mathbf{x}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\} = \mathbf{0}$ for all $\boldsymbol{\beta}$, we have $\mathbf{0} = d\left[\sum_{i=1}^{n}\mathbf{S}_{\boldsymbol{\zeta}}^{*}\{y_i, \mathbf{x}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\}\right]/d\boldsymbol{\beta}^{\mathrm{T}} = -n\widehat{\boldsymbol{\Omega}}_{21}^{*}\{\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\} - n\widehat{\boldsymbol{\Omega}}_{22}^{*}\{\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\}\partial\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}$ for all $\boldsymbol{\beta}$. Thus,

$$\frac{\partial\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}} = -\widehat{\boldsymbol{\Omega}}_{22}^{*-}\{\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\}\widehat{\boldsymbol{\Omega}}_{21}^{*}\{\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\} \tag{B.22}$$

since $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}) \in \boldsymbol{\zeta}$. For any $\boldsymbol{\beta}$, we also have

$$\begin{aligned}\frac{dl\{\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\}}{d\boldsymbol{\beta}} &= \sum_{i=1}^{n} \mathbf{S}^*_{\boldsymbol{\beta}}\{y_i, \mathbf{x}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\} \\ &= \sum_{i=1}^{n} \mathbf{S}^*_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\zeta}_0) - n\widehat{\boldsymbol{\Omega}}^*_{12}(\boldsymbol{\beta}, \boldsymbol{\zeta}^*)\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}) - \boldsymbol{\zeta}_0\},\end{aligned}$$

where $\boldsymbol{\zeta}^* \equiv \alpha\boldsymbol{\zeta}_0 + (1-\alpha)\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})$ for some $\alpha \in (0,1)$, and

$$\begin{aligned}\frac{d^2 l\{\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\}}{d\boldsymbol{\beta} d\boldsymbol{\beta}^{\mathrm{T}}} &= -n\widehat{\boldsymbol{\Omega}}^*_{11}\{\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\} - n\widehat{\boldsymbol{\Omega}}^*_{12}\{\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\}\frac{\partial\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}} \\ &= -n\widehat{\boldsymbol{\Omega}}^{*-1}_{\boldsymbol{\beta}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})\}.\end{aligned}$$

Now, since $\sum_{i=1}^{n} \mathbf{S}^*_{\boldsymbol{\beta}}\{y_i, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})\} = \mathbf{0}$, for some $\boldsymbol{\beta}^* \equiv \alpha_1\boldsymbol{\beta}_0 + (1-\alpha_1)\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\zeta}^* \equiv \alpha_2\boldsymbol{\zeta}_0 + (1-\alpha_2)\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)$ where $\alpha_1, \alpha_2 \in (0,1)$,

$$\begin{aligned}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &= -\left[\frac{d^2 l\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}}{d\boldsymbol{\beta} d\boldsymbol{\beta}^{\mathrm{T}}}\right]^{-1} \frac{dl\{\boldsymbol{\beta}_0, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)\}}{d\boldsymbol{\beta}} \qquad \text{(B.23)} \\ &= \widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\left[n^{-1}\sum_{i=1}^{n} \mathbf{S}^*_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \widehat{\boldsymbol{\Omega}}^*_{12}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*)\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}\right].\end{aligned}$$

Note that

$$\|\widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\zeta})\|_2 = O(1) \qquad \text{(B.24)}$$

by Lemma B.3.2. On the other hand, we have

$$\begin{aligned}\left\|n^{-1}\sum_{i=1}^{n}\{\mathbf{S}^*_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i)\}\right\|_2 &\leqslant \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_2 |E^*(Y \mid \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - E(Y \mid \mathbf{x})| \\ &\leqslant \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_2 \|f^*_{Y|\mathbf{X}}(\cdot, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - f_{Y|\mathbf{X}}(\cdot, \mathbf{x})\|_\infty \\ &= O(N^{-q}) = o(n^{-1/2}N^{-1/2}) \qquad \text{(B.25)}\end{aligned}$$

by Conditions (C1), (C3), and (C5). Also,

$$\left\|\sum_{i=1}^{n} \mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i)\right\|_2 \asymp_p n^{1/2} \qquad \text{(B.26)}$$

because $E\{\|\sum_{i=1}^n \mathbf{S}_{\boldsymbol{\beta}}(Y_i, \mathbf{X}_i)\|_2^2\} = nE\{\|\mathbf{S}_{\boldsymbol{\beta}}(Y, \mathbf{X})\|_2^2\} = n\,\mathrm{trace}(\boldsymbol{\Omega}_{11}) \asymp n$ by Condition (C6). Now we will prove

$$\|\{\widehat{\boldsymbol{\Omega}}_{21}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{21}\}\boldsymbol{\beta}\|_2 = o_p(N^{-1/2}\|\boldsymbol{\beta}\|_2) \tag{B.27}$$

for $\boldsymbol{\beta} \in \mathbb{R}^p$. Note that $\|f^*_{Y|\mathbf{X}}(\cdot, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(\cdot, \mathbf{x})\|_\infty = o_p(1)$ by the construction of $\boldsymbol{\zeta}^*$ and Proposition 3.3.1, so we have

$$\begin{aligned}
& |\mathrm{cov}^*\{B_k(Y), Y \mid \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*\} - \mathrm{cov}\{B_k(Y), Y \mid \mathbf{x}\}| \\
\leqslant\ & \left|\int B_k(y) y \{f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(y, \mathbf{x})\} dy\right| \\
& + \left|\int B_k(y)\{f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(y, \mathbf{x})\} dy\right| E^*(Y \mid \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) \\
& + E\{B_k(Y) \mid \mathbf{x}\} \left|\int y\{f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(y, \mathbf{x})\} dy\right| \\
=\ & O\left\{\|B_k(\cdot)\|_1 \|f^*_{Y|\mathbf{X}}(\cdot, \mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(\cdot, \mathbf{x})\|_\infty\right\} = o_p(N^{-1})
\end{aligned}$$

uniformly in $\mathbf{x}$ by Condition (C1) and Lemma B.3.4. This further implies

$$\begin{aligned}
& \left|n^{-1}\sum_{i=1}^n [\mathrm{cov}^*\{B_k(Y), Y \mid \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*\} - \mathrm{cov}\{B_k(Y), Y \mid \mathbf{x}_i\}]\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right| \\
\leqslant\ & \|\boldsymbol{\beta}\|_2 n^{-1}\sum_{i=1}^n \|\mathbf{x}_i\|_2 |\mathrm{cov}^*\{B_k(Y), Y \mid \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*\} - \mathrm{cov}\{B_k(Y), Y \mid \mathbf{x}_i\}| \\
=\ & o_p(N^{-1})
\end{aligned}$$

by Condition (C1). In addition, $n^{-1}\sum_{i=1}^n \mathrm{cov}\{B_k(Y), Y \mid \mathbf{x}_i\}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - E[\mathrm{cov}\{B_k(Y), Y \mid \mathbf{X}\}\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}] = O_p(n^{-1/2}N^{-1/2}\|\boldsymbol{\beta}\|_2)$ since $\mathrm{var}[\mathrm{cov}\{B_k(Y), Y \mid \mathbf{X}\}\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}] \leqslant E[\mathrm{var}\{B_k(Y) \mid \mathbf{X}\}\mathrm{var}(Y \mid \mathbf{X})(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})^2] = E[\mathbf{e}_k^{\mathrm{T}}\mathbf{I}_{\boldsymbol{\zeta}}(\mathbf{X})\mathbf{e}_k \mathrm{var}(Y \mid \mathbf{X})\|\mathbf{X}\|_2^2]\|\boldsymbol{\beta}\|_2^2 = O(N^{-1}\|\boldsymbol{\beta}\|_2^2)$ by Condition (C1) and Lemma B.3.5. Therefore, we get

$$\begin{aligned}
& \left\|\{\widehat{\boldsymbol{\Omega}}_{21}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{21}\}\boldsymbol{\beta}\right\|_\infty \\
=\ & \max_{k=1,\ldots,m+1} \left|\mathbf{e}_k^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}_{21}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{21}\}\boldsymbol{\beta}\right| \\
\leqslant\ & \max_{k=1,\ldots,m+1} \left|n^{-1}\sum_{i=1}^n [\mathrm{cov}^*\{B_k(Y), Y \mid \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}^*\} - \mathrm{cov}\{B_k(Y), Y \mid \mathbf{x}_i\}]\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right| \\
& + \max_{k=1,\ldots,m+1} \left|n^{-1}\sum_{i=1}^n \mathrm{cov}\{B_k(Y), Y \mid \mathbf{x}_i\}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - E[\mathrm{cov}\{B_k(Y), Y \mid \mathbf{X}\}\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}]\right|
\end{aligned}$$

$$= \quad o_p(N^{-1}\|\boldsymbol{\beta}\|_2)$$

by Condition (C3), and this leads to (B.27). In addition, we have

$$\|\boldsymbol{\Omega}_{21}\boldsymbol{\beta}\|_2 = O(N^{-1/2}\|\boldsymbol{\beta}\|_2) \tag{B.28}$$

for $\boldsymbol{\beta} \in \mathbb{R}^p$. This is because $\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\Omega}_{21})\boldsymbol{\beta} > 0$ for $\boldsymbol{\beta} \neq \mathbf{0}$ by Lemma B.3.2, which implies $N\|\boldsymbol{\Omega}_{21}\boldsymbol{\beta}\|_2^2 \asymp \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\Omega}_{21}\boldsymbol{\beta} < \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{11}\boldsymbol{\beta} \asymp \|\boldsymbol{\beta}\|_2^2$ by Lemma B.3.5 and Condition (C6). Then (B.21), (B.27), and (B.28) lead to

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\Omega}}_{12}^{*}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^{*})\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}\right\|_2^2 &= \sum_{j=1}^{p}\left[\mathbf{e}_j^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{12}^{*}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^{*})\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}\right]^2 \\
&\asymp_p n^{-1}N\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{12}^{*}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^{*})\mathbf{e}_j\|_2^2 \\
&= O_p(n^{-1}).
\end{aligned}$$

Therefore, combining this with (B.23), (B.24), (B.25), and (B.26), we obtain $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by the triangle inequality.

Now, we will show $\|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\beta}}^{*}\{\boldsymbol{\beta}^{*}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{*})\} - \boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 = o_p(1)$ where $\boldsymbol{\beta}^{*}$ is given in (B.23). First note that $\|\boldsymbol{\beta}^{*} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2}) = o_p(1)$ by the construction of $\boldsymbol{\beta}^{*}$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$. We now show $\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{*}) - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$. Note that $\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$ by Proposition 3.3.1. In addition, it is easy to show

$$\|\widehat{\boldsymbol{\Omega}}_{21}^{*}(\boldsymbol{\beta}^{*}, \boldsymbol{\zeta}^{*})\boldsymbol{\beta}\|_2 = O(N^{-1/2}\|\boldsymbol{\beta}\|_2) \tag{B.29}$$

for fixed $\boldsymbol{\beta}^{*}, \boldsymbol{\zeta}^{*}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ by a similar way to showing (B.28). Then this leads to $\|\partial\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2 = O_p(n^{-1/2}N^{1/2})$ by (B.22), Lemma B.3.5, and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$. Hence, by the Taylor expansion with $\boldsymbol{\beta}^{**} \equiv \alpha\boldsymbol{\beta}_0 + (1-\alpha)\boldsymbol{\beta}^{*}$ for some $\alpha \in (0, 1)$, we have

$$\left\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{*}) - \boldsymbol{\zeta}_0\right\|_2 \leqslant \left\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\right\|_2 + \left\|\frac{\partial\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{**})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right\|_2 = O_p(n^{-1/2}N) = o_p(1)$$

by Condition (C3). Then we have $\|f_{Y|\mathbf{X}}^{*}\{\cdot, \mathbf{x}, \boldsymbol{\beta}^{*}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{*})\} - f_{Y|\mathbf{X}}(\cdot, \mathbf{x})\|_\infty = o_p(1)$ uniformly

in $\mathbf{x}$ by Condition (C1), which leads to

$$\begin{aligned}\|\widehat{\boldsymbol{\Omega}}^*_{11}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\} - \boldsymbol{\Omega}_{11}\|_2 &\leqslant \left\| n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}[\mathrm{var}^*\{Y \mid \mathbf{x}_i, \boldsymbol{\beta}^*, \boldsymbol{\zeta}(\boldsymbol{\beta}^*)\} - \mathrm{var}(Y \mid \mathbf{x}_i)]\right\|_2 \\ &\quad + \left\| n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathrm{var}(Y \mid \mathbf{x}_i) - E\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathrm{var}(Y \mid \mathbf{X})\}\right\|_2 \\ &= o_p(1) \end{aligned} \tag{B.30}$$

by Condition (C1). In addition, using $\|f^*_{Y|\mathbf{X}}\{\cdot, \mathbf{x}, \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\} - f_{Y|\mathbf{X}}(\cdot, \mathbf{x})\|_\infty = o_p(1)$, we have

$$\boldsymbol{\zeta}_1^{\mathrm{T}}[\widehat{\boldsymbol{\Omega}}^{*-}_{22}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\} - \boldsymbol{\Omega}^-_{22}]\boldsymbol{\zeta}_2 = o_p(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2) \tag{B.31}$$

for $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2 \in \mathbb{R}^{m+1}$ from a similar argument to showing (B.18), and

$$\|\{\widehat{\boldsymbol{\Omega}}_{21}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\} - \boldsymbol{\Omega}_{21}\}\boldsymbol{\beta}\|_2 = o_p(N^{-1/2}\|\boldsymbol{\beta}\|_2) \tag{B.32}$$

for $\boldsymbol{\beta} \in \mathbb{R}^p$ from a similar argument to showing (B.27). Therefore,

$$\begin{aligned} &\|\widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\} - \boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 \\ \leqslant\ &\|\widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\|_2\|\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Omega}}^{*-1}_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\|_2\|\boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 \\ =\ &\|\widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\|_2 \\ &\times\|\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}^-_{22}\boldsymbol{\Omega}_{21} - \widehat{\boldsymbol{\Omega}}^*_{11}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\} + \widehat{\boldsymbol{\Omega}}^*_{12}\widehat{\boldsymbol{\Omega}}^{*-}_{22}\widehat{\boldsymbol{\Omega}}^*_{21}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\|_2\|\boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 \\ =\ &O\left(\left\|\boldsymbol{\Omega}_{11} - \widehat{\boldsymbol{\Omega}}^*_{11}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\right\|_2 + \left\|\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}^-_{22}\boldsymbol{\Omega}_{21} - (\widehat{\boldsymbol{\Omega}}^*_{12}\widehat{\boldsymbol{\Omega}}^{*-}_{22}\widehat{\boldsymbol{\Omega}}^*_{21})\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\right\|_2\right) \\ =\ &O\left(\left\|\boldsymbol{\Omega}_{11} - \widehat{\boldsymbol{\Omega}}^*_{11}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\right\|_2 + \left\|\left[\boldsymbol{\Omega}_{12} - \widehat{\boldsymbol{\Omega}}^*_{12}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\right]\boldsymbol{\Omega}^-_{22}\boldsymbol{\Omega}_{21}\right\|_2\right. \\ &+ \left\|\widehat{\boldsymbol{\Omega}}^*_{12}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\left[\boldsymbol{\Omega}^-_{22} - \widehat{\boldsymbol{\Omega}}^{*-}_{22}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\right]\boldsymbol{\Omega}_{21}\right\|_2 \\ &\left.+ \left\|(\widehat{\boldsymbol{\Omega}}^*_{12}\widehat{\boldsymbol{\Omega}}^{*-}_{22})\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\left[\boldsymbol{\Omega}_{21} - \widehat{\boldsymbol{\Omega}}^*_{21}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\right]\right\|_2\right) \\ =\ &o_p(1), \end{aligned} \tag{B.33}$$

where the first equality used the definition, the second equality is by Lemma B.3.2, and the last equality follows from Lemma B.3.5, (B.28), (B.29), (B.30), (B.31), and (B.32).

Finally, we show $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{\Omega}_{\boldsymbol{\beta}}\, n^{-1}\sum_{i=1}^n\{\mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}^-_{22}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\} + \mathbf{r}_2$ where $\mathbf{r}_2$

satisfies $\|\mathbf{r}_2\|_2 = o_p(n^{-1/2})$. From (B.23), we have $\mathbf{r}_2 = \mathbf{r}_{21} + \mathbf{r}_{22} + \mathbf{r}_{23}$ where

$$\begin{aligned}
\mathbf{r}_{21} &\equiv \left[\widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\} - \boldsymbol{\Omega}_{\boldsymbol{\beta}}\right] n^{-1}\sum_{i=1}^{n}\{\mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\}, \\
\mathbf{r}_{22} &\equiv \widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\} n^{-1}\sum_{i=1}^{n}\{\mathbf{S}^*_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - \mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i)\}, \\
\mathbf{r}_{23} &\equiv \widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\left[-\widehat{\boldsymbol{\Omega}}^*_{12}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*)\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\} + n^{-1}\sum_{i=1}^{n}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\right].
\end{aligned}$$

First, we get $\|\mathbf{r}_{21}\|_2 = o_p(n^{-1/2})$ from (B.33), (B.26), and $\|n^{-1}\sum_{i=1}^{n}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\}\|_2^2 = n^{-2}\sum_{j=1}^{p}\{\sum_{i=1}^{n}\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\}^2 \asymp_p n^{-1}N\sum_{j=1}^{p}\|\boldsymbol{\Omega}_{21}\mathbf{e}_j\|^2 = O_p(n^{-1})$ by (B.19) and (B.28). In addition, we get $\|\mathbf{r}_{22}\|_2 = o_p(n^{-1/2})$ from (B.24) and (B.25). Further, we have $\|\mathbf{r}_{23}\|_2 = o_p(n^{-1/2})$. This is because $\|\widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*)\}\|_2 = O(1)$ by (B.24),

$$\begin{aligned}
\left\|\{\widehat{\boldsymbol{\Omega}}^*_{12}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{12}\}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}\right\|_2^2 &= \sum_{j=1}^{p}\left[\mathbf{e}_j^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}^*_{12}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{12}\}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\}\right]^2 \\
&\asymp_p n^{-1}N\sum_{j=1}^{p}\|\{\widehat{\boldsymbol{\Omega}}^*_{21}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*) - \boldsymbol{\Omega}_{21}\}\mathbf{e}_j\|_2^2 \\
&= o_p(n^{-1})
\end{aligned}$$

by (B.21) and (B.27),

$$\begin{aligned}
&\left\|\boldsymbol{\Omega}_{12}\left\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0 - n^{-1}\sum_{i=1}^{n}\boldsymbol{\Omega}_{22}^{-}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\right\}\right\|_2^2 \\
= &\sum_{j=1}^{p}\left[\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Omega}_{12}\left\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0 - n^{-1}\sum_{i=1}^{n}\boldsymbol{\Omega}_{22}^{-}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\right\}\right]^2 \\
= &\, o_p(n^{-1}N)\sum_{j=1}^{p}\|\boldsymbol{\Omega}_{21}\mathbf{e}_j\|^2 \\
= &\, o_p(n^{-1})
\end{aligned}$$

by (B.20) and (B.28), and these lead to

$$\left\|\widehat{\boldsymbol{\Omega}}^*_{12}(\boldsymbol{\beta}_0, \boldsymbol{\zeta}^*)\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0) - \boldsymbol{\zeta}_0\} - n^{-1}\sum_{i=1}^{n}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\right\|_2$$

$$
\begin{aligned}
&\leqslant \left\|\{\widehat{\boldsymbol{\Omega}}^{*}_{12}(\boldsymbol{\beta}_0,\boldsymbol{\zeta}^{*})-\boldsymbol{\Omega}_{12}\}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)-\boldsymbol{\zeta}_0\}\right\|_2+\left\|\boldsymbol{\Omega}_{12}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)-\boldsymbol{\zeta}_0-n^{-1}\sum_{i=1}^{n}\boldsymbol{\Omega}^{-}_{22}\mathbf{S}_{\boldsymbol{\zeta}}(y_i,\mathbf{x}_i)\}\right\|_2\\
&= o_p(n^{-1/2}).
\end{aligned}
$$

Therefore, by Lemma B.3.3,

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 &= \boldsymbol{\Omega}_{\boldsymbol{\beta}}\, n^{-1}\sum_{i=1}^{n}\{\mathbf{S}_{\boldsymbol{\beta}}(y_i,\mathbf{x}_i)-\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}^{-}_{22}\mathbf{S}_{\boldsymbol{\zeta}}(y_i,\mathbf{x}_i)\}+\mathbf{r}_2\\
&= \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\, n^{-1}\sum_{i=1}^{n}(\mathbf{S}_{\boldsymbol{\beta}}(y_i,\mathbf{x}_i)-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}[\mathbf{B}(y_i)-E\{\mathbf{B}(Y)\mid\mathbf{x}_i)])+\mathbf{r}_2
\end{aligned}
$$

and $\mathbf{r}_2$ satisfies $\|\mathbf{r}_2\|_2 \leqslant \|\mathbf{r}_{21}\|_2+\|\mathbf{r}_{22}\|_2+\|\mathbf{r}_{23}\|_2 = o_p(n^{-1/2})$. Now, noting that $\mathrm{var}\{\mathbf{S}_{\boldsymbol{\beta}}(Y,\mathbf{X})-\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}^{-}_{22}\mathbf{S}_{\boldsymbol{\zeta}}(Y,\mathbf{X})\} = \boldsymbol{\Omega}_{11}+\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}^{-}_{22}\boldsymbol{\Omega}_{21}-2\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}^{-}_{22}\boldsymbol{\Omega}_{21} = \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$ by Lemma B.3.3, we get $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\to N(\mathbf{0},\mathbf{I})$ as $n\to\infty$. □

B.6 Addtional lemma

We introduce an additional lemma for the asymptotic properties of the profile estimators.

Lemma B.6.1. *Under Conditions (C1)-(C6),* $\|\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$, $\boldsymbol{\zeta}^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0) = O_p(n^{-1/2}N^{1/2}\|\boldsymbol{\zeta}\|_2)$ *for any conformal vector* $\boldsymbol{\zeta}$*, and*

$$
\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^{\mathrm{T}},(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0)^{\mathrm{T}}\}^{\mathrm{T}} = \boldsymbol{\Omega}^{-}n^{-1}\sum_{i=1}^{n}\mathbf{S}(y_i,\mathbf{x}_i)+(\mathbf{r}_2^{\mathrm{T}},\mathbf{r}_3^{\mathrm{T}})^{\mathrm{T}},
$$

where $\mathbf{r}_2\in\mathbb{R}^p$ *and* $\mathbf{r}_3\in\mathbb{R}^{m+1}$ *satisfy* $\|\mathbf{r}_2\|_2 = o_p(n^{-1/2})$, $\|\mathbf{r}_3\|_2 = o_p(n^{-1/2}N)$, *and* $|\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{r}_3| = o_p(n^{-1/2}N^{1/2}\|\boldsymbol{\zeta}\|_2)$.

Proof. Note that $\widehat{\boldsymbol{\zeta}} = \widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})$. The Taylor expansion with $\boldsymbol{\beta}^{*}\equiv\alpha\boldsymbol{\beta}_0+(1-\alpha)\widehat{\boldsymbol{\beta}}$ for some $\alpha\in(0,1)$ and (B.22) gives

$$
\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0 = \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)-\boldsymbol{\zeta}_0-\widehat{\boldsymbol{\Omega}}^{*-}_{22}\{\boldsymbol{\beta}^{*},\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{*})\}\widehat{\boldsymbol{\Omega}}^{*}_{21}\{\boldsymbol{\beta}^{*},\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{*})\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0).
$$

Then we get $\|\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$ because $\|\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)-\boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$ by Proposition 3.3.1, and $\|\widehat{\boldsymbol{\Omega}}^{*-}_{22}\{\boldsymbol{\beta}^{*},\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{*})\}\widehat{\boldsymbol{\Omega}}^{*}_{21}\{\boldsymbol{\beta}^{*},\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{*})\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2 \asymp N\|\widehat{\boldsymbol{\Omega}}^{*}_{21}\{\boldsymbol{\beta}^{*},\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{*})\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2 = O(N^{1/2}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2) = O_p(n^{-1/2}N^{1/2})$ by Lemma B.3.5, (B.29), and Proposition 3.3.2. Also, $\boldsymbol{\zeta}^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}_0)-\boldsymbol{\zeta}_0\}\asymp_p n^{-1/2}N^{1/2}\|\boldsymbol{\zeta}\|_2$ from (B.21) gives $\boldsymbol{\zeta}^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0) = O_p(n^{-1/2}N^{1/2}\|\boldsymbol{\zeta}\|_2)$.

Further, Propositions 3.3.1 and 3.3.2 lead to

$$\begin{aligned}
\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0 &= \boldsymbol{\Omega}_{22}^{-} n^{-1} \sum_{i=1}^{n} \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) + \mathbf{r}_1 \\
&\quad - \boldsymbol{\Omega}_{22}^{-} \boldsymbol{\Omega}_{21} \left[\boldsymbol{\Omega}_{\boldsymbol{\beta}}\, n^{-1} \sum_{i=1}^{n} \{ \mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) - \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-} \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) \} + \mathbf{r}_2 \right] \\
&\quad + \left[\boldsymbol{\Omega}_{22}^{-} \boldsymbol{\Omega}_{21} - \widehat{\boldsymbol{\Omega}}_{22}^{*-} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \} \widehat{\boldsymbol{\Omega}}_{21}^{*} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \} \right] (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&= n^{-1} \sum_{i=1}^{n} \{ - \boldsymbol{\Omega}_{22}^{-} \boldsymbol{\Omega}_{21} \boldsymbol{\Omega}_{\boldsymbol{\beta}} \mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) + (\boldsymbol{\Omega}_{22}^{-} + \boldsymbol{\Omega}_{22}^{-} \boldsymbol{\Omega}_{21} \boldsymbol{\Omega}_{\boldsymbol{\beta}} \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-}) \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) \} + \mathbf{r}_3
\end{aligned}$$

where $\mathbf{r}_3 = \mathbf{r}_1 - \boldsymbol{\Omega}_{22}^{-} \boldsymbol{\Omega}_{21} \mathbf{r}_2 + [\boldsymbol{\Omega}_{22}^{-} \boldsymbol{\Omega}_{21} - \widehat{\boldsymbol{\Omega}}_{22}^{*-} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \} \widehat{\boldsymbol{\Omega}}_{21}^{*} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \}] (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. First, we have $\|\mathbf{r}_1\|_2 = o_p(n^{-1/2} N)$ from Proposition 3.3.1 and $|\boldsymbol{\zeta}^{\mathrm{T}} \mathbf{r}_1| = o_p(n^{-1/2} N^{1/2} \|\boldsymbol{\zeta}\|_2)$ from (B.20). Also, $\|\boldsymbol{\Omega}_{22}^{-} \boldsymbol{\Omega}_{21} \mathbf{r}_2\|_2 \asymp N \|\boldsymbol{\Omega}_{21} \mathbf{r}_2\|_2 = O(N^{1/2} \|\mathbf{r}_2\|_2) = o_p(n^{-1/2} N^{1/2})$ by Lemma B.3.5, (B.28), and $\|\mathbf{r}_2\|_2 = o_p(n^{-1/2})$ from Proposition 3.3.2. In addition,

$$\begin{aligned}
& \|[\boldsymbol{\Omega}_{22}^{-} \boldsymbol{\Omega}_{21} - \widehat{\boldsymbol{\Omega}}_{22}^{*-} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \} \widehat{\boldsymbol{\Omega}}_{21}^{*} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \}] (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2 \\
\leqslant\ & \|[\boldsymbol{\Omega}_{22}^{-} - \widehat{\boldsymbol{\Omega}}_{22}^{*-} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \}] \boldsymbol{\Omega}_{21} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2 \\
& + \|\widehat{\boldsymbol{\Omega}}_{22}^{*-} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \} [\boldsymbol{\Omega}_{21} - \widehat{\boldsymbol{\Omega}}_{21}^{*} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \}] (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2 \\
\asymp\ & o_p(N) \|\boldsymbol{\Omega}_{21} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2 + O_p(N \|[\boldsymbol{\Omega}_{21} - \widehat{\boldsymbol{\Omega}}_{21}^{*} \{ \boldsymbol{\beta}^*, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^*) \}] (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2) \\
=\ & o_p(n^{-1/2} N^{1/2}).
\end{aligned}$$

The second argument is by (B.31) and Lemma B.3.5, and the last equality is by (B.28), (B.32), and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ from Proposition 3.3.2. Hence, we get $\|\mathbf{r}_3\|_2 = o_p(n^{-1/2} N)$ and $|\boldsymbol{\zeta}^{\mathrm{T}} \mathbf{r}_3| = o_p(n^{-1/2} N^{1/2} \|\boldsymbol{\zeta}\|_2)$ by the triangle inequality and the Cauchy-Schwarz inequality. Then since $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{\Omega}_{\boldsymbol{\beta}}\, n^{-1} \sum_{i=1}^{n} \{ \mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) - \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-} \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) \} + \mathbf{r}_2$ and $\|\mathbf{r}_2\|_2 = o_p(n^{-1/2})$ by Proposition 3.3.2, we have $\{ (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}, (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)^{\mathrm{T}} \}^{\mathrm{T}} = \boldsymbol{\Omega}^{-} n^{-1} \sum_{i=1}^{n} \mathbf{S}(y_i, \mathbf{x}_i) + (\mathbf{r}_2^{\mathrm{T}}, \mathbf{r}_3^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{r}_2, \mathbf{r}_3$ satisfy the properties stated in the lemma. □

B.7 Proof of Theorem 3.3.1

We can write $\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0$ as

$$\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0 = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\, n^{-1} \sum_{i=1}^{n} \mathrm{var}^*(Y | \mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}})$$

$$+\boldsymbol{\beta}_0\, n^{-1}\sum_{i=1}^{n}\Big\{\mathrm{var}^*(Y|\mathbf{x}_i,\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}})-\mathrm{var}(Y|\mathbf{x}_i)\Big\}$$
$$+\boldsymbol{\beta}_0\left[n^{-1}\sum_{i=1}^{n}\mathrm{var}(Y|\mathbf{x}_i)-E\{\mathrm{var}(Y|\mathbf{X})\}\right]. \tag{B.34}$$

Noting that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by Proposition 3.3.2 and $\|\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$ by Lemmas B.6.1, we have $\|f^*_{Y|\mathbf{X}}(\cdot,\mathbf{x},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}})-f_{Y|\mathbf{X}}(\cdot,\mathbf{x})\|_\infty = o_p(1)$ uniformly in $\mathbf{x}$ by Conditions (C1) and (C5), then it is easy to see that

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\mathrm{var}^*(Y|\mathbf{x}_i,\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}}) &= n^{-1}\sum_{i=1}^{n}\{\mathrm{var}^*(Y|\mathbf{x}_i,\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}})-\mathrm{var}(Y|\mathbf{x}_i)\} \\ &\quad+\left[n^{-1}\sum_{i=1}^{n}\mathrm{var}(Y|\mathbf{x}_i)-E\{\mathrm{var}(Y|\mathbf{X})\}\right]+E\{\mathrm{var}(Y|\mathbf{X})\} \\ &= E\{\mathrm{var}(Y|\mathbf{X})\}+o_p(1). \end{aligned} \tag{B.35}$$

Next, we have

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}\Big\{\mathrm{var}^*(Y|\mathbf{x}_i,\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}})-\mathrm{var}(Y|\mathbf{x}_i)\Big\} \qquad\qquad (B.36)\\ =\ & n^{-1}\sum_{i=1}^{n}\Big\{\mathrm{var}^*(Y|\mathbf{x}_i,\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}})-\mathrm{var}^*(Y|\mathbf{x}_i,\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)+\mathrm{var}^*(Y|\mathbf{x}_i,\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)-\mathrm{var}(Y|\mathbf{x}_i)\Big\} \\ =\ & n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial\mathrm{var}^*(Y|\mathbf{x}_i,\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\frac{\partial\mathrm{var}^*(Y|\mathbf{x}_i,\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial\boldsymbol{\zeta}^{\mathrm{T}}}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0)\right\}+O(N^{-q}), \end{aligned}$$

where $(\boldsymbol{\beta}^{*\mathrm{T}},\boldsymbol{\zeta}^{*\mathrm{T}})^{\mathrm{T}}$ is a point on the line connecting $(\widehat{\boldsymbol{\beta}}^{\mathrm{T}},\widehat{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}}$ and $(\boldsymbol{\beta}_0^{\mathrm{T}},\boldsymbol{\zeta}_0^{\mathrm{T}})^{\mathrm{T}}$. The last equality holds by Condition (C5). Then since $\|f^*_{Y|\mathbf{X}}(\cdot,\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)-f_{Y|\mathbf{X}}(\cdot,\mathbf{x})\|_\infty = o_p(1)$ uniformly in $\mathbf{x}$ by Conditions (C1) and (C5), it is easy to check that

$$\begin{aligned} & \left\|n^{-1}\sum_{i=1}^{n}\frac{\partial\mathrm{var}^*(Y|\mathbf{x}_i,\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial\boldsymbol{\beta}}-E[\{Y-E(Y|\mathbf{X})\}^3\mathbf{X}]\right\|_2 \\ =\ & \left\|n^{-1}\sum_{i=1}^{n}E^*[\{Y-E^*(Y|\mathbf{x}_i,\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)\}^3|\mathbf{x}_i,\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*]\mathbf{x}_i-E[\{Y-E(Y|\mathbf{X})\}^3\mathbf{X}]\right\|_2 \\ =\ & o_p(1). \end{aligned} \tag{B.37}$$

Furthermore, we have

$$\frac{\partial \text{var}^*(Y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} = E^*(\{Y - E^*(Y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta})\}^2[\mathbf{B}_+(Y) - E^*\{\mathbf{B}_+(Y)|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta}\}]|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta}) \quad \text{(B.38)}$$

Note that

$$\begin{aligned}
& \|E^*[\{Y - E^*(Y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta})\}^2\mathbf{B}_+(Y)|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta}]\|_2 \\
= & \left\{\sum_{k=1}^{m+1}\left(E^*[\{Y - E^*(Y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta})\}^2\mathbf{B}_+(Y)|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta}]^{\mathrm{T}}\mathbf{e}_k\right)^2\right\}^{1/2} \\
\leqslant & \sqrt{m+1}\max_{k=1,\ldots,m+1}\|\mathbf{B}_+(\cdot)^{\mathrm{T}}\mathbf{e}_k\|_1 \sup_{y\in[0,1],\mathbf{x}\in\mathcal{X}}\{y - E^*(Y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta})\}^2 f^*_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta}) \\
= & O(N^{-1/2})
\end{aligned}$$

by Conditions (C1), (C3), and Lemma B.3.4. Similar argument will show

$$\|E^*[\{Y - E^*(Y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta})\}^2 E^*\{\mathbf{B}_+(Y)|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta}\}|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta}]\|_2 = O(N^{-1/2}),$$

and

$$\|E(\{Y - E(Y|\mathbf{x})\}^2[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{x}\}]|\mathbf{x})\|_2 = O(N^{-1/2}). \quad \text{(B.39)}$$

Then (B.38), Condition (C1), and $\|f^*_{Y|\mathbf{X}}(\cdot,\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(\cdot,\mathbf{x})\|_\infty = o_p(1)$ lead to

$$\sup_{\mathbf{x}\in\mathcal{X}}\left\|\frac{\partial \text{var}^*(Y|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial \boldsymbol{\zeta}} - E(\{Y - E(Y|\mathbf{x})\}^2[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{x}\}]|\mathbf{x})\right\|_2 = o_p(N^{-1/2}).$$

Thus, we have

$$\begin{aligned}
& \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \text{var}^*(Y|\mathbf{x}_i,\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial \boldsymbol{\zeta}} - E(\{Y - E(Y|\mathbf{X})\}^2[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{X}\}])\right\|_2 \\
= & o_p(N^{-1/2}). \quad \text{(B.40)}
\end{aligned}$$

Using the fact that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by Proposition 3.3.2 and $\boldsymbol{\zeta}^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) = O_p(n^{-1/2}N^{1/2}\|\boldsymbol{\zeta}\|_2)$ by Lemma B.6.1, we combine the above with (B.36) and (B.37) to get

$$n^{-1}\sum_{i=1}^{n}\left\{\text{var}^*(Y|\mathbf{x}_i,\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}}) - \text{var}(Y|\mathbf{x}_i)\right\}$$

$$
\begin{aligned}
&= E[\{Y - E(Y|\mathbf{X})\}^3\mathbf{X}]^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\quad + E(\{Y - E(Y|\mathbf{X})\}^2[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{X}\}])^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) + o_p(n^{-1/2}).
\end{aligned}
$$

Then, we can rewrite (B.34) using (B.35) as

$$
\begin{aligned}
\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0 &= (E\{\mathrm{var}(Y|\mathbf{X})\}\mathbf{I} + \boldsymbol{\beta}_0 E[\{Y - E(Y|\mathbf{X})\}^3\mathbf{X}^{\mathrm{T}}])(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\quad + \boldsymbol{\beta}_0 E(\{Y - E(Y|\mathbf{X})\}^2[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{X}\}])^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \\
&\quad + \boldsymbol{\beta}_0 \left[n^{-1}\sum_{i=1}^{n} \mathrm{var}(Y|\mathbf{x}_i) - E\{\mathrm{var}(Y|\mathbf{X})\} \right] + \mathbf{r}, \qquad \text{(B.41)}
\end{aligned}
$$

where $\|\mathbf{r}\|_2 = o_p(n^{-1/2})$. Hence we get $\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0\|_2 = O_p(n^{-1/2})$ by (B.39).

Now we derive the asymptotic distribution of $\widehat{\boldsymbol{\xi}}$. By Lemma B.6.1, (B.41) equals to

$$
\begin{aligned}
\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0 &= \mathbf{A}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\boldsymbol{\Omega}^{-} n^{-1}\sum_{i=1}^{n} \mathbf{S}(y_i, \mathbf{x}_i) + \boldsymbol{\beta}_0 \left[n^{-1}\sum_{i=1}^{n} \mathrm{var}(Y|\mathbf{x}_i) - E\{\mathrm{var}(Y|\mathbf{X})\} \right] \\
&\quad + \mathbf{A}_1\mathbf{r}_2 + \mathbf{A}_2\mathbf{T}^{\mathrm{T}}\mathbf{r}_3 + \mathbf{r}.
\end{aligned}
$$

We obviously have

$$
\begin{aligned}
\mathrm{cov}[\mathbf{X}\{Y - E(Y|\mathbf{X})\}, \mathrm{var}(Y|\mathbf{X})] &= \mathbf{0}_p, \\
\mathrm{cov}[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{X}\}, \mathrm{var}(Y|\mathbf{X})] &= \mathbf{0}_{m+1}.
\end{aligned}
$$

In addition, recall that $\|\mathbf{r}_2\|_2 = o_p(n^{-1/2})$ and $|\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{r}_3| = o_p(n^{-1/2}N^{1/2}\|\boldsymbol{\zeta}\|_2)$ from Lemma B.6.1, and $\|\mathbf{A}_2\mathbf{T}^{\mathrm{T}}\|_2 = O(N^{-1/2})$ by (B.39), then the remainders satisfy

$$
\|\mathbf{A}_1\mathbf{r}_2 + \mathbf{A}_2\mathbf{T}^{\mathrm{T}}\mathbf{r}_3 + \mathbf{r}\|_2 \leqslant \|\mathbf{A}_1\|_2\|\mathbf{r}_2\|_2 + \|\mathbf{A}_2\mathbf{T}^{\mathrm{T}}\mathbf{r}_3\|_2 + \|\mathbf{r}\|_2 = o_p(n^{-1/2}).
$$

Hence, because

$$
\begin{aligned}
&\mathbf{A}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\boldsymbol{\Omega}^{-}\mathrm{diag}(\mathbf{I}_p, \mathbf{T})\mathbf{A}^{\mathrm{T}} + \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\mathrm{var}\{\mathrm{var}(Y|\mathbf{X})\} \\
= \;& \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\mathrm{var}\{\mathrm{var}(Y|\mathbf{X})\} = \boldsymbol{\Sigma}_{\boldsymbol{\xi}}
\end{aligned}
$$

by Lemma B.3.3, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0)$ converges to the normal distribution with mean $\mathbf{0}$ and variance $\mathbf{I}$. □

B.8 Proof of Theorem 3.3.2

For notational brevity, we denote $\nu \equiv \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, $q^*(\nu, \boldsymbol{\zeta})$ be such that

$$\int_0^{q^*(\nu,\boldsymbol{\zeta})} \exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt = \tau \int_0^1 \exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt,$$

and

$$\begin{aligned} q^{*\prime}(\nu,\boldsymbol{\zeta}) &\equiv \frac{\partial q^*(\nu,\boldsymbol{\zeta})}{\partial \nu} \\ &= \left.\frac{\tau\int_0^1 t\exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt - \int_0^y t\exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt}{\exp\{y\nu + \mathbf{B}_+(y)^{\mathrm{T}}\boldsymbol{\zeta}\}}\right|_{y=q^*(\nu,\boldsymbol{\zeta})} \\ &= \frac{E^*([\tau - I\{Y < q^*(\nu,\boldsymbol{\zeta})\}]Y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta})}{f^*_{Y|\mathbf{X}}\{q^*(\nu,\boldsymbol{\zeta}),\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta}\}}. \end{aligned} \tag{B.42}$$

Let $\boldsymbol{\beta}^*$ and $\boldsymbol{\zeta}^*$ be such that $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 = o_p(1)$ and $\|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}_0\|_2 = o_p(1)$ respectively. First note that

$$\sup_{\mathbf{x}\in\mathcal{X}} |q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) - q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)| = o_p(1). \tag{B.43}$$

This is because

$$\begin{aligned} 0 &= \int_0^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)} f^*_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)dy - \int_0^{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)} f_{Y|\mathbf{X}}(y|\mathbf{x})dy \\ &= \int_0^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)} \{f^*_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(y|\mathbf{x})\}dy + \int_{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)}^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)} f_{Y|\mathbf{X}}(y|\mathbf{x})dy, \end{aligned}$$

which implies, by Condition (C1) and $\|f^*_{Y|\mathbf{X}}(\cdot,\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(\cdot,\mathbf{x})\|_\infty = o_p(1)$ uniformly in $\mathbf{x}$ by Conditions (C1) and (C5), that

$$\begin{aligned} c_f|q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)| &\leqslant \left|\int_0^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)} \{f^*_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(y|\mathbf{x})\}dy\right| \\ &= o_p(1) \end{aligned}$$

uniformly in $\mathbf{x}$, where $c_f = \inf_{y\in[0,1],\mathbf{x}\in\mathcal{X}} f_{Y|\mathbf{X}}(y|\mathbf{x})$. Similarly, we also have

$$\sup_{\mathbf{x}\in\mathcal{X}} |q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)| = O(N^{-q}) \tag{B.44}$$

because

$$\begin{aligned} 0 &= \int_0^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)} f^*_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)dy - \int_0^{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)} f_{Y|\mathbf{X}}(y,\mathbf{x})dy \\ &= \int_0^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)} \{f^*_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0) - f_{Y|\mathbf{X}}(y,\mathbf{x})\}dy + \int_{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)}^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)} f_{Y|\mathbf{X}}(y,\mathbf{x})dy, \end{aligned}$$

which, together with $\|f^*_{Y|\mathbf{X}}(\cdot,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0) - f_{Y|\mathbf{X}}(\cdot,\mathbf{x})\|_\infty = O(N^{-q})$ uniformly in $\mathbf{x}$ by Conditions (C1) and (C5), implies

$$\begin{aligned} c_f|q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0) - q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)| &\leqslant \left|\int_0^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)} \{f^*_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0) - f_{Y|\mathbf{X}}(y,\mathbf{x})\}dy\right| \\ &= O(N^{-q}) \end{aligned}$$

uniformly in $\mathbf{x}$ by Condition (C1). Condition (C1) further leads to

$$\begin{aligned} & |q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)| \\ = & \left|\frac{E^*([\tau - I\{Y < q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)\}]Y|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*),\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\}} - \frac{E([\tau - I\{Y < q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0),\mathbf{x}\}}\right| \\ \leqslant & \left|\frac{1}{f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*),\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\}} - \frac{1}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0),\mathbf{x}\}}\right| \\ & \times|E^*([\tau - I\{Y < q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)\}]Y|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)| \\ & + \frac{1}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0),\mathbf{x}\}} \times (\tau\,|E^*(Y|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - E(Y|\mathbf{x})| \\ & + \left|E^*[I\{Y < q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)\}Y|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*] - E[I\{Y < q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}Y|\mathbf{x}]\right|\big) \\ = & o_p(1) \end{aligned} \tag{B.45}$$

$$= o_p(1) \tag{B.46}$$

uniformly in $\mathbf{x}$. The first term in (B.45) is $o_p(1)$ since using (B.43) and $\|f^*_{Y|\mathbf{X}}(\cdot,\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(\cdot,\mathbf{x})\|_\infty = o_p(1)$, we have

$$\begin{aligned} & \left|f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*),\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\} - f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}\right| \\ \leqslant & \left|f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*),\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\} - f_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)|\mathbf{x}\}\right| \\ & + \left|f_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)|\mathbf{x}\} - f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}\right| \\ = & o_p(1). \end{aligned} \tag{B.47}$$

The second term in (B.45) is $o_p(1)$ because $|E^*(Y|\mathbf{x}, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) - E(Y|\mathbf{x})| = o_p(1)$, and for an arbitrary $g(\cdot)$ such that $\|g(\cdot)\|_1 < \infty$,

$$\begin{aligned}
& \left|E^*\left[I\{Y < q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}g(Y)|\mathbf{x}, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*\right] - E\left[I\{Y < q(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\}g(Y)|\mathbf{x}\right]\right| \\
\leqslant & \left|\int_0^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)} g(y)\{f^*_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) - f_{Y|\mathbf{X}}(y|\mathbf{x})\}dy\right| + \left|\int_{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)}^{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)} g(y)f_{Y|\mathbf{X}}(y|\mathbf{x})dy\right| \\
\leqslant & o_p(1)\|g(\cdot)\|_1 + C_f|q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) - q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\|g(\cdot)\|_1 \\
= & o_p(1)\|g(\cdot)\|_1. \qquad \text{(B.48)}
\end{aligned}$$

Similarly, we also have

$$\sup_{\mathbf{x}\in\mathcal{X}} |q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) - q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)| = O(N^{-q}) \qquad \text{(B.49)}$$

by Conditions (C1) and (C5).

In addition, we can show that

$$\begin{aligned}
& \frac{\partial q^{*\prime}(\nu, \boldsymbol{\zeta})}{\partial \boldsymbol{\beta}} \\
= & \frac{\tau\int_0^1 \mathbf{x}t^2\exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt}{\exp[q^*(\nu, \boldsymbol{\zeta})\nu + \mathbf{B}_+\{q^*(\nu, \boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}]} \\
& - \frac{\int_0^{q^*(\nu,\boldsymbol{\zeta})} \mathbf{x}t^2\exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt + q^*(\nu, \boldsymbol{\zeta})\exp[q^*(\nu, \boldsymbol{\zeta})\nu + \mathbf{B}_+\{q^*(\nu, \boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}]\mathbf{x}q^{*\prime}(\nu, \boldsymbol{\zeta})}{\exp[q^*(\nu, \boldsymbol{\zeta})\nu + \mathbf{B}_+\{q^*(\nu, \boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}]} \\
& - \frac{\tau\int_0^1 t\exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt - \int_0^{q^*(\nu,\boldsymbol{\zeta})} t\exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt}{\exp[q^*(\nu, \boldsymbol{\zeta})\nu + \mathbf{B}_+\{q^*(\nu, \boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}]} \\
& \times \left[\mathbf{x}q^{*\prime}(\nu, \boldsymbol{\zeta})\nu + q^*(\nu, \boldsymbol{\zeta})\mathbf{x} + \mathbf{B}'_+\{q^*(\nu, \boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}\mathbf{x}q^{*\prime}(\nu, \boldsymbol{\zeta})\right] \\
= & \mathbf{x}\left\{\frac{E^*([\tau - I\{Y \leqslant q^*(\nu, \boldsymbol{\zeta})\}]Y^2|\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta})}{f^*_{Y|\mathbf{X}}\{q^*(\nu, \boldsymbol{\zeta}), \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}\}}\right. \\
& \left. -2q^{*\prime}(\nu, \boldsymbol{\zeta})q^*(\nu, \boldsymbol{\zeta}) - \{q^{*\prime}(\nu, \boldsymbol{\zeta})\}^2[\nu + \mathbf{B}'_+\{q^*(\nu, \boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}]\right\}, \qquad \text{(B.50)} \\
& \frac{\partial q^*(\nu, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} \\
= & \left.\frac{\tau\int_0^1 \mathbf{B}_+(t)\exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt - \int_0^y \mathbf{B}_+(t)\exp\{t\nu + \mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt}{\exp\{y\nu + \mathbf{B}_+(y)^{\mathrm{T}}\boldsymbol{\zeta}\}}\right|_{y=q^*(\nu,\boldsymbol{\zeta})} \\
= & \frac{E^*([\tau - I\{Y \leqslant q^*(\nu, \boldsymbol{\zeta})\}]\mathbf{B}_+(Y)|\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta})}{f^*_{Y|\mathbf{X}}\{q^*(\nu, \boldsymbol{\zeta}), \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\zeta}\}}, \qquad \text{(B.51)}
\end{aligned}$$

and

$$
\begin{aligned}
&\frac{\partial q^{*\prime}(\nu,\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} \\
= \;& \frac{\tau\int_0^1 t\mathbf{B}_+(t)\exp\{t\nu+\mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt}{\exp\{q^*(\nu,\boldsymbol{\zeta})\nu+\mathbf{B}_+(y)^{\mathrm{T}}\boldsymbol{\zeta}\}} \\
& -\frac{\int_0^{q^*(\nu,\boldsymbol{\zeta})} t\mathbf{B}_+(t)\exp\{t\nu+\mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt + q^*(\nu,\boldsymbol{\zeta})\exp[q^*(\nu,\boldsymbol{\zeta})\nu+\mathbf{B}_+\{q^*(\nu,\boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}]\frac{\partial q^*(\nu,\boldsymbol{\zeta})}{\partial\boldsymbol{\zeta}}}{\exp[q^*(\nu,\boldsymbol{\zeta})\nu+\mathbf{B}_+\{q^*(\nu,\boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}]} \\
& -\frac{\tau\int_0^1 t\exp\{t\nu+\mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt - \int_0^{q^*(\nu,\boldsymbol{\zeta})} t\exp\{t\nu+\mathbf{B}_+(t)^{\mathrm{T}}\boldsymbol{\zeta}\}dt}{\exp[q^*(\nu,\boldsymbol{\zeta})\nu+\mathbf{B}_+\{q^*(\nu,\boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}]} \\
& \times\left[\frac{\partial q^*(\nu,\boldsymbol{\zeta})}{\partial\boldsymbol{\zeta}}\nu+\mathbf{B}'_+\{q^*(\nu,\boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}\frac{\partial q^*(\nu,\boldsymbol{\zeta})}{\partial\boldsymbol{\zeta}}+\mathbf{B}_+\{q^*(\nu,\boldsymbol{\zeta})\}\right] \\
= \;& \frac{E^*([\tau-I\{Y\leqslant q^*(\nu,\boldsymbol{\zeta})\}]Y\mathbf{B}_+(Y)|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta})}{f^*_{Y|\mathbf{X}}\{q^*(\nu,\boldsymbol{\zeta}),\mathbf{x},\boldsymbol{\beta},\boldsymbol{\zeta}\}} \qquad \text{(B.52)} \\
& -q^{*\prime}(\nu,\boldsymbol{\zeta})\mathbf{B}_+\{q^*(\nu,\boldsymbol{\zeta})\}-\frac{\partial q^*(\nu,\boldsymbol{\zeta})}{\partial\boldsymbol{\zeta}}[q^*(\nu,\boldsymbol{\zeta})+q^{*\prime}(\nu,\boldsymbol{\zeta})\nu+q^{*\prime}(\nu,\boldsymbol{\zeta})\mathbf{B}'_+\{q^*(\nu,\boldsymbol{\zeta})\}^{\mathrm{T}}\boldsymbol{\zeta}].
\end{aligned}
$$

Now using (B.49), we write $\widehat{\boldsymbol{\eta}_\tau}-\boldsymbol{\eta}_{\tau 0}$ as

$$
\begin{aligned}
&\widehat{\boldsymbol{\eta}_\tau}-\boldsymbol{\eta}_{\tau 0} \\
= \;& (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)n^{-1}\sum_{i=1}^n q^{*\prime}\{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})\}+\boldsymbol{\beta}_0 n^{-1}\sum_{i=1}^n\left[q^{*\prime}\{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})\}-q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\right] \\
& +\boldsymbol{\beta}_0 n^{-1}\sum_{i=1}^n\left\{q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)-q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\right\}+\boldsymbol{\beta}_0\left[n^{-1}\sum_{i=1}^n q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)-E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right] \\
= \;& \left[n^{-1}\sum_{i=1}^n q^{*\prime}\{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})\}\mathbf{I}+\boldsymbol{\beta}_0 n^{-1}\sum_{i=1}^n\frac{\partial q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right](\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\
& +\boldsymbol{\beta}_0 n^{-1}\sum_{i=1}^n\frac{\partial q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial\boldsymbol{\zeta}^{\mathrm{T}}}\{\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})-\boldsymbol{\zeta}_0\} \\
& +\boldsymbol{\beta}_0\left[n^{-1}\sum_{i=1}^n q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)-E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]+O(N^{-q}), \qquad \text{(B.53)}
\end{aligned}
$$

where $(\boldsymbol{\beta}^{*\mathrm{T}},\boldsymbol{\zeta}^{*\mathrm{T}})^{\mathrm{T}}$ is a point on the line connecting $\{\widehat{\boldsymbol{\beta}}^{\mathrm{T}},\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\}^{\mathrm{T}}$ and $(\boldsymbol{\beta}_0^{\mathrm{T}},\boldsymbol{\zeta}_0^{\mathrm{T}})^{\mathrm{T}}$. To treat the first term in (B.53), we first obtain

$$
\left|n^{-1}\sum_{i=1}^n q^{*\prime}\{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})\}-E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right|
$$

$$
\begin{aligned}
&\leqslant n^{-1}\sum_{i=1}^{n}\left|q^{*\prime}\{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})\}-q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\right|+\left|n^{-1}\sum_{i=1}^{n}q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)-E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right| \\
&= o_p(1) \qquad \text{(B.54)}
\end{aligned}
$$

by (B.46). Now we will show

$$
\begin{aligned}
&\left\|n^{-1}\sum_{i=1}^{n}\frac{\partial q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial\boldsymbol{\beta}}-E\left[\mathbf{X}\left\{\frac{E([\tau-I\{Y\leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y^2|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}}\right.\right.\right. \\
&\left.\left.\left.-2q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)-\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2[\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0+c'\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\}\right]\right\|_2 \\
&= o_p(1). \qquad \text{(B.55)}
\end{aligned}
$$

We prove (B.55) through proving

$$
\begin{aligned}
&\left\|\frac{\partial q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial\boldsymbol{\beta}}-\mathbf{x}\left\{\frac{E([\tau-I\{Y\leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y^2|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}}\right.\right. \\
&\left.\left.-2q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)-\{q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2[\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0+c'\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\right\}\right\|_2 \\
&= o_p(1) \qquad \text{(B.56)}
\end{aligned}
$$

uniformly in $\mathbf{x}$ under Condition (C1), where $\partial q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)/\partial\boldsymbol{\beta}$ is given in (B.50). To prove (B.56), first note that

$$
\begin{aligned}
&\left|\frac{E^*([\tau-I\{Y\leqslant q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)\}]Y^2|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*),\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\}}-\frac{E([\tau-I\{Y\leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y^2|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}}\right| \\
&= o_p(1), \qquad \text{(B.57)}
\end{aligned}
$$

because $|E^*(Y^2|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)-E(Y^2|\mathbf{x})|=o_p(1)$ by $\|f^*_{Y|\mathbf{X}}(\cdot,\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)-f_{Y|\mathbf{X}}(\cdot,\mathbf{x})\|_\infty = o_p(1)$,

$$
|E^*[I\{Y\leqslant q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)\}Y^2|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*]-E[I\{Y\leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}Y^2|\mathbf{x}]|=o_p(1)
$$

by (B.48), and

$$
\left|\frac{1}{f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*),\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\}}-\frac{1}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}}\right|=o_p(1)
$$

by (B.47). Furthermore, we have

$$|q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) - q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)| = o_p(1) \tag{B.58}$$

by (B.43) and (B.46). We additionally get

$$\begin{aligned} & |\mathbf{B}'_+\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^{\mathrm{T}}\boldsymbol{\zeta}^* - c'\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}| \\ \leqslant\ & \|\mathbf{B}'_+\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}\|_2\|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}_0\|_2 + |\mathbf{B}'_+\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^{\mathrm{T}}\boldsymbol{\zeta}_0 - c'\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}| \\ & +|c'\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\} - c'\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}| \\ =\ & o_p(1), \end{aligned} \tag{B.59}$$

where, to bound the first term, we used $\|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$ by Lemma B.6.1, for the second term, we used $|\mathbf{B}'_+\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^{\mathrm{T}}\boldsymbol{\zeta}_0 - c'\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}| = O(N^{-q/2})$ under Conditions (C1), (C2), (C5), and for the third term, we noted that $c'(\cdot)$ is continuous under Condition (C1) and used (B.43). The results in (B.59) and (B.46), together with the fact $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ directly lead to

$$\begin{aligned} & |q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^2[\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \mathbf{B}'_+\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^{\mathrm{T}}\boldsymbol{\zeta}^*] \\ & -\{q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2[\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + c'\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]| \\ =\ & o_p(1). \end{aligned} \tag{B.60}$$

Combining the results in (B.57), (B.58), and (B.60), and taking into account the form in (B.50) lead to (B.56), and subsequently (B.55). We now combine (B.54) and (B.55) to get

$$\left\| n^{-1}\sum_{i=1}^n q^{*\prime}\{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})\}\mathbf{I} + \boldsymbol{\beta}_0 n^{-1}\sum_{i=1}^n \frac{\partial q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} - \mathbf{C}_1 \right\|_2 = o_p(1). \tag{B.61}$$

Next, we handle the second term in (B.53). By (B.51), we have

$$\begin{aligned} & \left\| \frac{\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)}{\partial \boldsymbol{\zeta}} - \frac{E([\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+(Y)|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}} \right\|_\infty \\ =\ & \sup_{k=1,\ldots,m+1} \left| \left\{ \frac{\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)}{\partial \boldsymbol{\zeta}} - \frac{E([\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+(Y)|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}} \right\}^{\mathrm{T}} \mathbf{e}_k \right| \\ \leqslant\ & \sup_{k=1,\ldots,m+1} \left| \frac{1}{f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*), \mathbf{x}, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*\}} \right. \end{aligned}$$

$$
\begin{aligned}
&\times\big\{\tau[E^*\{\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\}-E\{\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x}\}]\\
&-(E^*[I\{Y\leqslant q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)\}\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*]\\
&-E[I\{Y\leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x}])\big\}\Big|\\
&+\sup_{k=1,\ldots,m+1}\left|\left[\frac{1}{f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*),\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\}}-\frac{1}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}}\right]\right.\\
&\left.\times E([\tau-I\{Y\leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x})\right|\\
=\;& o_p(N^{-1}).
\end{aligned}
$$

The last equality holds because $\sup_{k=1,\ldots,m+1}\left|E^*\{\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\}-E\{\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x}\}\right|=o_p(N^{-1})$ by Lemma B.3.4 and $\|f^*_{Y|\mathbf{X}}(\cdot,\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)-f_{Y|\mathbf{X}}(\cdot,\mathbf{x})\|_\infty=o_p(1)$,

$$
\begin{aligned}
&\sup_{k=1,\ldots,m+1}\left|E^*[I\{Y\leqslant q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)\}\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*]\right.\\
&\left.-E[I\{Y\leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x}]\right|\\
=\;& o_p(N^{-1})
\end{aligned}
$$

by (B.48) and Lemma B.3.4,

$$
\left|\frac{1}{f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*),\mathbf{x},\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*\}}-\frac{1}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}}\right|=o_p(1)
$$

by (B.47), and

$$
\begin{aligned}
&\sup_{k=1,\ldots,m+1}|E([\tau-I\{Y\leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mathbf{x})|\\
\leqslant\;&\sup_{k=1,\ldots,m+1}E\{|\mathbf{B}_+(Y)^{\mathrm{T}}\mathbf{e}_k|\mid\mathbf{x}\}=O(N^{-1})
\end{aligned}
\tag{B.62}
$$

by Lemma B.3.4 and Condition (C1). Since $\|\boldsymbol{\zeta}\|_2=O(N^{1/2}\|\boldsymbol{\zeta}\|_\infty)$ for any $\boldsymbol{\zeta}\in\mathbb{R}^{m+1}$ by Condition (C3), we get

$$
\left\|\frac{\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)}{\partial\boldsymbol{\zeta}}-\frac{E([\tau-I\{Y\leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+(Y)|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}}\right\|_2=o_p(N^{-1/2})
\tag{B.63}
$$

uniformly in $\mathbf{x}$ under Condition (C1). Similarly, we have

$$\|E([\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y\mathbf{B}_+(Y)|\mathbf{x})\|_2 = O(N^{-1/2}) \tag{B.64}$$

by (B.62), and

$$\begin{aligned} & \left\| \frac{E^*([\tau - I\{Y \leqslant q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}]Y\mathbf{B}_+(Y)|\mathbf{x}, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)}{f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*), \mathbf{x}, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*\}} \right. \\ & \left. - \frac{E([\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y\mathbf{B}_+(Y)|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}} \right\|_2 \\ = \; & o_p(N^{-1/2}) \end{aligned} \tag{B.65}$$

uniformly in $\mathbf{x}$. (B.65) can be shown similarly to showing (B.63) through replacing $\mathbf{B}_+(y)$ with $y\mathbf{B}_+(y)$ and using that the response $y \in [0,1]$ from Condition (C1). On the other hand, we have

$$\begin{aligned} & \left\|E\left[q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right\|_\infty \\ = \; & \max_{k=1,\ldots,m+1} \left| \int_{\{\nu:\nu=\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \mathbf{x}\in\mathcal{X}\}} \mathbf{B}_+\{q(\nu)\}^{\mathrm{T}}\mathbf{e}_k \left[\int_{\{\mathbf{x}:\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0=\nu\}} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \right] q'(\nu)d\nu \right| \\ = \; & \max_{k=1,\ldots,m+1} \left| \int_0^1 \mathbf{B}_+(t)^{\mathrm{T}}\mathbf{e}_k \left[\int_{\{\mathbf{x}\in\mathcal{X}:q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)=t\}} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \right] dt \right| \\ \leqslant \; & \max_{k=1,\ldots,m+1} \|\mathbf{B}_+(\cdot)^{\mathrm{T}}\mathbf{e}_k\|_1 \sup_{t\in[0,1]} \left[\int_{\{\mathbf{x}\in\mathcal{X}:q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)=t\}} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \right] \\ \leqslant \; & C_{\mathcal{X}} \max_{k=1,\ldots,m+1} \|\mathbf{B}_+(\cdot)^{\mathrm{T}}\mathbf{e}_k\|_1 \\ = \; & O(N^{-1}) \end{aligned} \tag{B.66}$$

by Lemma B.3.4. The last inequality holds for some constant $C_{\mathcal{X}} > 0$ since $f_{\mathbf{X}}(\mathbf{x})$ is bounded on $\mathcal{X}$ by Condition (C1), and $\{\mathbf{x} \in \mathcal{X} : q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = t\}$ is bounded for any t hence has finite volume because $\mathcal{X}$ is bounded by Condition (C1). Similarly, we can show

$$\left\|E\left[q^{*\prime}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\zeta})\mathbf{B}_+\{q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\zeta})\}\right]\right\|_\infty = O(N^{-1}). \tag{B.67}$$

In addition, it is easy to see that $|q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\} - E[q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]| \leqslant C_1$, and

$$\mathrm{var}[q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}] \;\leqslant\; E([q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]^2)$$

$$
\begin{aligned}
&\leqslant C_2' E[q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) B_k^2\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}] \\
&= C_2' \int B_k^2(t) \left[\int_{\{\mathbf{x}: q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)=t\}} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \right] dt \\
&\leqslant C_2'' \|B_k(\cdot)\|_2^2 \\
&\leqslant C_2 N^{-1}
\end{aligned}
$$

for some constants $C_1, C_2, C_2', C_2'' > 0$ by Lemma B.3.4, where $q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = E([\tau - I\{Y < q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)/f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0), \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0\}$ is bounded because the response Y is bounded and $f_{Y|\mathbf{X}}(y, \mathbf{x})$ is positive by Condition (C1). Then by Bernstein's inequality,

$$
\begin{aligned}
&\mathrm{pr}\left(\left| n^{-1} \sum_{i=1}^n q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) B_k\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\} - E[q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) B_k\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}] \right| \geqslant \epsilon \right) \\
\leqslant\ & 2\exp\left(\frac{-n^2\epsilon^2/2}{C_1 n\epsilon/3 + C_2 n N^{-1}} \right)
\end{aligned}
$$

which can be arbitrarily small when $\epsilon = C_\epsilon n^{-1/2} N^{-1/2}$ for sufficiently large constant $C_\epsilon > 0$. Noting that the result holds for any k due to the property of the B-spline bases. Using Boole's inequality and $m \asymp N$ from Condition (C3), we further get

$$
\begin{aligned}
&\left\| n^{-1} \sum_{i=1}^n q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) \mathbf{B}_+\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\} - E[q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) \mathbf{B}_+\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}] \right\|_\infty \\
=\ & O_p\{n^{-1/2} N^{-1/2} (\log N)^{1/2}\} \qquad \text{(B.68)}
\end{aligned}
$$

by Condition (C3). Similarly, we can show

$$
\begin{aligned}
&\left\| n^{-1} \sum_{i=1}^n q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\zeta}) \mathbf{B}_+\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\zeta})\} - E[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\zeta}) \mathbf{B}_+\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\zeta})\}] \right\|_\infty \\
=\ & O_p\{n^{-1/2} N^{-1/2} (\log N)^{1/2}\}. \qquad \text{(B.69)}
\end{aligned}
$$

Now, we have

$$
\begin{aligned}
&\left| n^{-1} \sum_{i=1}^n \left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\} - q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) B_k\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\} \right] \right| \\
\leqslant\ & \left| n^{-1} \sum_{i=1}^n \left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\} - q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\} \right] \right|
\end{aligned}
$$

$$
\begin{aligned}
&+\left|n^{-1}\sum_{i=1}^{n}\left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\}-q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right| \\
= &\left|\widehat{\mathbf{d}}_{k1}^{\mathrm{T}}(\boldsymbol{\beta}^*-\boldsymbol{\beta}_0)+\widehat{\mathbf{d}}_{k2}^{\mathrm{T}}(\boldsymbol{\zeta}^*-\boldsymbol{\zeta}_0)\right| \\
&+\left|n^{-1}\sum_{i=1}^{n}\left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\}-q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right| \\
\leqslant &\left|\widehat{\mathbf{d}}_{k1}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right|+\left|\widehat{\mathbf{d}}_{k2}^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0)\right| \\
&+\left|n^{-1}\sum_{i=1}^{n}\left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\}-q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right|, \\
\leqslant &\left|\mathbf{d}_{k1}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right|+\left|(\widehat{\mathbf{d}}_{k1}-\mathbf{d}_{k1})^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right| \\
&+\left|\mathbf{d}_{k2}^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0)\right|+\left|(\widehat{\mathbf{d}}_{k2}-\mathbf{d}_{k2})^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0)\right| \\
&+\left|n^{-1}\sum_{i=1}^{n}\left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\}-q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right|, \qquad \text{(B.70)}
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{\mathbf{d}}_{k1} &\equiv n^{-1}\sum_{i=1}^{n}\left[\frac{\partial q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{**},\boldsymbol{\zeta}^{**})}{\partial\boldsymbol{\beta}}B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{**},\boldsymbol{\zeta}^{**})\}\right. \\
&\quad \left.+q^{*\prime 2}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{**},\boldsymbol{\zeta}^{**})B_k'\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{**},\boldsymbol{\zeta}^{**})\}\mathbf{x}_i\right], \\
\mathbf{d}_{k1} &\equiv E\left[\frac{\partial q^{*\prime}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)}{\partial\boldsymbol{\beta}}B_k\{q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\}+q^{*\prime 2}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)B_k'\{q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\}\mathbf{X}\right], \\
\widehat{\mathbf{d}}_{k2} &\equiv n^{-1}\sum_{i=1}^{n}\left[\frac{\partial q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{**},\boldsymbol{\zeta}^{**})}{\partial\boldsymbol{\zeta}}B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{**},\boldsymbol{\zeta}^{**})\}\right. \\
&\quad \left.+q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{**},\boldsymbol{\zeta}^{**})B_k'\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{**},\boldsymbol{\zeta}^{**})\}\frac{\partial q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{**},\boldsymbol{\zeta}^{**})}{\partial\boldsymbol{\zeta}}\right], \\
\mathbf{d}_{k2} &\equiv E\left[\frac{\partial q^{*\prime}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)}{\partial\boldsymbol{\zeta}}B_k\{q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\}\right. \\
&\quad \left.+q^{*\prime}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)B_k'\{q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)\}\frac{\partial q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)}{\partial\boldsymbol{\zeta}}\right].
\end{aligned}
$$

We first claim $\|\mathbf{d}_{k1}\|_2 = O(1)$. Note that $\|\mathbf{x}\|_2 = O(1)$ and $Y \in [0,1]$ by Condition (C1), $q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\zeta}) \in [0,1]$, and $q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)$ defined in (B.42) is bounded because $Y \in [0,1]$ by Condition (C1) and $f^*_{Y|\mathbf{X}}$ is strictly positive. Then we can show from (B.50) that $\|\partial q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0,\boldsymbol{\zeta}_0)/\partial\boldsymbol{\beta}\|_2 = O(1)$ because $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$ is bounded by Condition (C1) and $\mathbf{B}_+'(\cdot)^{\mathrm{T}}\boldsymbol{\zeta}_0$

is bounded by Lemma B.3.4 and $\|\boldsymbol{\zeta}_0\|_\infty \leqslant C_{\boldsymbol{\zeta}}$. Hence, using $B_k(\cdot) \leqslant 1$, we get

$$\left\| \frac{\partial q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)}{\partial \boldsymbol{\beta}} B_k\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\} \right\|_2 = O(1). \tag{B.71}$$

Then we analyze the second term in $\mathbf{d}_{k1}$. Note that $B_k'(\cdot) = q\{B_{k,q-1}(\cdot)/(t_{k+q} - t_k) - B_{k+1,q-1}(\cdot)/(t_{k+q+1} - t_{k+1})\}$ by the equation (8) on page 115 of De Boor (1978), where $B_{k,q-1}$ is the kth B-spline basis of order $q-1$. Then we get

$$\|B_k'(\cdot)\|_1 = \left\| q \left\{ \frac{B_{k,q-1}(\cdot)}{t_{k+q} - t_k} - \frac{B_{k+1,q-1}(\cdot)}{t_{k+q+1} - t_{k+1}} \right\} \right\|_1 = O(1) \tag{B.72}$$

using $\|B_{k,q-1}(\cdot)\|_1, \|B_{k+1,q-1}(\cdot)\|_1 \asymp N^{-1}$ by Lemma B.3.4, $(t_{k+q} - t_k), (t_{k+q+1} - t_{k+1}) \asymp N^{-1}$ by Condition (C4). This leads to

$$\begin{aligned}
& \left\| E\left[q^{*\prime 2}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) B_k'\{q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\}\mathbf{X} \right] \right\|_2 \\
= & \left\| \int_{\mathcal{X}} q^{*\prime 2}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) B_k'\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\}\mathbf{x} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \right\|_2 \\
= & \left\| \int_0^1 B_k'(t) \left[\int_{\{\mathbf{x}\in\mathcal{X}: q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) = t\}} q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\mathbf{x} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \right] dt \right\|_2 \\
\leqslant & \left\| q \left\{ \frac{B_{k,q-1}(\cdot)}{t_{k+q} - t_k} - \frac{B_{k+1,q-1}(\cdot)}{t_{k+q+1} - t_{k+1}} \right\} \right\|_1 \sup_{t\in[0,1]} \left[\int_{\{\mathbf{x}\in\mathcal{X}: q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) = t\}} \|q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\mathbf{x} f_{\mathbf{X}}(\mathbf{x})\|_2 d\mathbf{x} \right] \\
= & O(1).
\end{aligned} \tag{B.73}$$

The last equality holds because $\|q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\mathbf{x} f_{\mathbf{X}}(\mathbf{x})\|_2$ is bounded on $\mathcal{X}$ by Condition (C1), and $\{\mathbf{x} \in \mathcal{X} : q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) = t\}$ is bounded for any t hence has finite volume because $\mathcal{X}$ is bounded by Condition (C1). Therefore, by (B.71) and (B.73), we have $\|\mathbf{d}_{k1}\|_2 = O(1)$. Using Lemma B.6.1, we get $\mathbf{d}_{k1}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^n v_{1i} + o_p(n^{-1/2})$ where $v_{1i} \equiv \mathbf{d}_{k1}^{\mathrm{T}}\boldsymbol{\Omega}_{\boldsymbol{\beta}}\{\mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i)\}$. Note that v_{1i} satisfies $E(V_{1i}) = 0$, $\mathrm{var}(V_{1i}) = \mathbf{d}_{k1}^{\mathrm{T}}\boldsymbol{\Omega}_{\boldsymbol{\beta}}\mathbf{d}_{k1} \leqslant C_{11}$ and $|v_{1i}| \leqslant C_{12}N^{1/2}$ for some constants $C_{11}, C_{12} > 0$, because $\|\mathbf{S}_{\boldsymbol{\beta}}(y, \mathbf{x})\|_2 = O(1)$ from Condition (C1), $\|\boldsymbol{\Omega}_{21}\boldsymbol{\beta}\|_2 = O(N^{-1/2}\|\boldsymbol{\beta}\|_2)$ by (B.28), $\|\boldsymbol{\zeta}_1^{\mathrm{T}}\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\zeta}_2\|_2 = O(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2)$ by Lemma B.3.4, and $\|\mathbf{S}_{\boldsymbol{\zeta}}(y, \mathbf{x})\|_2 \leqslant \|\mathbf{B}_+(y)\|_2 + E\{\|\mathbf{B}_+(Y)\|_2 \mid \mathbf{x}\} \leqslant 2$. Then by Bernstein's inequality, for any positive δ, we can choose $\epsilon = C_1 n^{-1/2}$ where C_1 is a sufficiently large constant, so that

$$\mathrm{pr}\left\{ \left| \mathbf{d}_{k1}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right| \geqslant \epsilon \right\} = \mathrm{pr}\left\{ \left| n^{-1}\sum_{i=1}^n v_{1i} + o_p(n^{-1/2}) \right| \geqslant \epsilon \right\}$$

$$\begin{aligned} &\leqslant \quad \mathrm{pr}\left\{\left|n^{-1}\sum_{i=1}^{n} v_{1i}\right| \geqslant \epsilon/2\right\} \\ &\leqslant \quad 2\exp\left(\frac{-n^2\epsilon^2/8}{C_{11}n + C_{12}nN^{1/2}\epsilon/6}\right) < \delta \end{aligned} \tag{B.74}$$

when $n \to \infty$ by Condition (C3). Also, note that

$$\mathrm{pr}\left\{\left|(\widehat{\mathbf{d}}_{k1} - \mathbf{d}_{k1})^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right| \geqslant \epsilon\right\} \leqslant \mathrm{pr}\left\{\left|\mathbf{d}_{k1}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right| \geqslant \epsilon\right\} \tag{B.75}$$

because of the law of large numbers, $\|\boldsymbol{\beta}^{**} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ and $\|\boldsymbol{\zeta}^{**} - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$.

In addition, we will show $\|\mathbf{d}_{k2}\|_2 = O(N^{-1/2})$. To show this, noting that $\|\boldsymbol{\zeta}\|_2 = O(N^{1/2}\|\boldsymbol{\zeta}\|_\infty)$ for any $\boldsymbol{\zeta} \in \mathbb{R}^{m+1}$ by Condition (C3), first we have $\|\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)/\partial\boldsymbol{\zeta}^{\mathrm{T}}\|_2 = O(N^{-1/2})$ by (B.51) and (B.62). Also,

$$\left\| E\left[\frac{\partial q^{*\prime}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)}{\partial\boldsymbol{\zeta}^{\mathrm{T}}} B_k\{q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\}\right]\right\|_2 = O(N^{-1/2}) \tag{B.76}$$

based on (B.52). Specifically, we have $\|E^*([\tau - I\{Y \leqslant q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\}]Y\mathbf{B}_+(Y)|\mathbf{x}, \boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\|_2 = O(N^{-1/2})$ by (B.64) and (B.65), $\|E[q^{*\prime}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\mathbf{B}_+\{q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\}]\|_2 = O(N^{-1/2})$ by (B.67) and (B.69), and $\|\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)/\partial\boldsymbol{\zeta}^{\mathrm{T}}\|_2 = O(N^{-1/2})$. In addition, to handle the last term in (B.52), note that $q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\zeta}) \in [0, 1]$, $q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)$ defined in (B.42) is bounded because $Y \in [0, 1]$ by Condition (C1) and $f^*_{Y|\mathbf{X}}$ is positive, $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$ is bounded by Condition (C1), and $\mathbf{B}'_+(\cdot)^{\mathrm{T}}\boldsymbol{\zeta}_0$ is bounded because $\|\mathbf{B}'_+(\cdot)^{\mathrm{T}}\boldsymbol{\zeta}_0 - c'(\cdot)\|_\infty = O(N^{-q/2}) = o(1)$ by the Landau-Kolmogorov inequality and Conditions (C3), (C5), and $c'(\cdot)$ is bounded by Condition (C1). Then we have

$$q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) + q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\mathbf{B}'_+\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\}^{\mathrm{T}}\boldsymbol{\zeta}_0$$

bounded in probability. This shows (B.76). Also, the second term of $\mathbf{d}_{k2}$ satisfies

$$\begin{aligned} &\left\| E\left[q^{*\prime}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)B'_k\{q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\}\frac{\partial q^*(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)}{\partial\boldsymbol{\zeta}}\right]\right\|_2 \\ = \quad &\left\| \int_{\mathcal{X}} q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)B'_k\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\}\frac{\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)}{\partial\boldsymbol{\zeta}} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}\right\|_2 \\ = \quad &\left\| \int_0^1 B'_k(t)\left[\int_{\{\mathbf{x}\in\mathcal{X}: q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) = t\}} \frac{\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)}{\partial\boldsymbol{\zeta}} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}\right] dt\right\|_2 \end{aligned}$$

$$
\begin{aligned}
&\leqslant \left\| q \left\{ \frac{B_{k,q-1}(\cdot)}{t_{k+q} - t_k} - \frac{B_{k+1,q-1}(\cdot)}{t_{k+q+1} - t_{k+1}} \right\} \right\|_1 \\
&\quad \times \sup_{t \in [0,1]} \left[\int_{\{\mathbf{x} \in \mathcal{X} : q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) = t\}} \left\| \frac{\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)}{\partial \boldsymbol{\zeta}} f_{\mathbf{X}}(\mathbf{x}) \right\|_2 d\mathbf{x} \right] \\
&= O(N^{-1/2}). \qquad \text{(B.77)}
\end{aligned}
$$

The last equality holds by (B.72) and $\|\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)/\partial \boldsymbol{\zeta}^{\mathrm{T}}\|_2 = O(N^{-1/2})$ and $\{\mathbf{x} \in \mathcal{X} : q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0) = t\}$ is bounded for any t hence has finite volume because $\mathcal{X}$ is bounded by Condition (C1). Then from (B.76) and (B.77), we get $\|\mathbf{d}_{k2}\|_2 = O(N^{-1/2})$.

By Lemma B.6.1, we get $\mathbf{d}_{k2}^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) = n^{-1} \sum_{i=1}^n v_{2i} + o_p(n^{-1/2})$ where

$$
v_{2i} \equiv \mathbf{d}_{k2}^{\mathrm{T}} \{ -\boldsymbol{\Omega}_{22}^- \boldsymbol{\Omega}_{21} \boldsymbol{\Omega}_{\boldsymbol{\beta}} \mathbf{S}_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) + (\boldsymbol{\Omega}_{22}^- + \boldsymbol{\Omega}_{22}^- \boldsymbol{\Omega}_{21} \boldsymbol{\Omega}_{\boldsymbol{\beta}} \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^-) \mathbf{S}_{\boldsymbol{\zeta}}(y_i, \mathbf{x}_i) \}.
$$

Using this expression, we analyze the order of $\mathbf{d}_{k2}^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$ using Bernstein's inequality. First note that v_{2i} satisfies $E(V_{2i}) = 0$ and $\mathrm{var}(V_{2i}) = \mathbf{d}_{k2}^{\mathrm{T}}(\boldsymbol{\Omega}_{22}^- + \boldsymbol{\Omega}_{22}^- \boldsymbol{\Omega}_{21} \boldsymbol{\Omega}_{\boldsymbol{\beta}} \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^-) \mathbf{d}_{k2} \leqslant C_{21}$ for some constant $C_{21} > 0$ because $\|\mathbf{d}_{k2}\|_2 = O(N^{-1/2})$, $\|\boldsymbol{\zeta}_1^{\mathrm{T}} \boldsymbol{\Omega}_{22}^- \boldsymbol{\zeta}_2\|_2 = O(N \|\boldsymbol{\zeta}_1\|_2 \|\boldsymbol{\zeta}_2\|_2)$ by Lemma B.3.4, and $\|\boldsymbol{\Omega}_{21}\boldsymbol{\beta}\|_2 = O(N^{-1/2}\|\boldsymbol{\beta}\|_2)$ by (B.28). Also, $|v_{2i}| \leqslant C_{22} N^{1/2}$ for some constant $C_{22} > 0$ because $\|\mathbf{d}_{k2}\|_2 = O(N^{-1/2})$, $\|\boldsymbol{\zeta}_1^{\mathrm{T}} \boldsymbol{\Omega}_{22}^- \boldsymbol{\zeta}_2\|_2 = O(N \|\boldsymbol{\zeta}_1\|_2 \|\boldsymbol{\zeta}_2\|_2)$ by Lemma B.3.4, $\|\boldsymbol{\Omega}_{21}\boldsymbol{\beta}\|_2 = O(N^{-1/2}\|\boldsymbol{\beta}\|_2)$ by (B.28), $\|\mathbf{S}_{\boldsymbol{\beta}}(y, \mathbf{x})\|_2 = O(1)$ from Condition (C1), and $\|\mathbf{S}_{\boldsymbol{\zeta}}(y, \mathbf{x})\|_2 \leqslant \|\mathbf{B}_+(y)\|_2 + E\{\|\mathbf{B}_+(Y)\|_2 \mid \mathbf{x}\} \leqslant 2$. Then by Bernstein's inequality, for an arbitrary $\delta > 0$, we can choose $\epsilon = C_2 n^{-1/2}$ where C_2 is a sufficiently large constant so that

$$
\begin{aligned}
\mathrm{pr}\left\{ \left| \mathbf{d}_{k2}^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \right| \geqslant \epsilon \right\} &= \mathrm{pr}\left\{ \left| n^{-1} \sum_{i=1}^n v_{2i} + o_p(n^{-1/2}) \right| \geqslant \epsilon \right\} \\
&\leqslant \mathrm{pr}\left\{ \left| n^{-1} \sum_{i=1}^n v_{2i} \right| \geqslant \epsilon/2 \right\} \\
&\leqslant 2 \exp\left(\frac{-n^2\epsilon^2/8}{C_{21} n + C_{22} n N^{1/2} \epsilon/6} \right) < \delta \qquad \text{(B.78)}
\end{aligned}
$$

when $n \to \infty$ by Condition (C3). Also, note that

$$
\mathrm{pr}\left\{ \left| (\widehat{\mathbf{d}}_{k2} - \mathbf{d}_{k2})^{\mathrm{T}} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \right| \geqslant \epsilon \right\} \leqslant \mathrm{pr}\left\{ \left| \mathbf{d}_{k2}^{\mathrm{T}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \right| \geqslant \epsilon \right\} \qquad \text{(B.79)}
$$

because of the law of large numbers, $\|\boldsymbol{\beta}^{**} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ and $\|\boldsymbol{\zeta}^{**} - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$.

Now, the last term in (B.70) is $O(N^{-q}) = o(n^{-1/2})$ by (B.44), (B.49), Condition

(C3). Therefore, for ϵ such that $\epsilon n^{1/2} \to \infty$ as $n \to \infty$, we have

$$\text{pr}\left(\left|n^{-1}\sum_{i=1}^{n}\left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\} - q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right| \geqslant \epsilon\right) \to 0.$$

Then combining this with (B.70), (B.74), (B.75), (B.78), (B.79), we obtain that for any $\delta > 0$, there exists sufficiently large C_ϵ, so that for $\epsilon = C_\epsilon n^{-1/2}(\log N)^{1/2}$,

$$\begin{aligned}
& \text{pr}\left(\left\|n^{-1}\sum_{i=1}^{n}\left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{B}_+\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\} - q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right\|_\infty \geqslant \epsilon\right) \\
\leqslant & \sum_{k=1}^{m+1}\text{pr}\left(\left|n^{-1}\sum_{i=1}^{n}\left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\} - q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right| \geqslant \epsilon\right) \\
\leqslant & 4(m+1)\exp\left(\frac{-n^2\epsilon^2/128}{C_{11}n + C_{12}nN^{1/2}\epsilon/24}\right) + 4(m+1)\exp\left(\frac{-n^2\epsilon^2/128}{C_{21}n + C_{22}nN^{1/2}\epsilon/24}\right) \\
& + \text{pr}\left(\left|n^{-1}\sum_{i=1}^{n}\left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)B_k\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0, \boldsymbol{\zeta}_0)\} - q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)B_k\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right| \geqslant \epsilon\right) \\
< & \delta. \qquad \text{(B.80)}
\end{aligned}$$

Now, note that $\|\boldsymbol{\zeta}\|_2 = O(N^{1/2}\|\boldsymbol{\zeta}\|_\infty)$ for any $\boldsymbol{\zeta} \in \mathbb{R}^{m+1}$ by Condition (C3). Then we get

$$\begin{aligned}
& \left\|n^{-1}\sum_{i=1}^{n}[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{B}_+\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\} - E\left[q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right\|_2 \\
\leqslant & O(N^{1/2})\left\|n^{-1}\sum_{i=1}^{n}\left[q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{B}_+\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\} - q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right]\right\|_\infty \\
& + O(N^{1/2})\left\|n^{-1}\sum_{i=1}^{n}q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+\{q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\} - E[q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\right\|_\infty \\
= & O(N^{1/2})[O_p\{n^{-1/2}(\log N)^{1/2}\} + O_p\{n^{-1/2}N^{-1/2}(\log N)^{1/2}\}] \\
= & o_p(N^{-1/2}), \qquad \text{(B.81)}
\end{aligned}$$

where the first equality holds by (B.68) and (B.80) and the second equality holds by Condition (C3). Furthermore, we have

$$\begin{aligned}
& |[q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) + q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* + q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{B}_+'\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^{\mathrm{T}}\boldsymbol{\zeta}^*] \\
& -[q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]| \\
= & o_p(1) \qquad \text{(B.82)}
\end{aligned}$$

by (B.43), (B.46), (B.59), and $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$.

Now, (B.64), (B.66) and (B.62) lead to

$$\begin{aligned}
\frac{\|E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{x})\|_2}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}} &= O(N^{-1/2}), && \text{(B.83)}\\
\|E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+^{\mathrm{T}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\|_2 &= O(N^{1/2})\|E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+^{\mathrm{T}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\|_\infty\\
&= O(N^{-1/2}), && \text{(B.84)}\\
\frac{\|E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{x})\|_2}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}} &= O(N^{1/2})\\
&\quad \times \frac{\|E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{x})\|_\infty}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}}\\
&= O(N^{-1/2}), && \text{(B.85)}
\end{aligned}$$

because $\|\boldsymbol{\zeta}\|_2 = O(N^{1/2}\|\boldsymbol{\zeta}\|_\infty)$ for $\boldsymbol{\zeta} \in \mathbb{R}^{m+1}$ and $f_{Y|\mathbf{X}}(\cdot \mid \mathbf{x})$ is bounded from below by a positive constant by Condition (C1). In addition, we have

$$|q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}| \leqslant C \qquad \text{(B.86)}$$

for some constant $C > 0$ because $q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \in [0,1]$, $q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = E([\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}Y \mid \mathbf{x})/f_{Y|\mathbf{X}}(y, \mathbf{x})$ is bounded because $Y \in [0,1]$ and $f_{Y|\mathbf{X}}(\cdot \mid \mathbf{x})$ is strictly positive by Condition (C1), $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$ and $c'(\cdot)$ are bounded by Condition (C1). Then we further get

$$\begin{aligned}
\|\mathbf{C}_2\mathbf{T}^{\mathrm{T}}\|_2 &= \Bigg\| \boldsymbol{\beta}_0 E\Bigg\{ \frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y\{\mathbf{B}_+^{\mathrm{T}}(Y) - \mathbf{e}_1^{\mathrm{T}}\}|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}}\\
&\quad -q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)[\mathbf{B}_+^{\mathrm{T}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} - \mathbf{e}_1^{\mathrm{T}}]\\
&\quad -\frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\{\mathbf{B}_+^{\mathrm{T}}(Y) - \mathbf{e}_1^{\mathrm{T}}\}|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}}\\
&\quad \times[q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\}\Bigg\|_2\\
&= \Bigg\| \boldsymbol{\beta}_0 E\Bigg\{ \frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}}\\
&\quad -q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+^{\mathrm{T}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} - \frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}}\\
&\quad \times[q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\}\Bigg\|_2\\
&\leqslant \|\boldsymbol{\beta}_0\|_2 \Bigg\| E\Bigg\{ \frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}}
\end{aligned}$$

$$
\begin{aligned}
& -q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+^{\mathrm{T}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} - \frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}} \\
& \times [q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\big\|_2 \\
\leqslant\ & \|\boldsymbol{\beta}_0\|_2 E\Bigg\{ \frac{\|E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{X})\|_2}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}} \\
& + \|q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+^{\mathrm{T}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\|_2 + \frac{\|E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{X})\|_2}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}} \\
& \times |q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}|\Bigg\} \\
=\ & O(N^{-1/2}), \qquad \text{(B.87)}
\end{aligned}
$$

where the second equality holds because $-E([\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y|\mathbf{x})/f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\} + q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = 0$ and $E[\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}|\mathbf{x}] = 0$ by the definition of $q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)$, and the last equality holds since the first term is $O(N^{-1/2})$ by (B.83), the second term is $O(N^{-1/2})$ by (B.84), and the last term is $O(N^{-1/2})$ by (B.85), and because of (B.86).

Using (B.52), we further get

$$
\begin{aligned}
& \left\| \boldsymbol{\beta}_0 n^{-1} \sum_{i=1}^{n} \frac{\partial q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} - \mathbf{C}_2\mathbf{T}^{\mathrm{T}} \right\|_2 \\
\leqslant\ & \|\boldsymbol{\beta}_0\|_2 \Bigg\| n^{-1} \sum_{i=1}^{n} \frac{E^*([\tau - I\{Y \leqslant q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}]Y\mathbf{B}_+(Y)|\mathbf{x}_i, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)}{f^*_{Y|\mathbf{X}}\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*), \mathbf{x}_i, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*\}} \\
& - E\left\{ \frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]Y\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}} \right\} \Bigg\|_2 \\
& + \|\boldsymbol{\beta}_0\|_2 \left\| n^{-1} \sum_{i=1}^{n} q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{B}_+\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\} - E\left[q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{B}_+^{\mathrm{T}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right] \right\|_2 \\
& + \|\boldsymbol{\beta}_0\|_2 \Bigg\| n^{-1} \sum_{i=1}^{n} \frac{\partial q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)}{\partial \boldsymbol{\zeta}} [q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) + q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* \\
& + q^{*\prime}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{B}'_+\{q^*(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^{\mathrm{T}}\boldsymbol{\zeta}^*] - E\left\{ \frac{E([\tau - I\{Y \leqslant q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{X})}{f_{Y|\mathbf{X}}\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{X}\}} \right. \\
& \left. \times [q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}] \right\} \Big\|_2 \qquad \text{(B.88)} \\
=\ & o_p(N^{-1/2}), \qquad \text{(B.89)}
\end{aligned}
$$

where the last equality holds because the first term in (B.88) is $o_p(N^{-1/2})$ by (B.65), the second term in (B.88) is $o_p(N^{-1/2})$ by (B.81), and the last term in (B.88) is $o_p(N^{-1/2})$

since

$$
\begin{aligned}
&\left\| \frac{\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)}{\partial \boldsymbol{\zeta}} [q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) + q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* \right. \\
&+ q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{B}_+'\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^{\mathrm{T}}\boldsymbol{\zeta}^*] \\
&- \frac{E([\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}} \\
&\left. \times [q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}] \right\|_2 \\
\leqslant\ & \left\| \frac{\partial q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)}{\partial \boldsymbol{\zeta}} - \frac{E([\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}} \right\|_2 \\
&\times |q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) + q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* + q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{B}_+'\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^{\mathrm{T}}\boldsymbol{\zeta}^*| \\
&+ \left\| \frac{E([\tau - I\{Y \leqslant q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\mathbf{B}_+^{\mathrm{T}}(Y)|\mathbf{x})}{f_{Y|\mathbf{X}}\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{x}\}} \right\|_2 \\
&\times |[q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) + q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* + q^{*\prime}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\mathbf{B}_+'\{q^*(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\}^{\mathrm{T}}\boldsymbol{\zeta}^*] \\
&- [q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) + q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 + q'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)c'\{q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]|,
\end{aligned}
$$

where the first term is $o_p(N^{-1/2})$ by (B.63), (B.82), (B.86), and the second term is $o_p(N^{-1/2})$ by (B.82) and (B.85).

Inserting (B.61) and (B.89) in (B.53), using Lemma B.6.1, we get

$$
\begin{aligned}
\widehat{\boldsymbol{\eta}_\tau} - \boldsymbol{\eta}_{\tau 0} &= \mathbf{C}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \mathbf{C}_2\mathbf{T}^{\mathrm{T}}\{\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\zeta}_0\} \\
&\quad + \boldsymbol{\beta}_0 \left[n^{-1}\sum_{i=1}^n q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) - E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} \right] + \mathbf{r} \\
&= \mathbf{C}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\boldsymbol{\Omega}^- n^{-1}\sum_{i=1}^n \mathbf{S}(y_i, \mathbf{x}_i) \\
&\quad + \boldsymbol{\beta}_0 \left[n^{-1}\sum_{i=1}^n q'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) - E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} \right] + \mathbf{C}_1\mathbf{r}_2 + \mathbf{C}_2\mathbf{T}^{\mathrm{T}}\mathbf{r}_3 + \mathbf{r},
\end{aligned}
$$

where $\|\mathbf{r}\|_2 = o_p(n^{-1/2})$ by Proposition 3.3.2 and Lemma B.6.1, hence

$$
\|\mathbf{C}_1\mathbf{r}_2 + \mathbf{C}_2\mathbf{T}^{\mathrm{T}}\mathbf{r}_3 + \mathbf{r}\|_2 = o_p(n^{-1/2})
$$

by Lemma B.6.1 and (B.87). In addition, we obviously have

$$
\begin{aligned}
\mathrm{cov}[\mathbf{X}\{Y - E(Y|\mathbf{X})\}, q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)] &= \mathbf{0}_p, \\
\mathrm{cov}[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{X}\}, q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)] &= \mathbf{0}_m.
\end{aligned}
$$

Thus with

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\eta}_\tau} &= \mathbf{C}\boldsymbol{\Sigma}^{-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\mathrm{var}\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} \\ &= \mathbf{C}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\boldsymbol{\Omega}^{-}\mathrm{diag}(\mathbf{I}_p, \mathbf{T})\mathbf{C}^{\mathrm{T}} + \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\mathrm{var}\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\end{aligned}$$

by Lemma B.3.3, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_\tau}^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\eta}}_\tau - \boldsymbol{\eta}_{\tau 0})$ converges to the normal distribution with mean $\mathbf{0}$ and variance $\mathbf{I}$. □

B.9 Proof of Theorem 3.3.3

First note that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by Proposition 3.3.2 and $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N)$ by Lemma B.6.1. We have $\|\boldsymbol{\Omega}_{11}\|_2 = O(1)$ by Condition (C1), and $\|\widehat{\boldsymbol{\Omega}}_{11} - \boldsymbol{\Omega}_{11}\|_2 = \|\widehat{\boldsymbol{\Omega}}^*_{11}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) - \boldsymbol{\Omega}_{11}\|_2 = o_p(1)$ by (B.30) with replacing $\boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\beta}}$. In addition, we have $\|\boldsymbol{\Omega}_{21}\boldsymbol{\beta}\|_2 = O(N^{-1/2}\|\boldsymbol{\beta}\|_2)$ by (B.28), and $\|(\widehat{\boldsymbol{\Omega}}_{21} - \boldsymbol{\Omega}_{21})\boldsymbol{\beta}\|_2 = \|\{\widehat{\boldsymbol{\Omega}}^*_{21}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) - \boldsymbol{\Omega}_{21}\}\boldsymbol{\beta}\|_2 = o_p(N^{-1/2}\|\boldsymbol{\beta}\|_2)$ by (B.32) with replacing $\boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\beta}}$. In addition, $\boldsymbol{\zeta}_1^{\mathrm{T}}\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\zeta}_2 = O(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2)$ by Lemma B.3.5, and $\boldsymbol{\zeta}_1(\widehat{\boldsymbol{\Omega}}_{22}^{-} - \boldsymbol{\Omega}_{22}^{-})\boldsymbol{\zeta}_2 = \boldsymbol{\zeta}_1\{\widehat{\boldsymbol{\Omega}}^*_{22}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) - \boldsymbol{\Omega}_{22}^{-}\}\boldsymbol{\zeta}_2 = o_p(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2)$ by (B.31) with replacing $\boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\beta}}$. Lastly, we have $\|\boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 = O(1)$ by Lemma B.3.2, and $\|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\beta}} - \boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 = \|\widehat{\boldsymbol{\Omega}}^*_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) - \boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 = o_p(1)$ by (B.33) with replacing $\boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\beta}}$. To summarize, we have

$$\begin{aligned}\|\boldsymbol{\Omega}_{11}\|_2 &= O(1), \qquad\qquad (\text{B.90}) \\ \|\widehat{\boldsymbol{\Omega}}_{11} - \boldsymbol{\Omega}_{11}\|_2 &= o_p(1), \\ \|\widehat{\boldsymbol{\Omega}}_{11}\|_2 &\leqslant \|\boldsymbol{\Omega}_{11}\|_2 + \|\widehat{\boldsymbol{\Omega}}_{11} - \boldsymbol{\Omega}_{11}\|_2 = O_p(1), \\ \|\boldsymbol{\Omega}_{21}\boldsymbol{\beta}\|_2 &= O(N^{-1/2}\|\boldsymbol{\beta}\|_2), \\ \|(\widehat{\boldsymbol{\Omega}}_{21} - \boldsymbol{\Omega}_{21})\boldsymbol{\beta}\|_2 &= o_p(N^{-1/2}\|\boldsymbol{\beta}\|_2), \\ \|\widehat{\boldsymbol{\Omega}}_{21}\boldsymbol{\beta}\|_2 &\leqslant \|\boldsymbol{\Omega}_{21}\boldsymbol{\beta}\|_2 + \|(\widehat{\boldsymbol{\Omega}}_{21} - \boldsymbol{\Omega}_{21})\boldsymbol{\beta}\|_2 = O_p(N^{-1/2}\|\boldsymbol{\beta}\|_2), \\ \boldsymbol{\zeta}_1\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\zeta}_2 &= O(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2), \\ \boldsymbol{\zeta}_1(\widehat{\boldsymbol{\Omega}}_{22}^{-} - \boldsymbol{\Omega}_{22}^{-})\boldsymbol{\zeta}_2 &= o_p(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2), \\ \boldsymbol{\zeta}_1\widehat{\boldsymbol{\Omega}}_{22}^{-}\boldsymbol{\zeta}_2 &\leqslant |\boldsymbol{\zeta}_1\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\zeta}_2| + |\boldsymbol{\zeta}_1(\widehat{\boldsymbol{\Omega}}_{22}^{-} - \boldsymbol{\Omega}_{22}^{-})\boldsymbol{\zeta}_2| = O_p(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2), \\ \|\boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 &= O(1), \\ \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\beta}} - \boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 &= o_p(1), \\ \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\beta}}\|_2 &\leqslant \|\boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 + \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\beta}} - \boldsymbol{\Omega}_{\boldsymbol{\beta}}\|_2 = O_p(1).\end{aligned}$$

Since $\mathbf{\Sigma}_{\boldsymbol{\beta}} = \mathbf{\Omega}_{\boldsymbol{\beta}}$ by Lemma B.3.3 and similarly $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\beta}} = \widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}}$, we have

$$\|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\beta}} - \mathbf{\Sigma}_{\boldsymbol{\beta}}\|_2 = \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}} - \mathbf{\Omega}_{\boldsymbol{\beta}}\|_2 = o_p(1).$$

In addition, using Lemma B.3.3, $\mathbf{\Sigma}^{-1}$ can be expressed as

$$\begin{aligned} \mathbf{\Sigma}^{-1} &= \mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\mathbf{\Omega}^{-}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}) \\ &= \mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}}) \begin{bmatrix} \mathbf{\Omega}_{\boldsymbol{\beta}} & -\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^{-} \\ -\mathbf{\Omega}_{22}^{-}\mathbf{\Omega}_{21}\mathbf{\Omega}_{\boldsymbol{\beta}} & \mathbf{\Omega}_{22}^{-} + \mathbf{\Omega}_{22}^{-}\mathbf{\Omega}_{21}\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^{-} \end{bmatrix} \mathrm{diag}(\mathbf{I}_p, \mathbf{T}). \end{aligned} \tag{B.91}$$

Using the results in (B.90), we can easily show that

$$\begin{aligned} \|\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^{-}\boldsymbol{\zeta}\|_2 &= O(N^{1/2}\|\boldsymbol{\zeta}\|_2), \\ \|(\widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}}\widehat{\mathbf{\Omega}}_{12}\widehat{\mathbf{\Omega}}_{22}^{-} - \mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^{-})\boldsymbol{\zeta}\|_2 &= o_p(N^{1/2}\|\boldsymbol{\zeta}\|_2), \\ \|\boldsymbol{\zeta}_1(\mathbf{\Omega}_{22}^{-} + \mathbf{\Omega}_{22}^{-}\mathbf{\Omega}_{21}\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^{-})\boldsymbol{\zeta}_2\|_2 &= O(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2), \end{aligned}$$

and

$$\|\boldsymbol{\zeta}_1\{(\widehat{\mathbf{\Omega}}_{22}^{-} + \widehat{\mathbf{\Omega}}_{22}^{-}\widehat{\mathbf{\Omega}}_{21}\widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}}\widehat{\mathbf{\Omega}}_{12}\widehat{\mathbf{\Omega}}_{22}^{-}) - (\mathbf{\Omega}_{22}^{-} + \mathbf{\Omega}_{22}^{-}\mathbf{\Omega}_{21}\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^{-})\}\boldsymbol{\zeta}_2\|_2 = o_p(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2).$$

This implies

$$\begin{aligned} & (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\zeta}_1^{\mathrm{T}})\mathbf{\Omega}^{-}(\boldsymbol{\beta}_2^{\mathrm{T}}, \boldsymbol{\zeta}_2^{\mathrm{T}})^{\mathrm{T}} \\ = \; & O\{\|\boldsymbol{\beta}_1\|_2\|\boldsymbol{\beta}_2\|_2 + N^{1/2}(\|\boldsymbol{\beta}_1\|_2\|\boldsymbol{\zeta}_2\|_2 + \|\boldsymbol{\beta}_2\|_2\|\boldsymbol{\zeta}_1\|_2) + N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2\}, \\ & (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\zeta}_1^{\mathrm{T}})(\widehat{\mathbf{\Omega}}^{-} - \mathbf{\Omega}^{-})(\boldsymbol{\beta}_2^{\mathrm{T}}, \boldsymbol{\zeta}_2^{\mathrm{T}})^{\mathrm{T}} \\ = \; & o_p\{\|\boldsymbol{\beta}_1\|_2\|\boldsymbol{\beta}_2\|_2 + N^{1/2}(\|\boldsymbol{\beta}_1\|_2\|\boldsymbol{\zeta}_2\|_2 + \|\boldsymbol{\beta}_2\|_2\|\boldsymbol{\zeta}_1\|_2) + N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2\}, \\ & (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\zeta}_1^{\mathrm{T}})\widehat{\mathbf{\Omega}}^{-}(\boldsymbol{\beta}_2^{\mathrm{T}}, \boldsymbol{\zeta}_2^{\mathrm{T}})^{\mathrm{T}} \\ = \; & O_p\{\|\boldsymbol{\beta}_1\|_2\|\boldsymbol{\beta}_2\|_2 + N^{1/2}(\|\boldsymbol{\beta}_1\|_2\|\boldsymbol{\zeta}_2\|_2 + \|\boldsymbol{\beta}_2\|_2\|\boldsymbol{\zeta}_1\|_2) + N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2\}. \end{aligned} \tag{B.92}$$

Now we will show $\|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}} - \mathbf{\Sigma}_{\boldsymbol{\xi}}\|_2 = o_p(1)$. (B.91) leads to

$$\begin{aligned} & \|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}} - \mathbf{\Sigma}_{\boldsymbol{\xi}}\|_2 \\ \leqslant \; & \|\widehat{\mathbf{A}}\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{A}}^{\mathrm{T}} - \mathbf{A}\mathbf{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}}\|_2 + \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\mathrm{var}}\{\widehat{\mathrm{var}}(Y \mid \mathbf{X})\} - \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\mathrm{var}\{\mathrm{var}(Y \mid \mathbf{X})\}\|_2 \\ = \; & \|\widehat{\mathbf{A}}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\widehat{\mathbf{\Omega}}^{-}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\widehat{\mathbf{A}}^{\mathrm{T}} - \mathbf{A}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\mathbf{\Omega}^{-}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\mathbf{A}^{\mathrm{T}}\|_2 \\ & + \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\mathrm{var}}\{\widehat{\mathrm{var}}(Y \mid \mathbf{X})\} - \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\mathrm{var}\{\mathrm{var}(Y \mid \mathbf{X})\}\|_2 \end{aligned}$$

$$
\begin{aligned}
= \;& \|(\widehat{\mathbf{A}} - \mathbf{A})\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\widehat{\boldsymbol{\Omega}}^{-}\mathrm{diag}(\mathbf{I}_p, \mathbf{T})\widehat{\mathbf{A}}^{\mathrm{T}}\|_2 \\
& + \|\mathbf{A}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})(\widehat{\boldsymbol{\Omega}}^{-} - \boldsymbol{\Omega}^{-})\mathrm{diag}(\mathbf{I}_p, \mathbf{T})\widehat{\mathbf{A}}^{\mathrm{T}}\|_2 \\
& + \|\mathbf{A}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\boldsymbol{\Omega}^{-}\mathrm{diag}(\mathbf{I}_p, \mathbf{T})(\widehat{\mathbf{A}} - \mathbf{A})^{\mathrm{T}}\|_2 \\
& + \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\mathrm{var}}\{\widehat{\mathrm{var}}(Y \mid \mathbf{X})\}\|_2 \\
& + \|\boldsymbol{\beta}_0(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\widehat{\mathrm{var}}\{\widehat{\mathrm{var}}(Y \mid \mathbf{X})\}\|_2 \\
& + \|\boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}[\widehat{\mathrm{var}}\{\widehat{\mathrm{var}}(Y \mid \mathbf{X})\} - \mathrm{var}\{\mathrm{var}(Y \mid \mathbf{X})\}]\|_2. \qquad \text{(B.93)}
\end{aligned}
$$

First note that $\|\mathbf{A}_1\|_2$ is of constant order by Condition (C1), and

$$
\begin{aligned}
\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 \;\leqslant\; & |\widehat{E}\{\widehat{\mathrm{var}}(Y|\mathbf{X})\} - E\{\mathrm{var}(Y|\mathbf{X})\}| \\
& + \|\widehat{\boldsymbol{\beta}}\|_2\|\widehat{E}[\{Y - \widehat{E}(Y|\mathbf{X})\}^3\mathbf{X}] - E[\{Y - E(Y|\mathbf{X})\}^3\mathbf{X}]\|_2 \\
& + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2\|E[\{Y - \widehat{E}(Y|\mathbf{X})\}^3\mathbf{X}]\|_2 \\
= \;& o_p(1)
\end{aligned}
$$

because $|\widehat{E}\{\widehat{\mathrm{var}}(Y|\mathbf{X})\} - E\{\mathrm{var}(Y|\mathbf{X})\}| = o_p(1)$ by (B.35), $\|\widehat{E}[\{Y - \widehat{E}(Y|\mathbf{X})\}^3\mathbf{X}] - E[\{Y - E(Y|\mathbf{X})\}^3\mathbf{X}]\|_2 = o_p(1)$ by (B.37), and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by Proposition 3.3.2. Also, $\|\mathbf{A}_2\mathbf{T}^{\mathrm{T}}\|_2 = O(N^{-1/2})$ by (B.39), and

$$
\begin{aligned}
\|(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)\mathbf{T}^{\mathrm{T}}\|_2 \;\leqslant\; & \|\widehat{\boldsymbol{\beta}}\|_2\|\widehat{E}(\{Y - \widehat{E}(Y|\mathbf{X})\}^2[\mathbf{B}_+(Y) - \widehat{E}\{\mathbf{B}_+(Y)|\mathbf{X}\}]) \\
& - E(\{Y - E(Y|\mathbf{X})\}^2[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{X}\}])\|_2 \\
& + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2\|E(\{Y - E(Y|\mathbf{X})\}^2[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{X}\}])\|_2 \\
= \;& o_p(N^{-1/2})
\end{aligned}
$$

since

$$
\begin{aligned}
& \|\widehat{E}(\{Y - \widehat{E}(Y|\mathbf{X})\}^2[\mathbf{B}_+(Y) - \widehat{E}\{\mathbf{B}_+(Y)|\mathbf{X}\}]) \\
& - E(\{Y - E(Y|\mathbf{X})\}^2[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{X}\}])\|_2 \\
= \;& o_p(N^{-1/2})
\end{aligned}
$$

by (B.38) and (B.40), $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by Proposition 3.3.2, and $\|E(\{Y - E(Y|\mathbf{X})\}^2[\mathbf{B}_+(Y) - E\{\mathbf{B}_+(Y)|\mathbf{X}\}])\|_2 = O_p(N^{-1/2}) = o_p(1)$ by (B.39). Further, we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ from Proposition 3.3.2 and $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\|_2 = O_p(n^{-1/2}N) = o_p(1)$ from Lemma B.6.1 and Condition (C3), and these further lead to $\|f^*_{Y|\mathbf{X}}(\cdot, \mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) - f_{Y|\mathbf{X}}(\cdot, \mathbf{x})\|_\infty = o_p(1)$ since y is bounded between 0 and 1 by Condition (C1). Then we

can show that $\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\text{var}}(Y|\mathbf{x}) - \text{var}(Y|\mathbf{x})| = o_p(1)$ since the support of $\mathbf{x}$ is compact by Condition (C1). Thus,

$$\begin{aligned}
& \widehat{\text{var}}\{\widehat{\text{var}}(Y|\mathbf{X})\} - \text{var}\{\text{var}(Y|\mathbf{X})\} \\
= \; & \widehat{E}[\{\widehat{\text{var}}(Y|\mathbf{X})\}^2 - \{\text{var}(Y|\mathbf{X})\}^2] + \widehat{E}[\{\text{var}(Y|\mathbf{X})\}^2] - E[\{\text{var}(Y|\mathbf{X})\}^2] \\
& -[\widehat{E}\{\widehat{\text{var}}(Y|\mathbf{X})\}]^2 + [\widehat{E}\{\text{var}(Y|\mathbf{X})\}]^2 - [\widehat{E}\{\text{var}(Y|\mathbf{X})\}]^2 + [E\{\text{var}(Y|\mathbf{X})\}]^2 \\
= \; & o_p(1).
\end{aligned}$$

To summarize, we have

$$\begin{aligned}
\|\mathbf{A}_1\|_2 &= O(1), \\
\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 &= o_p(1), \\
\|\widehat{\mathbf{A}}_1\|_2 &\leqslant \|\mathbf{A}_1\|_2 + \|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(1), \\
\|\mathbf{A}_2\mathbf{T}^{\mathrm{T}}\|_2 &= O(N^{-1/2}), \\
\|(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)\mathbf{T}^{\mathrm{T}}\|_2 &= o_p(N^{-1/2}), \\
\|\widehat{\mathbf{A}}_2\mathbf{T}^{\mathrm{T}}\|_2 &\leqslant \|\mathbf{A}_2\mathbf{T}^{\mathrm{T}}\|_2 + \|(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)\mathbf{T}^{\mathrm{T}}\|_2 = O_p(N^{-1/2}), \\
\widehat{\text{var}}\{\widehat{\text{var}}(Y|\mathbf{X})\} - \text{var}\{\text{var}(Y|\mathbf{X})\} &= o_p(1).
\end{aligned}$$

Therefore, the above results, (B.92), and (B.93) lead to

$$\begin{aligned}
& \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}}\|_2 \\
\leqslant \; & O_p[o_p(1)O_p(1) + N^{1/2}\{o_p(1)O_p(N^{-1/2}) + O_p(1)o_p(N^{-1/2})\} + No_p(N^{-1/2})O_p(N^{-1/2}\}] \\
& + o_p[O(1)O_p(1) + N^{1/2}\{O(1)O_p(N^{-1/2}) + O_p(1)O(N^{-1/2})\} + NO(N^{-1/2})O_p(N^{-1/2}\}] \\
& + O[O(1)o_p(1) + N^{1/2}\{O(1)o_p(N^{-1/2}) + o_p(1)O(N^{-1/2})\} + NO(N^{-1/2})o_p(N^{-1/2}\}] \\
& + O_p(n^{-1/2}) + O_p(n^{-1/2}) + o_p(1) \\
= \; & o_p(1),
\end{aligned}$$

where we used $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by Proposition 3.3.2.

Lastly, we will show $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_\tau} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}_\tau}\|_2 = o_p(1)$. (B.91) leads to

$$\begin{aligned}
\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\|_2 &\leqslant \|\widehat{\mathbf{C}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}}^{\mathrm{T}} - \mathbf{C}\boldsymbol{\Sigma}^{-1}\mathbf{C}^{\mathrm{T}}\|_2 + \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\text{var}}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\} - \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\text{var}\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\|_2 \\
&= \|\widehat{\mathbf{C}}\text{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\widehat{\boldsymbol{\Omega}}^{-}\text{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\widehat{\mathbf{C}}^{\mathrm{T}} - \mathbf{C}\text{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\boldsymbol{\Omega}^{-}\text{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\mathbf{C}^{\mathrm{T}}\|_2 \\
&\quad + \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\text{var}}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\} - \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\text{var}\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\|_2 \\
&\leqslant \|(\widehat{\mathbf{C}} - \mathbf{C})\text{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\widehat{\boldsymbol{\Omega}}^{-}\text{diag}(\mathbf{I}_p, \mathbf{T})\widehat{\mathbf{C}}^{\mathrm{T}}\|_2
\end{aligned}$$

$$
\begin{aligned}
&+\|\mathbf{C}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})(\widehat{\boldsymbol{\Omega}}^{-} - \boldsymbol{\Omega}^{-})\mathrm{diag}(\mathbf{I}_p, \mathbf{T})\widehat{\mathbf{C}}^{\mathrm{T}}\|_2 \\
&+\|\mathbf{C}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}})\boldsymbol{\Omega}^{-}\mathrm{diag}(\mathbf{I}_p, \mathbf{T})(\widehat{\mathbf{C}} - \mathbf{C})^{\mathrm{T}}\|_2 \\
&+\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\mathrm{var}}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}\|_2 \\
&+\|\boldsymbol{\beta}_0(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\widehat{\mathrm{var}}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}\|_2 \\
&+\|\boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}[\widehat{\mathrm{var}}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\} - \mathrm{var}\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\|_2. \qquad \text{(B.94)}
\end{aligned}
$$

The boundedness on $\mathbf{X}$ and Y from Condition (C1) immediately implies $\|\mathbf{C}_1\|_2$ is of constant order, and $\|\widehat{\mathbf{C}}_1 - \mathbf{C}_1\|_2 = o_p(1)$ by (B.61). Also, we have $\|\mathbf{C}_2\mathbf{T}^{\mathrm{T}}\|_2 = O(N^{-1/2})$ from (B.87), and $\|(\widehat{\mathbf{C}}_2 - \mathbf{C}_2)\mathbf{T}^{\mathrm{T}}\|_2 = o_p(N^{-1/2})$ by (B.89). Now, (B.43) implies

$$
\begin{aligned}
&\widehat{\mathrm{var}}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\} - \mathrm{var}\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} \\
= \;& \widehat{E}[\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}^2 - \{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2] + \widehat{E}[\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2] - E[\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2] \\
& -[\widehat{E}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}]^2 + [\widehat{E}\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]^2 - [\widehat{E}\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]^2 + [E\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}]^2 \\
= \;& o_p(1).
\end{aligned}
$$

To summarize, we have

$$
\begin{aligned}
\|\mathbf{C}_1\|_2 &= O(1), \\
\|\widehat{\mathbf{C}}_1 - \mathbf{C}_1\|_2 &= o_p(1), \\
\|\widehat{\mathbf{C}}_1\|_2 &\leqslant \|\mathbf{C}_1\|_2 + \|\widehat{\mathbf{C}}_1 - \mathbf{C}_1\|_2 = O_p(1), \\
\|\mathbf{C}_2\mathbf{T}^{\mathrm{T}}\|_2 &= O(N^{-1/2}), \\
\|(\widehat{\mathbf{C}}_2 - \mathbf{C}_2)\mathbf{T}^{\mathrm{T}}\|_2 &= o_p(N^{-1/2}), \\
\|\widehat{\mathbf{C}}_2\mathbf{T}^{\mathrm{T}}\|_2 &\leqslant \|\mathbf{C}_2\mathbf{T}^{\mathrm{T}}\|_2 + \|(\widehat{\mathbf{C}}_2 - \mathbf{C}_2)\mathbf{T}^{\mathrm{T}}\|_2 = O_p(N^{-1/2}), \\
\widehat{\mathrm{var}}\{\widehat{q}'(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\} - \mathrm{var}\{q'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\} &= o_p(1).
\end{aligned}
$$

Therefore, the above results, (B.92), and (B.94) lead to

$$
\begin{aligned}
&\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\|_2 \\
\leqslant \;& O_p[o_p(1)O_p(1) + N^{1/2}\{o_p(1)O_p(N^{-1/2}) + O_p(1)o_p(N^{-1/2})\} + No_p(N^{-1/2})O_p(N^{-1/2}\}] \\
& +o_p[O(1)O_p(1) + N^{1/2}\{O(1)O_p(N^{-1/2}) + O_p(1)O(N^{-1/2})\} + NO(N^{-1/2})O_p(N^{-1/2}\}] \\
& +O[O(1)o_p(1) + N^{1/2}\{O(1)o_p(N^{-1/2}) + o_p(1)O(N^{-1/2})\} + NO(N^{-1/2})o_p(N^{-1/2}\}] \\
& +O_p(n^{-1/2}) + O_p(n^{-1/2}) + o_p(1) \\
= \;& o_p(1),
\end{aligned}
$$

where we used $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by Proposition 3.3.2.

B.10 Proof of Proposition 3.3.3

In terms of estimating $\boldsymbol{\beta}$, the score function is $\mathbf{S}_{\boldsymbol{\beta}} = y\mathbf{x} - E(Y\mathbf{x} \mid \mathbf{x})$ and the nuisance tangent space is

$$\Lambda \;=\; [\mathbf{a}(y) - E\{\mathbf{a}(Y) \mid \mathbf{x}\} + \mathbf{b}(\mathbf{x}) : \forall \mathbf{a}(y), \mathbf{b}(\mathbf{x}) \in \mathcal{R}^p, E\{\mathbf{b}(\mathbf{X})\} = \mathbf{0}].$$

Its orthogonal complement is

$$\Lambda^{\perp} = (\mathbf{a}(y, \mathbf{x}) - E\{\mathbf{a}(Y, \mathbf{x}) \mid \mathbf{x}\} : E\{\mathbf{a}(y, \mathbf{X}) \mid y\} = E[E\{\mathbf{a}(Y, \mathbf{X}) \mid \mathbf{X}\} \mid y]).$$

Thus, the efficient score is $\mathbf{S}_{\text{eff}} = y\mathbf{x} - \mathbf{a}_0(y) - E\{Y\mathbf{x} - \mathbf{a}_0(Y) \mid \mathbf{x}\}$, where $\mathbf{a}_0(y)$ satisfies

$$\mathbf{a}_0(y) - E[E\{\mathbf{a}_0(Y) \mid \mathbf{X}\} \mid y] = E(y\mathbf{X} \mid y) - E\{E(Y\mathbf{X} \mid \mathbf{X}) \mid y\}. \tag{B.95}$$

Thus, the efficient variance is $\{E(\mathbf{S}_{\text{eff}}^{\otimes 2})\}^{-1}$.

To show that the MLE estimator $\widehat{\boldsymbol{\beta}}$ is actually efficient, we only need to show $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \to E(\mathbf{S}_{\text{eff}}^{\otimes 2})$ when $n \to \infty$. Let $\mathbf{a}_0(y) = \Lambda \mathbf{B}_+(y) + O(N^{-q})$, where Λ is a $p \times (m+1)$ coefficient matrix. Then on (B.95), we right-multiply $\mathbf{B}_+^{\mathrm{T}}(y)$ and take expectation to get

$$\begin{aligned} & \Lambda E\{\mathbf{B}_+(Y)^{\otimes 2}\} - \Lambda E[E\{\mathbf{B}_+(Y) \mid \mathbf{X}\}^{\otimes 2}] \\ = \;& E[E\{Y\mathbf{X} - E(Y\mathbf{X} \mid \mathbf{X})\mathbf{B}_+^{\mathrm{T}}(Y) \mid \mathbf{X}\}] + O(N^{-q}) E\{\mathbf{B}_+^{\mathrm{T}}(Y)\} \\ = \;& E[\mathbf{X}\mathrm{cov}\{Y, \mathbf{B}_+(Y) \mid \mathbf{X}\}] + O(N^{-q}) E\{\mathbf{B}_+^{\mathrm{T}}(Y)\}, \end{aligned}$$

i.e.,

$$\begin{aligned} \mathbf{a}_0(y) \;&=\; \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\mathbf{B}_+(y) + O(N^{-q}) E\{\mathbf{B}_+^{\mathrm{T}}(Y)\}\boldsymbol{\Omega}_{22}^{-}\mathbf{B}_+(y) \\ &=\; \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}\mathbf{B}_+(y) + O(N^{1/2-q}), \end{aligned}$$

where we used $\|E\{\mathbf{B}_+(Y)\}\|_2 = O[N^{1/2}\|E\{\mathbf{B}_+(Y)\}\|_\infty] = O\{N^{1/2} \max_{k=1,\ldots,m+1} \|B_k(\cdot)\|_1\} = O(N^{-1/2})$ by Lemma B.3.4, $\|\boldsymbol{\zeta}_1\boldsymbol{\Omega}_{22}^{-}\boldsymbol{\zeta}_2\|_2 = O(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2)$ by Lemma B.3.5, and $\|\mathbf{B}_+(y)\|_2 = O(1)$. Therefore, since $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{\Omega}_{\boldsymbol{\beta}}$ by Lemma B.3.3,

$$E(\mathbf{S}_{\text{eff}}^{\otimes 2}) - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$$

$$
\begin{aligned}
&= E\left\{(Y\mathbf{X} - E(Y\mathbf{X} \mid \mathbf{X}) - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-}[\mathbf{B}_{+}(y) - E\{\mathbf{B}_{+}(Y) \mid \mathbf{X}\}])^{\otimes 2}\right\} - \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1} + O(N^{1/2-q}) \\
&= o(1).
\end{aligned}
$$

□

B.11 Proof of Theorem 3.3.4

Since

$$
\begin{aligned}
E(\boldsymbol{\phi}_{\text{eff}}^{\otimes 2}) &= E\left([\boldsymbol{\beta} v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) - \boldsymbol{\beta} E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}]^{\otimes 2}\right) \\
&\quad + E\left([\boldsymbol{\beta} Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y - E\{\boldsymbol{\beta} Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y \mid \mathbf{X}\}]^{\otimes 2}\right)
\end{aligned}
$$

and

$$
\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\beta}^{\otimes 2}\mathrm{var}\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\},
$$

we only need to show

$$
E\left([\boldsymbol{\beta} Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y - E\{\boldsymbol{\beta} Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y \mid \mathbf{X}\}]^{\otimes 2}\right) - \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}} \to \mathbf{0}.
$$

Now we have $\boldsymbol{\beta} y^2 + \mathbf{a}(y) = \boldsymbol{\Lambda}\mathbf{B}_{+}(y) + O(N^{-q})$, where $\boldsymbol{\Lambda} \in \mathcal{R}^{p\times(m+1)}$. Then (B.5) and (B.6) imply

$$
\begin{aligned}
&-\boldsymbol{\Lambda} E[\mathbf{B}_{+}(y) - E\{\mathbf{B}_{+}(Y) \mid \mathbf{X}\} \mid y] + \boldsymbol{\beta} E\{y^2 - E(Y^2 \mid \mathbf{X}) \mid y\} + O(N^{-q}) \\
&= 2\boldsymbol{\beta} E[yE(Y \mid \mathbf{X}) - E\{YE(Y \mid \mathbf{X}) \mid \mathbf{X}\} \mid y] + \mathbf{M} E[\mathbf{X}\{y - E(Y \mid \mathbf{X})\} \mid y] \\
&= 2\boldsymbol{\beta} E[yE(Y \mid \mathbf{X}) - E\{YE(Y \mid \mathbf{X}) \mid \mathbf{X}\} \mid y] \\
&\quad + \left(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\left[2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}} Y v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \{\boldsymbol{\Lambda}\mathbf{B}_{+}(Y) - \boldsymbol{\beta} Y^2\}\{Y - E(Y \mid \mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right) \\
&\quad \times \boldsymbol{\Omega}_{11}^{-1} E[\mathbf{X}\{y - E(Y \mid \mathbf{X})\} \mid y].
\end{aligned}
$$

Multiplying $\mathbf{B}_{+}^{\mathrm{T}}(y)$ on both sides and taking expectation lead to

$$
\begin{aligned}
&-\boldsymbol{\Lambda}\boldsymbol{\Omega}_{22} + \boldsymbol{\beta} E[\{Y^2 - E(Y^2 \mid \mathbf{X})\}\mathbf{B}_{+}^{\mathrm{T}}(Y)] + O(N^{-q}) E\{\mathbf{B}_{+}^{\mathrm{T}}(Y)\} \\
&= 2\boldsymbol{\beta} E[YE(Y \mid \mathbf{X})\mathbf{B}_{+}^{\mathrm{T}}(Y) - \{E(Y \mid \mathbf{X})\}^2 \mathbf{B}_{+}^{\mathrm{T}}(Y)] \\
&\quad + \left(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\left[2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}} Y v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \{\boldsymbol{\Lambda}\mathbf{B}_{+}(Y) - \boldsymbol{\beta} Y^2\}\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right) \\
&\quad \times \boldsymbol{\Omega}_{11}^{-1} E[\mathbf{X}\{Y - E(Y \mid \mathbf{X})\}\mathbf{B}_{+}^{\mathrm{T}}(Y)]
\end{aligned}
$$

$$
\begin{aligned}
&= 2\boldsymbol{\beta}E[Y\mathrm{cov}\{Y,\mathbf{B}_+(Y)\mid\mathbf{X}\}]+E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}-E\left\{2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}}Yv(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right\}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\\
&\quad -\boldsymbol{\Lambda}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}+\boldsymbol{\beta}E[Y^2\{Y-E(Y\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}]\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\\
&= 2\boldsymbol{\beta}E[Y\mathrm{cov}\{Y,\mathbf{B}_+(Y)\mid\mathbf{X}\}]+\mathbf{A}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}-\boldsymbol{\Lambda}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12},
\end{aligned}
$$

hence

$$
\begin{aligned}
&-\boldsymbol{\Lambda}\boldsymbol{\Omega}_{22}+\boldsymbol{\Lambda}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\\
&= 2\boldsymbol{\beta}E[Y\mathrm{cov}\{Y,\mathbf{B}_+(Y)\mid\mathbf{X}\}]+\mathbf{A}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}-\boldsymbol{\beta}E[\{Y^2-E(Y^2\mid\mathbf{X})\}\mathbf{B}_+^{\mathrm{T}}(Y)]\\
&\quad +O(N^{-q})E\{\mathbf{B}_+^{\mathrm{T}}(Y)\}\\
&= \mathbf{A}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}-\mathbf{A}_2\mathbf{T}^{\mathrm{T}}+O(N^{-q})E\{\mathbf{B}_+^{\mathrm{T}}(Y)\},
\end{aligned}
$$

and

$$
\begin{aligned}
\boldsymbol{\beta}y^2+\mathbf{a}(y) &= (\mathbf{A}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}-\mathbf{A}_2\mathbf{T}^{\mathrm{T}})(\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}-\boldsymbol{\Omega}_{22})^{-}\mathbf{B}_+(y)\\
&\quad +O(N^{-q})E\{\mathbf{B}_+^{\mathrm{T}}(Y)\}(\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}-\boldsymbol{\Omega}_{22})^{-}\mathbf{B}_+(y)\\
&= \mathbf{U}\mathbf{B}_+(y)+O(N^{1/2-q}),
\end{aligned}
$$

where, for notational brevity,

$$
\mathbf{U} \equiv (\mathbf{A}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}-\mathbf{A}_2\mathbf{T}^{\mathrm{T}})(\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}-\boldsymbol{\Omega}_{22})^{-}.
$$

The above holds because

$$
\|E\{\mathbf{B}_+(Y)\}\|_2=O[N^{1/2}\|E\{\mathbf{B}_+(Y)\}\|_\infty]=O\{N^{1/2}\max_{k=1,\ldots,m+1}\|B_k(\cdot)\|_1\}=O(N^{-1/2})
$$

by Lemma B.3.4, $\|\mathbf{B}_+(y)\|_2\leqslant\|\mathbf{B}_+(y)\|_1=1$, and

$$
\boldsymbol{\zeta}_1(\boldsymbol{\Omega}_{22}-\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-}\boldsymbol{\zeta}_2=O(N\|\boldsymbol{\zeta}_1\|_2\|\boldsymbol{\zeta}_2\|_2). \tag{B.96}
$$

This is because by Condition (C6') and Lemma B.3.5, all nonzero eigenvalues of $\boldsymbol{\Omega}$ are of order either 1 or N^{-1}, which implies $\boldsymbol{\Omega}^{-}$ has all nonzero eigenvalues of order either 1 or N. Since $(\boldsymbol{\Omega}_{22}-\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-}$ is a block diagonal element of $\boldsymbol{\Omega}^{-}$, its eigenvalues are of order at most N. Then

$$
\begin{aligned}
\mathbf{M} &= \big(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I}-E\{2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}}Yv(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}-\mathbf{U}E\left[\mathbf{B}_+(Y)\{Y-E(Y\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\\
&\quad +\boldsymbol{\beta}E\left[Y^2\{Y-E(Y\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\big)\boldsymbol{\Omega}_{11}^{-1}+O(N^{1/2-q})
\end{aligned}
$$

$$
\begin{aligned}
&= \left(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\{2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}}Yv(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} - \mathbf{U}\boldsymbol{\Omega}_{21} + \boldsymbol{\beta}E\left[Y^2\{Y - E(Y \mid \mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right) \\
&\quad \times\boldsymbol{\Omega}_{11}^{-1} + O(N^{1/2-q}) \\
&= (\mathbf{A}_1 - \mathbf{U}\boldsymbol{\Omega}_{21})\boldsymbol{\Omega}_{11}^{-1} + O(N^{1/2-q}).
\end{aligned}
$$

Hence

$$
\boldsymbol{\beta}y^2 + \mathbf{a}(y) - E\{\boldsymbol{\beta}Y^2 + \mathbf{a}(Y) \mid \mathbf{x}\} = \mathbf{U}[\mathbf{B}_+(y) - E\{\mathbf{B}_+(Y) \mid \mathbf{x}\}] + O(N^{1/2-q}),
$$

and

$$
\begin{aligned}
E\{\mathrm{var}(\mathbf{M}\mathbf{X}Y \mid \mathbf{X})\} &= \mathbf{M}\boldsymbol{\Omega}_{11}\mathbf{M}^{\mathrm{T}}, \\
E[\mathrm{var}\{\boldsymbol{\beta}Y^2 + \mathbf{a}(Y) \mid \mathbf{X}\}] &= \mathbf{U}\boldsymbol{\Omega}_{22}\mathbf{U}^{\mathrm{T}} + O(N^{1/2-q}), \\
E[\mathrm{cov}\{\mathbf{M}\mathbf{X}Y, \boldsymbol{\beta}Y^2 + \mathbf{a}(Y) \mid \mathbf{X}\}] &= \mathbf{M}\boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}} + O(N^{1/2-q}).
\end{aligned}
$$

Thus, noting that $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$,

$$
\begin{aligned}
& E\left([\boldsymbol{\beta}Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y - E\{\boldsymbol{\beta}Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y \mid \mathbf{X}\}]^{\otimes 2}\right) \\
&= \mathbf{M}\boldsymbol{\Omega}_{11}\mathbf{M}^{\mathrm{T}} + \mathbf{U}\boldsymbol{\Omega}_{22}\mathbf{U}^{\mathrm{T}} + \mathbf{M}\boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}} + \mathbf{U}\boldsymbol{\Omega}_{21}\mathbf{M}^{\mathrm{T}} + O(N^{1/2-q}) \\
&= (\mathbf{A}_1 - \mathbf{U}\boldsymbol{\Omega}_{21})\boldsymbol{\Omega}_{11}^{-1}(\mathbf{A}_1^{\mathrm{T}} - \boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}}) + \mathbf{U}\boldsymbol{\Omega}_{22}\mathbf{U}^{\mathrm{T}} + (\mathbf{A}_1 - \mathbf{U}\boldsymbol{\Omega}_{21})\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}} \\
&\quad +\mathbf{U}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}(\mathbf{A}_1^{\mathrm{T}} - \boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}}) + O(N^{1/2-q}) \\
&= \mathbf{A}_1\boldsymbol{\Omega}_{11}^{-1}\mathbf{A}_1^{\mathrm{T}} + \mathbf{U}(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})\mathbf{U}^{\mathrm{T}} + O(N^{1/2-q}) \\
&= \mathbf{A}_1\boldsymbol{\Omega}_{11}^{-1}\mathbf{A}_1^{\mathrm{T}} + (\mathbf{A}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} - \mathbf{A}_2\mathbf{T}^{\mathrm{T}})(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1}(\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\mathbf{A}_1^{\mathrm{T}} - \mathbf{T}\mathbf{A}_2^{\mathrm{T}}) \\
&\quad +O(N^{1/2-q}) \\
&= \mathbf{A}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}}) \\
&\quad \times \begin{pmatrix} \boldsymbol{\Omega}_{11}^{-1} + \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & -\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1} \\ -(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & (\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1} \end{pmatrix} \\
&\quad \times\mathrm{diag}(\mathbf{I}_p, \mathbf{T})\mathbf{A}^{\mathrm{T}} \\
&\quad +O(N^{1/2-q}) \\
&= \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}} + o(1),
\end{aligned}
$$

where the last equality holds by Lemma B.3.3. □

B.12 Proof of Theorem 3.3.5

By (B.12)

$$\begin{aligned} E(\boldsymbol{\phi}_{\text{eff}}^{\otimes 2}) &= E\left([\boldsymbol{\beta} q'(\nu) - \boldsymbol{\beta} E\{q'(\nu)\}]^{\otimes 2}\right) \\ &\quad + E\left([\mathbf{M}_1 \mathbf{X} Y + \mathbf{a}(Y) - E\{\mathbf{M}_1 \mathbf{X} Y + \mathbf{a}(Y) \mid \mathbf{X}\}]^{\otimes 2}\right), \end{aligned}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}_\tau} = \mathbf{C}\boldsymbol{\Sigma}^{-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\beta}^{\otimes 2}\mathrm{var}\{q'(\nu)\}$$

in Theorem 3.3.2, we only need to show

$$E\left([\mathbf{M}_1 \mathbf{X} Y + \mathbf{a}(Y) - E\{\mathbf{M}_1 \mathbf{X} Y + \mathbf{a}(Y) \mid \mathbf{X}\}]^{\otimes 2}\right) - \mathbf{C}\boldsymbol{\Sigma}^{-1}\mathbf{C}^{\mathrm{T}} \to \mathbf{0}.$$

Now we have $\mathbf{a}(y) = \boldsymbol{\Lambda}\mathbf{B}_+(y) + O(N^{-q})$ where $\boldsymbol{\Lambda} \in \mathcal{R}^{p\times(m+1)}$. Then (B.11) and (B.13) imply

$$\begin{aligned} & -\boldsymbol{\Lambda} E[\mathbf{B}_+(y) - E\{\mathbf{B}_+(Y) \mid \mathbf{X}\} \mid y] + O(N^{-q}) \\ = \; & -\boldsymbol{\beta} E\{r(y,\nu) \mid y\} + \mathbf{M}_1 E[\mathbf{X}\{y - E(Y \mid \mathbf{X})\} \mid y] \\ = \; & -\boldsymbol{\beta} E\{r(y,\nu) \mid y\} \\ & + \left(E\{q'(\nu)\}\mathbf{I} + \boldsymbol{\beta} E\{\mathbf{X}^{\mathrm{T}} q''(\nu)\} - E\left[\{\boldsymbol{\Lambda}\mathbf{B}_+(Y) + O(N^{-q})\}\mathbf{X}^{\mathrm{T}}\{Y - E(Y \mid \mathbf{X})\}\right]\right) \\ & \times \boldsymbol{\Omega}_{11}^{-1} E[\mathbf{X}\{y - E(Y \mid \mathbf{X})\} \mid y]. \end{aligned}$$

Multiplying $\mathbf{B}_+^{\mathrm{T}}(y)$ on both sides above and taking expectation, incorporating (B.7), we get

$$\begin{aligned} & -\boldsymbol{\Lambda}\boldsymbol{\Omega}_{22} + O(N^{-q}) E\{\mathbf{B}_+^{\mathrm{T}}(Y)\} \\ = \; & -\boldsymbol{\beta} E\{r(Y,\nu)\mathbf{B}_+^{\mathrm{T}}(Y)\} \\ & + \left(E\{q'(\nu)\}\mathbf{I} + \boldsymbol{\beta} E\{\mathbf{X}^{\mathrm{T}} q''(\nu)\} - E\left[\{\boldsymbol{\Lambda}\mathbf{B}_+(Y) + O(N^{-q})\}\mathbf{X}^{\mathrm{T}}\{Y - E(Y \mid \mathbf{X})\}\right]\right) \\ & \times \boldsymbol{\Omega}_{11}^{-1} E[\mathbf{X}\{Y - E(Y \mid \mathbf{X})\}\mathbf{B}_+^{\mathrm{T}}(Y)] \\ = \; & -\boldsymbol{\beta} E\left(\frac{E\{\epsilon Y \mathbf{B}_+^{\mathrm{T}}(Y) | \mathbf{X}\}}{f\{q(\nu), \nu\}} - q'(\nu)\mathbf{B}_+^{\mathrm{T}}\{q(\nu)\} \right. \\ & \left. - \frac{E\{\epsilon \mathbf{B}_+^{\mathrm{T}}(Y) | \mathbf{X}\}}{f\{q(\nu), \nu\}}[q(\nu) + q'(\nu)\nu + q'(\nu)c'\{q(\nu)\}] \right) \end{aligned}$$

$$
\begin{aligned}
& +\mathbf{C}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} - \boldsymbol{\Lambda}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} + O(N^{-q})\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \\
= \quad & -\boldsymbol{\Lambda}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} + \mathbf{C}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} - \mathbf{C}_2\mathbf{T}^{\mathrm{T}} + O(N^{-q})\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12},
\end{aligned}
$$

hence

$$
\boldsymbol{\Lambda}(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}) = -\mathbf{C}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} + \mathbf{C}_2\mathbf{T}^{\mathrm{T}} + O(N^{-q-1/2}),
$$

since

$$
\|E\{\mathbf{B}_+(Y)\}\|_2 = O[N^{1/2}\|E\{\mathbf{B}_+(Y)\}\|_\infty] = O\{N^{1/2}\max_{k=1,\ldots,m+1}\|B_k(\cdot)\|_1\} = O(N^{-1/2})
$$

by Lemma B.3.4, and $\|\boldsymbol{\Omega}_{21}\boldsymbol{\beta}\|_2 = O(N^{-1/2}\|\boldsymbol{\beta}\|_2)$ by (B.28). Then by (B.96) and $\|\mathbf{B}_+(y)\|_2 \leqslant \|\mathbf{B}_+(y)\|_1 = 1$,

$$
\mathbf{a}(y) = \mathbf{U}\mathbf{B}_+(y) + O(N^{1/2-q}),
$$

where for notational brevity,

$$
\mathbf{U} \equiv (-\mathbf{C}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} + \mathbf{C}_2\mathbf{T}^{\mathrm{T}})(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^-.
$$

Then by (B.13),

$$
\begin{aligned}
& \mathbf{M}_1 \\
= \quad & (E\{q'(\nu)\}\mathbf{I} + \boldsymbol{\beta}E\{\mathbf{X}^{\mathrm{T}}q''(\nu)\} - \mathbf{U}E\left[\mathbf{B}_+(Y)\mathbf{X}^{\mathrm{T}}\{Y - E(Y \mid \mathbf{X})\}\right])\boldsymbol{\Omega}_{11}^{-1} + O(N^{1/2-q}) \\
= \quad & (\mathbf{C}_1 - \mathbf{U}\boldsymbol{\Omega}_{21})\boldsymbol{\Omega}_{11}^{-1} + O(N^{1/2-q}).
\end{aligned}
$$

Hence

$$
\mathbf{a}(y) - E\{\mathbf{a}(Y) \mid \mathbf{x}\} = \mathbf{U}[\mathbf{B}_+(y) - E\{\mathbf{B}_+(Y) \mid \mathbf{x}\}] + O(N^{1/2-q}),
$$

and

$$
\begin{aligned}
E\{\mathrm{var}(\mathbf{M}_1\mathbf{X}Y \mid \mathbf{X})\} &= \mathbf{M}_1\boldsymbol{\Omega}_{11}\mathbf{M}_1^{\mathrm{T}}, \\
E[\mathrm{cov}\{\mathbf{M}\mathbf{X}Y, \mathbf{a}(Y) \mid \mathbf{X}\}] &= \mathbf{M}_1\boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}} + O(N^{1/2-q}), \\
E[\mathrm{var}\{\mathbf{a}(Y) \mid \mathbf{X}\}] &= \mathbf{U}\boldsymbol{\Omega}_{22}\mathbf{U}^{\mathrm{T}} + O(N^{1/2-q}).
\end{aligned}
$$

Thus noting that $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2]$, we obtain

$$
\begin{aligned}
& E\left([\mathbf{M}_1\mathbf{X}Y + \mathbf{a}(Y) - E\{\mathbf{M}_1\mathbf{X}Y + \mathbf{a}(Y) \mid \mathbf{X}\}]^{\otimes 2}\right) \\
= \;& \mathbf{M}_1\boldsymbol{\Omega}_{11}\mathbf{M}_1^{\mathrm{T}} + \mathbf{M}_1\boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}} + \mathbf{U}\boldsymbol{\Omega}_{21}\mathbf{M}_1^{\mathrm{T}} + \mathbf{U}\boldsymbol{\Omega}_{22}\mathbf{U}^{\mathrm{T}} + O(N^{1/2-q}) \\
= \;& (\mathbf{C}_1 - \mathbf{U}\boldsymbol{\Omega}_{21})\boldsymbol{\Omega}_{11}^{-1}(\mathbf{C}_1^{\mathrm{T}} - \boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}}) + (\mathbf{C}_1 - \mathbf{U}\boldsymbol{\Omega}_{21})\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}} \\
& + \mathbf{U}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}(\mathbf{C}_1^{\mathrm{T}} - \boldsymbol{\Omega}_{12}\mathbf{U}^{\mathrm{T}}) + \mathbf{U}\boldsymbol{\Omega}_{22}\mathbf{U}^{\mathrm{T}} + O(N^{1/2-q}) \\
= \;& \mathbf{C}_1\boldsymbol{\Omega}_{11}^{-1}\mathbf{C}_1^{\mathrm{T}} + \mathbf{U}(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})\mathbf{U}^{\mathrm{T}} + O(N^{1/2-q}) \\
= \;& \mathbf{C}_1\boldsymbol{\Omega}_{11}^{-1}\mathbf{C}_1^{\mathrm{T}} + (-\mathbf{C}_1\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} + \mathbf{C}_2\mathbf{T}^{\mathrm{T}})(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1}(-\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\mathbf{C}_1^{\mathrm{T}} + \mathbf{T}\mathbf{C}_2^{\mathrm{T}}) \\
& + O(N^{1/2-q}) \\
= \;& \mathbf{C}\mathrm{diag}(\mathbf{I}_p, \mathbf{T}^{\mathrm{T}}) \\
& \times \begin{pmatrix} \boldsymbol{\Omega}_{11}^{-1} + \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & -\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1} \\ -(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & (\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1} \end{pmatrix} \\
& \times \mathrm{diag}(\mathbf{I}_p, \mathbf{T})\mathbf{C}^{\mathrm{T}} \\
= \;& \mathbf{C}\boldsymbol{\Sigma}^{-1}\mathbf{C}^{\mathrm{T}} + o(1),
\end{aligned}
$$

where the last equality holds by Lemma B.3.3. □

B.13 Additional lemma

We now introduce a lemma for the analysis of the discrete response case.

Lemma B.13.1. *Let* $\boldsymbol{\theta}^* \equiv (\boldsymbol{\beta}^{*\mathrm{T}}, \boldsymbol{\gamma}^{*\mathrm{T}})^{\mathrm{T}}$ *and* $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 = o(1)$. *Under Conditions (D1)-(D2), uniformly with respect to* $\mathbf{x}$,

(i) $\|\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\|_2 = O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 + \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2)$,

(ii) $\|\mathbf{p}_1(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}_1(\mathbf{x}, \boldsymbol{\theta}_0)\|_2 = O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 + \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2)$,

(iii) $\|\mathbf{p}_2(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}_2(\mathbf{x}, \boldsymbol{\theta}_0)\|_2 = O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 + \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2)$,

(iv) $\|var\{\mathbf{D}(Y) \mid \mathbf{x}, \boldsymbol{\theta}^*\} - var\{\mathbf{D}(Y) \mid \mathbf{x}, \boldsymbol{\theta}_0\}\|_2 = O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 + \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2)$.

If $\boldsymbol{\theta}^*$ *satisfies* $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 = o_p(1)$, *the above results hold in probability.*

Proof. For some $\widetilde{\boldsymbol{\theta}}$ on the line connecting $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_0$, we have

$$
\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0) \;=\; \frac{\partial \mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \frac{\partial \mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}(\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0).
$$

First, by Conditions (D1) and (D2),

$$\begin{aligned}
\left\| \frac{\partial \mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}^{\mathrm{T}}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) \right\|_2 &= \left\| \left\{ \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) - E(Y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) \mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\} \mathbf{x}^{\mathrm{T}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) \right\|_2 \\
&\leqslant \left[\sum_{y=1}^{M} \left\{ y - E(Y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\}^2 \mathrm{pr}(Y = y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}})^2 \right]^{1/2} O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2) \\
&\leqslant \sum_{y=1}^{M} \left| y - E(Y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right| \mathrm{pr}(Y = y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2) \\
&\leqslant 2E(Y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2) \\
&\leqslant 2\{E(Y \mid \mathbf{x}, \boldsymbol{\theta}_0) + O(\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2)\} O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2) \\
&= O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2).
\end{aligned}$$

The last inequality holds since $E(Y \mid \mathbf{x}, \boldsymbol{\theta}_0)$ is bounded by Condition (D1), and $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = o(1)$ by the definition of $\widetilde{\boldsymbol{\theta}}$. In addition, for a vector $\mathbf{a}$, let $\mathrm{diag}(\mathbf{a})$ be the diagonal matrix with entries equal to the elements of $\mathbf{a}$, then

$$\begin{aligned}
\left\| \frac{\partial \mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} (\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0) \right\|_2 &= \left\| \left[\mathrm{diag} \left\{ \mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\} - \mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \mathbf{p}^{\mathrm{T}}(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right] (\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0) \right\|_2 \\
&\leqslant \left[\|\mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})\|_\infty + \|\mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})\|_2^2 \right] \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2 \\
&= O(\|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2)
\end{aligned}$$

because $\|\mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})\|_\infty \leqslant \|\mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})\|_2 \leqslant 1$. Hence, we get

$$\|\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\|_2 = O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 + \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2).$$

Similarly for some $\widetilde{\boldsymbol{\theta}}$ on the line connecting $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_0$,

$$\begin{aligned}
\left\| \frac{\partial \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}^{\mathrm{T}}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) \right\|_2 &= \left\| \left\{ \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) - E(Y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\} \mathbf{x}^{\mathrm{T}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) \right\|_2 \\
&\leqslant \sum_{y=1}^{M} \left| y - E(Y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right| y \mathrm{pr}(Y = y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2) \\
&\leqslant \left[E(Y^2 \mid \mathbf{x}, \boldsymbol{\theta}_0) + \{E(Y \mid \mathbf{x}, \boldsymbol{\theta}_0)\}^2 + O(\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2) \right] \\
&\quad \times O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2) \\
&= O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2)
\end{aligned}$$

by Conditions (D1) and (D2). In addition,

$$\begin{aligned}
\left\| \frac{\partial \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} (\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0) \right\|_2 &= \left\| \left[\operatorname{diag} \left\{ \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\} - \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \mathbf{p}^{\mathrm{T}}(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right] (\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0) \right\|_2 \\
&\leqslant \left[\left\| \operatorname{diag} \left\{ \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\} \right\|_2 + \left\| \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \mathbf{p}^{\mathrm{T}}(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\|_2 \right] \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2 \\
&\leqslant \left[\|\mathbf{p}_1(\mathbf{x}, \widetilde{\theta})\|_\infty + \|\mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}})\|_2 \|\mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})\|_2 \right] \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2 \\
&\leqslant O(\|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2),
\end{aligned}$$

because $\|\mathbf{p}_1(\mathbf{x}, \widetilde{\theta})\|_\infty \leqslant \|\mathbf{p}_1(\mathbf{x}, \widetilde{\theta})\|_2 \leqslant \|\mathbf{p}_1(\mathbf{x}, \widetilde{\theta})\|_1 = E(Y \mid \mathbf{x}, \widetilde{\theta}) = E(Y \mid \mathbf{x}) + o(1)$ and $E(Y \mid \mathbf{x})$ is bounded by Condition (D1). Therefore,

$$\begin{aligned}
\|\mathbf{p}_1(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}_1(\mathbf{x}, \boldsymbol{\theta}_0)\|_2 &= \left\| \frac{\partial \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}^{\mathrm{T}}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \frac{\partial \mathbf{p}_1(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} (\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0) \right\|_2 \\
&= O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 + \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2).
\end{aligned}$$

Also, by Conditions (D1) and (D2),

$$\begin{aligned}
\left\| \frac{\partial \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}^{\mathrm{T}}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) \right\|_2 &= \left\| \left\{ \mathbf{p}_3(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) - E(Y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\} \mathbf{x}^{\mathrm{T}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) \right\|_2 \\
&\leqslant \sum_{y=1}^{M} \left| y - E(Y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right| y^2 \mathrm{pr}(Y = y \mid \mathbf{x}, \widetilde{\boldsymbol{\theta}}) O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2) \\
&\leqslant \left[E(Y^3 \mid \mathbf{x}) + E(Y \mid \mathbf{x}) E(Y^2 \mid \mathbf{x}) + O(\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2) \right] \\
&\quad \times O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2) \\
&= O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2),
\end{aligned}$$

and

$$\begin{aligned}
\left\| \frac{\partial \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} (\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0) \right\|_2 &= \left\| \left[\operatorname{diag} \left\{ \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\} - \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \mathbf{p}^{\mathrm{T}}(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right] (\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0) \right\|_2 \\
&\leqslant \left[\left\| \operatorname{diag} \left\{ \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\} \right\|_2 + \left\| \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \mathbf{p}^{\mathrm{T}}(\mathbf{x}, \widetilde{\boldsymbol{\theta}}) \right\|_2 \right] \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2 \\
&\leqslant \left[\|\mathbf{p}_2(\mathbf{x}, \widetilde{\theta})\|_\infty + \|\mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}})\|_2 \|\mathbf{p}(\mathbf{x}, \widetilde{\boldsymbol{\theta}})\|_2 \right] \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2 \\
&\leqslant O(\|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2).
\end{aligned}$$

The last inequality holds because $\|\mathbf{p}_2(\mathbf{x}, \widetilde{\theta})\|_\infty \leqslant \|\mathbf{p}_2(\mathbf{x}, \widetilde{\theta})\|_2 \leqslant \|\mathbf{p}_2(\mathbf{x}, \widetilde{\theta})\|_1 = E(Y^2 \mid$

$\mathbf{x}, \widetilde{\theta}) = E(Y^2 \mid \mathbf{x}) + o(1)$ and $E(Y^2 \mid \mathbf{x})$ is bounded by Condition (D1). Then we get

$$\begin{aligned}\|\mathbf{p}_2(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}_2(\mathbf{x}, \boldsymbol{\theta}_0)\|_2 &= \left\| \frac{\partial \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \frac{\partial \mathbf{p}_2(\mathbf{x}, \widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}(\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0) \right\|_2 \\ &= O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 + \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2).\end{aligned}$$

Now,

$$\begin{aligned}& \|\mathrm{var}\{\mathbf{D}(Y) \mid \mathbf{x}, \boldsymbol{\theta}^*\} - \mathrm{var}\{\mathbf{D}(Y) \mid \mathbf{x}, \boldsymbol{\theta}_0\}\|_2 \\ \leqslant\ & \|\mathrm{diag}\,\{\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\}\|_2 + \left\|\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*)\mathbf{p}^{\mathrm{T}}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\mathbf{p}^{\mathrm{T}}(\mathbf{x}, \boldsymbol{\theta}_0)\right\|_2 \\ =\ & \|\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\|_\infty \\ & + \left\|\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*)\{\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\}^{\mathrm{T}} + \{\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\}\mathbf{p}^{\mathrm{T}}(\mathbf{x}, \boldsymbol{\theta}_0)\right\|_2 \\ \leqslant\ & \|\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\|_2 \\ & + \left|\{\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\}^{\mathrm{T}}\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*)\right| + \left|\mathbf{p}^{\mathrm{T}}(\mathbf{x}, \boldsymbol{\theta}_0)\{\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\}\right| \\ \leqslant\ & 3\,\|\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}^*) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta}_0)\|_2 \\ =\ & O(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 + \|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2).\end{aligned}$$

Since the support of $\mathbf{x}$ is compact by Condition (D1), the about results hold uniformly with respect to $\mathbf{x}$. □

B.14 Proof of Proposition 3.3.4

First, we will show that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 = O_p(n^{-1/2}M^{1/2})$. Note that the Hessian of $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$, i.e.

$$\begin{aligned}& \begin{bmatrix} \partial^2 l(\boldsymbol{\beta}, \boldsymbol{\gamma})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}} & \partial^2 l(\boldsymbol{\beta}, \boldsymbol{\gamma})/\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^{\mathrm{T}} \\ \partial^2 l(\boldsymbol{\beta}, \boldsymbol{\gamma})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}^{\mathrm{T}} & \partial^2 l(\boldsymbol{\beta}, \boldsymbol{\gamma})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}} \end{bmatrix} \\ =\ & -\sum_{i=1}^{n} \begin{bmatrix} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathrm{var}(Y \mid \mathbf{x}_i, \boldsymbol{\theta}) & \mathbf{x}_i\mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{x}_i, \boldsymbol{\theta}\} \\ [\mathbf{x}_i\mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{x}_i, \boldsymbol{\theta}\}]^{\mathrm{T}} & \mathrm{var}\{\mathbf{D}(Y) \mid \mathbf{x}_i, \boldsymbol{\theta}\} \end{bmatrix} \\ =\ & -\sum_{i=1}^{n} \mathrm{cov}\left\{ \begin{bmatrix} \mathbf{x}_i Y \\ \mathbf{D}(Y) \end{bmatrix}, \begin{bmatrix} \mathbf{x}_i Y \\ \mathbf{D}(Y) \end{bmatrix} \mid \mathbf{x}_i, \boldsymbol{\theta} \right\}\end{aligned}$$

is negative definite for any $\boldsymbol{\theta}$, which implies that a local maximizer is the global maximizer. Hence similarly to the proof of Proposition 3.3.1, it suffices to show for any $\epsilon > 0$

there exists constants $C_{\boldsymbol{\beta}}, C_{\boldsymbol{\gamma}} > 0$ such that

$$\text{pr}\left\{ l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) > \sup_{\|\mathbf{v}_{\boldsymbol{\beta}}\|_2 = C_{\boldsymbol{\beta}}, \|\mathbf{v}_{\boldsymbol{\gamma}}\|_2 = C_{\boldsymbol{\gamma}}} l(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{v}_{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0 + n^{-1/2}M^{1/2}\mathbf{v}_{\boldsymbol{\gamma}}) \right\} \geqslant 1 - 5\epsilon \quad \text{(B.97)}$$

for a sufficiently large n. Now, by the Taylor expansion,

$$\begin{aligned} & l(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{v}_{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0 + n^{-1/2}M^{1/2}\mathbf{v}_{\boldsymbol{\gamma}}) - l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) \qquad \text{(B.98)} \\ = \;& n^{-1/2}\frac{\partial l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\mathbf{v}_{\boldsymbol{\beta}} + n^{-1/2}M^{1/2}\frac{\partial l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{v}_{\boldsymbol{\gamma}} \\ & + \frac{1}{2}n^{-1}\mathbf{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}\mathbf{v}_{\boldsymbol{\beta}} + \frac{1}{2}n^{-1}M\mathbf{v}_{\boldsymbol{\gamma}}^{\mathrm{T}}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{v}_{\boldsymbol{\gamma}} + n^{-1}M^{1/2}\mathbf{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{v}_{\boldsymbol{\gamma}} \end{aligned}$$

where $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0 + \alpha_1 n^{-1/2}\mathbf{v}_{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0 + \alpha_2 n^{-1/2}M^{1/2}\mathbf{v}_{\boldsymbol{\gamma}}$ for some $\mathbf{v}_{\boldsymbol{\beta}}, \mathbf{v}_{\boldsymbol{\gamma}}$ such that $\|\mathbf{v}_{\boldsymbol{\beta}}\|_2 = C_{\boldsymbol{\beta}}, \|\mathbf{v}_{\boldsymbol{\gamma}}\|_2 = C_{\boldsymbol{\gamma}}$ and some $\alpha_1 \in (0,1), \alpha_2 \in (0,1)$. We first have

$$\left\| \frac{\partial l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\beta}} \right\|_2 = \left\| \sum_{i=1}^{n} \mathbf{x}_i \{ y_i - E(Y \mid \mathbf{x}_i) \} \right\|_2 \asymp_p n^{1/2}$$

by Conditions (D1) and (D5). Therefore, for any $\epsilon > 0$ there exists a constant $0 < C_1 < \infty$ such that

$$\text{pr}\left\{ \left\| \frac{\partial l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\beta}} \right\|_2 \leqslant C_1 n^{1/2} \right\} \geqslant 1 - \epsilon. \quad \text{(B.99)}$$

Also,

$$\left\| \frac{\partial l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} \right\|_2 = \left\| \sum_{i=1}^{n} \{ \mathbf{D}(y_i) - E\{\mathbf{D}(Y) \mid \mathbf{x}_i\} \} \right\|_2 \asymp_p n^{1/2}M^{-1/2},$$

because

$$\begin{aligned} \left\| E\left\{ \left(\sum_{i=1}^{n} [\mathbf{D}(Y_i) - E\{\mathbf{D}(Y) \mid \mathbf{X}_i\}] \right)^{\otimes 2} \right\} \right\|_2 &= \left\| nE\left([\mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\}]^{\otimes 2} \right) \right\|_2 \\ &= n \left\| E[\text{var}\{\mathbf{D}(Y) \mid \mathbf{X}\}] \right\|_2 \\ &\asymp nM^{-1} \end{aligned}$$

by Remark 3.3.1. Hence, for any $\epsilon > 0$ there exists a constant $0 < C_2 < \infty$ such that

$$\text{pr}\left\{\left\|\frac{\partial l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}\right\|_2 \leqslant C_2 n^{1/2} M^{-1/2}\right\} \geqslant 1 - \epsilon. \tag{B.100}$$

In addition, noting that $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 = O(n^{-1/2}M^{1/2}) = o(1)$ by Condition (D3), we get

$$\begin{aligned}\left\|-n^{-1}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}} - \boldsymbol{\Sigma}_{11}\right\|_2 &\leqslant \left\|n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\{\text{var}(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*) - \text{var}(Y \mid \mathbf{x}_i)\}\right\|_2 \\ &\quad + \left\|n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\text{var}(Y \mid \mathbf{x}_i) - E\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\text{var}(Y \mid \mathbf{X})\}\right\|_2 \\ &= o_p(1)\end{aligned} \tag{B.101}$$

by Conditions (D1) and (D2). Since all eigenvalues of $\boldsymbol{\Sigma}_{11}$ are of constant order by Conditions (D1) and (D5), the above implies there exists a constant $0 < C_3 < \infty$ such that

$$\text{pr}\left\{\mathbf{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}\mathbf{v}_{\boldsymbol{\beta}} \leqslant -C_3 C_{\boldsymbol{\beta}}^2 n\right\} \geqslant 1 - \epsilon. \tag{B.102}$$

Further, since $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 = O(n^{-1/2})$ and $\|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2 = O(n^{-1/2}M^{1/2})$,

$$\begin{aligned}\left\|-n^{-1}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^{\mathrm{T}}} - \boldsymbol{\Sigma}_{22}\right\|_2 &\leqslant \left\|n^{-1}\sum_{i=1}^n [\text{var}\{\mathbf{D}(Y) \mid \mathbf{x}_i, \boldsymbol{\theta}^*\} - \text{var}\{\mathbf{D}(Y) \mid \mathbf{x}_i\}]\right\|_2 \\ &\quad + \left\|n^{-1}\sum_{i=1}^n \text{var}\{\mathbf{D}(Y) \mid \mathbf{x}_i\} - E[\text{var}\{\mathbf{D}(Y) \mid \mathbf{X}\}]\right\|_2 \\ &= O(n^{-1/2}M^{1/2}) + O_p(n^{-1/2}M^{-1}) \\ &= O_p(n^{-1/2}M^{1/2}) \\ &= o_p(M^{-1})\end{aligned} \tag{B.103}$$

by Lemma B.13.1, Remark 3.3.1, and $n^{-1/2}M^{1/2} = o(M^{-1})$ by Condition (D3). Together with Remark 3.3.1, this implies

$$\text{pr}\left\{\mathbf{v}_{\boldsymbol{\gamma}}^{\mathrm{T}}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{v}_{\boldsymbol{\gamma}} \leqslant -C_4 C_{\boldsymbol{\gamma}}^2 n M^{-1}\right\} \geqslant 1 - \epsilon \tag{B.104}$$

for some constant $0 < C_4 < \infty$. Now, we have $\|\boldsymbol{\Sigma}_{12}\|_2 = O(M^{-1/2})$, because noting that $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ are positive definite by Condition (D5), for any $\mathbf{u} \in \mathbb{R}^M$ such

that $\|\mathbf{u}\|_2 = 1$ we have

$$c\|\boldsymbol{\Sigma}_{12}\mathbf{u}\|_2^2 \leqslant \|\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\mathbf{u}\|_2^2 = \mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{u} < \mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{22}\mathbf{u} \leqslant \|\boldsymbol{\Sigma}_{22}\|_2 \asymp M^{-1} \quad \text{(B.105)}$$

by Remark 3.3.1 for some constant $c > 0$. Also, since $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 = O(n^{-1/2})$ and $\|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2 = O(n^{-1/2}M^{1/2})$,

$$\begin{aligned}
& \left\| -n^{-1}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^{\mathrm{T}}} - \boldsymbol{\Sigma}_{12} \right\|_2 \\
\leqslant & \left\| n^{-1}\sum_{i=1}^n \mathbf{x}_i[\mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{x}_i, \boldsymbol{\theta}^*\} - \mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{x}_i\}] \right\|_2 \\
& + \left\| n^{-1}\sum_{i=1}^n \mathbf{x}_i \mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{x}_i\} - E[\mathbf{X}\mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{X}\}] \right\|_2 \\
\leqslant & \left\| n^{-1}\sum_{i=1}^n \mathbf{x}_i\{\mathbf{p}_1^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\theta}^*) - \mathbf{p}_1^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\theta}_0)\} \right\|_2 \\
& + \left\| n^{-1}\sum_{i=1}^n \mathbf{x}_i\{E(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*)\mathbf{p}^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\theta}^*) - E(Y \mid \mathbf{x}_i, \boldsymbol{\theta}_0)\mathbf{p}^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\theta}_0)\} \right\|_2 \\
& + \left\| n^{-1}\sum_{i=1}^n \mathbf{x}_i \mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{x}_i\} - E[\mathbf{X}\mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{X}\}] \right\|_2 \\
= & O(n^{-1/2}M^{1/2}) + O_p(n^{-1/2}M^{-1/2}) \\
= & o_p(M^{-1/2}) \qquad \text{(B.106)}
\end{aligned}$$

by Lemma B.13.1, Condition (D1), and $n^{-1/2}M^{1/2} = o(M^{-1/2})$ by Condition (D3). Thus we have

$$\mathrm{pr}\left\{ \mathbf{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{v}_{\boldsymbol{\gamma}} \leqslant C_5 C_{\boldsymbol{\beta}} C_{\boldsymbol{\gamma}} n M^{-1/2} \right\} \geqslant 1 - \epsilon \qquad \text{(B.107)}$$

for some constant $0 < C_5 < \infty$. Combining (B.98), (B.99), (B.100), (B.102), (B.104), and (B.107), with probability at least $1 - 5\epsilon$,

$$\begin{aligned}
& l(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{v}_{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0 + n^{-1/2}M^{1/2}\mathbf{v}_{\boldsymbol{\gamma}}) - l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) \\
\leqslant & C_1 C_{\boldsymbol{\beta}} + C_2 C_{\boldsymbol{\gamma}} - \frac{C_3}{2}C_{\boldsymbol{\beta}}^2 - \frac{C_4}{2}C_{\boldsymbol{\gamma}}^2 + C_5 C_{\boldsymbol{\beta}} C_{\boldsymbol{\gamma}} \\
= & C_{\boldsymbol{\beta}}\left(C_1 - \frac{C_3}{2}C_{\boldsymbol{\beta}} + C_5 C_{\boldsymbol{\gamma}} \right) + C_{\boldsymbol{\gamma}}\left(C_2 - \frac{C_4}{2}C_{\boldsymbol{\gamma}} \right) \\
< & 0
\end{aligned}$$

when $C_{\boldsymbol{\beta}} > 2(C_1 + C_5 C_{\boldsymbol{\gamma}})/C_3$ and $C_{\boldsymbol{\gamma}} > 2C_2/C_4$, and this proves (B.97). Therefore, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 = O_p(n^{-1/2}M^{1/2})$.

Now, we analyze the asymptotic behavior of $(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$. Since $(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$ is the maximizer of $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$, letting $(\boldsymbol{\beta}^{*\mathrm{T}}, \boldsymbol{\gamma}^{*\mathrm{T}})^{\mathrm{T}}$ be on the line connecting $(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$ and $(\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\gamma}_0^{\mathrm{T}})^{\mathrm{T}}$,

$$\begin{aligned}
\mathbf{0} &= n^{-1}\frac{\partial l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\theta}} - \left\{-n^{-1}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\}\begin{bmatrix}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\end{bmatrix} \\
&= n^{-1}\sum_{i=1}^{n}\begin{bmatrix}\mathbf{x}_i\{y_i - E(Y \mid \mathbf{x}_i)\} \\ \mathbf{D}(y_i) - E\{\mathbf{D}(Y) \mid \mathbf{x}_i\}\end{bmatrix} - \left(\boldsymbol{\Sigma} + \begin{bmatrix}\mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{12}^{\mathrm{T}} & \mathbf{R}_{22}\end{bmatrix}\right)\begin{bmatrix}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\end{bmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{R}_{11} &\equiv -n^{-1}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}} - \boldsymbol{\Sigma}_{11}, \\
\mathbf{R}_{12} &\equiv -n^{-1}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\gamma}^{\mathrm{T}}} - \boldsymbol{\Sigma}_{12}, \\
\mathbf{R}_{22} &\equiv -n^{-1}\frac{\partial^2 l(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^{\mathrm{T}}} - \boldsymbol{\Sigma}_{22}.
\end{aligned}$$

Since $\boldsymbol{\Sigma}$ is invertible by Condition (D5), we have

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\end{bmatrix} = \boldsymbol{\Sigma}^{-1}n^{-1}\sum_{i=1}^{n}\begin{bmatrix}\mathbf{x}_i\{y_i - E(Y \mid \mathbf{x}_i)\} \\ \mathbf{D}(y_i) - E\{\mathbf{D}(Y) \mid \mathbf{x}_i\}\end{bmatrix} - \boldsymbol{\Sigma}^{-1}\begin{bmatrix}\mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{12}^{\mathrm{T}} & \mathbf{R}_{22}\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\end{bmatrix}.$$

Now, note that

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\beta}} & -\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{22}^{-1} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\end{bmatrix} = \begin{bmatrix}O(1) & O(M^{1/2}) \\ O(M^{1/2}) & O(M)\end{bmatrix} \quad \text{(B.108)}$$

in terms of the 2-norms of the block matrices. This is because $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ is positive definite by Condition (D5), $\|\boldsymbol{\Sigma}_{12}\|_2 = O(M^{-1/2})$ by (B.105), and $\|\boldsymbol{\Sigma}_{22}^{-1}\|_2 \asymp M$ by Remark 3.3.1. In addition, using the fact that $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ and $\|\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0\|_2 = O_p(n^{-1/2}M^{1/2})$, similar arguments to (B.101), (B.103), and (B.106) leads to $\|\mathbf{R}_{11}\|_2 = o_p(1)$, $\|\mathbf{R}_{22}\|_2 = o_p(M^{-1})$ and $\|\mathbf{R}_{12}\|_2 = o_p(M^{-1/2})$. Hence, in terms of their 2-norms,

$$-\boldsymbol{\Sigma}^{-1}\begin{bmatrix}\mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{12}^{\mathrm{T}} & \mathbf{R}_{22}\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} O(1) & O(M^{1/2}) \\ O(M^{1/2}) & O(M) \end{bmatrix} \begin{bmatrix} o_p(1) & o_p(M^{-1/2}) \\ o_p(M^{-1/2}) & o_p(M^{-1}) \end{bmatrix} \begin{bmatrix} O_p(n^{-1/2}) \\ O_p(n^{-1/2}M^{1/2}) \end{bmatrix} \\
&= \begin{bmatrix} o_p(n^{-1/2}) \\ o_p(n^{-1/2}M^{1/2}) \end{bmatrix}.
\end{aligned}
$$

Therefore we get

$$
\begin{bmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \end{bmatrix} = \boldsymbol{\Sigma}^{-1} n^{-1} \sum_{i=1}^{n} \begin{bmatrix} \mathbf{x}_i \{ y_i - E(Y \mid \mathbf{x}_i) \} \\ \mathbf{D}(y_i) - E\{ \mathbf{D}(Y) \mid \mathbf{x}_i \} \end{bmatrix} + \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix},
$$

where $\|\mathbf{r}_1\|_2 = o_p(n^{-1/2})$ and $\|\mathbf{r}_2\|_2 = o_p(n^{-1/2}M^{1/2})$, which proves the second result of Proposition 3.3.4. Also, we can express $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ as

$$
\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \begin{bmatrix} \mathbf{I}_p & \mathbf{0}_M \end{bmatrix} \boldsymbol{\Sigma}^{-1} n^{-1} \sum_{i=1}^{n} \begin{bmatrix} \mathbf{x}_i \{ y_i - E(Y \mid \mathbf{x}_i) \} \\ \mathbf{D}(y_i) - E\{ \mathbf{D}(Y) \mid \mathbf{x}_i \} \end{bmatrix} + \mathbf{r}_1.
$$

Then since

$$
E\left(\begin{bmatrix} \mathbf{X} \{ Y - E(Y \mid \mathbf{X}) \} \\ \mathbf{D}(Y) - E\{ \mathbf{D}(Y) \mid \mathbf{X} \} \end{bmatrix}^{\otimes 2} \right) = \boldsymbol{\Sigma},
$$

using (B.108) we can conclude $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \sqrt{n} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}_p, \mathbf{I}_p)$ in distribution as $n \to \infty$. □

B.15 Proof of Theorem 3.3.6

We can express $\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0$ as

$$
\begin{aligned}
\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0 &= (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) n^{-1} \sum_{i=1}^{n} \mathrm{var}(Y \mid \mathbf{x}_i, \widehat{\boldsymbol{\theta}}) + \boldsymbol{\beta}_0 n^{-1} \sum_{i=1}^{n} \{ \mathrm{var}(Y \mid \mathbf{x}_i, \widehat{\boldsymbol{\theta}}) - \mathrm{var}(Y \mid \mathbf{x}_i) \} \\
&\quad + \boldsymbol{\beta}_0 \left[n^{-1} \sum_{i=1}^{n} \mathrm{var}(Y \mid \mathbf{x}_i) - E\{ \mathrm{var}(Y \mid \mathbf{X}) \} \right] \\
&= \left\{ n^{-1} \sum_{i=1}^{n} \mathrm{var}(Y \mid \mathbf{x}_i, \widehat{\boldsymbol{\theta}}) \mathbf{I} + \boldsymbol{\beta}_0 n^{-1} \sum_{i=1}^{n} \frac{\partial \mathrm{var}(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \right\} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\quad + \boldsymbol{\beta}_0 n^{-1} \sum_{i=1}^{n} \frac{\partial \mathrm{var}(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)
\end{aligned}
$$

$$+\boldsymbol{\beta}_0 \left[n^{-1} \sum_{i=1}^{n} \text{var}(Y \mid \mathbf{x}_i) - E\{\text{var}(Y \mid \mathbf{X})\} \right], \tag{B.109}$$

where $\boldsymbol{\theta}^*$ is on the line connecting $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Since $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_p(n^{-1/2}M^{1/2})$ by Proposition 3.3.4 and $\text{var}(Y \mid \mathbf{x}_i, \boldsymbol{\theta})$ is Lipschitz continuous at $\boldsymbol{\theta}_0$ uniformly with respect to $\mathbf{x}$ by Conditions (D1) and (D2), it is easy to see that

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} \text{var}(Y \mid \mathbf{x}_i, \widehat{\boldsymbol{\theta}}) &= n^{-1} \sum_{i=1}^{n} \text{var}(Y \mid \mathbf{x}_i) + O_p(n^{-1/2}M^{1/2}) \\ &= E\{\text{var}(Y \mid \mathbf{X})\} + O_p(n^{-1/2}M^{1/2}). \end{aligned} \tag{B.110}$$

Similarly, Conditions (D1) and (D2) lead to

$$\begin{aligned} & \left\| n^{-1} \sum_{i=1}^{n} \frac{\partial \text{var}(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\beta}} - E\left[\mathbf{X}\{Y - E(Y \mid \mathbf{X})\}^3 \right] \right\|_2 \\ = & \left\| n^{-1} \sum_{i=1}^{n} \mathbf{x}_i E[\{Y - E(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*)\}^3 \mid \mathbf{x}_i, \boldsymbol{\theta}^*] - E\left[\mathbf{X}\{Y - E(Y \mid \mathbf{X})\}^3 \right] \right\|_2 \\ = & O_p(n^{-1/2}M^{1/2}). \end{aligned} \tag{B.111}$$

In addition, in terms of the 2-norm, we have

$$\begin{aligned} & n^{-1} \sum_{i=1}^{n} \frac{\partial \text{var}(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\gamma}} \\ = & n^{-1} \sum_{i=1}^{n} \big[E\{\mathbf{D}(Y)Y^2 \mid \mathbf{x}_i, \boldsymbol{\theta}^*\} - 2E\{\mathbf{D}(Y)Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*\} E(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*) \\ & + 2E\{\mathbf{D}(Y) \mid \mathbf{x}_i, \boldsymbol{\theta}^*\} E(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*)^2 - E\{\mathbf{D}(Y) \mid \mathbf{x}_i, \boldsymbol{\theta}^*\} E(Y^2 \mid \mathbf{x}_i, \boldsymbol{\theta}^*) \big] \\ = & n^{-1} \sum_{i=1}^{n} \{ \mathbf{p}_2(\mathbf{x}_i, \boldsymbol{\theta}^*) - 2\mathbf{p}_1(\mathbf{x}_i, \boldsymbol{\theta}^*) E(Y \mid \mathbf{x}_i) \\ & + 2\mathbf{p}(\mathbf{x}_i, \boldsymbol{\theta}^*) E(Y \mid \mathbf{x}_i)^2 - \mathbf{p}(\mathbf{x}_i, \boldsymbol{\theta}^*) E(Y^2 \mid \mathbf{x}_i) \big\} \\ & + O_p(n^{-1/2}M^{1/2}) \\ = & n^{-1} \sum_{i=1}^{n} \{ \mathbf{p}_2(\mathbf{x}_i, \boldsymbol{\theta}_0) - 2\mathbf{p}_1(\mathbf{x}_i, \boldsymbol{\theta}_0) E(Y \mid \mathbf{x}_i) \\ & + 2\mathbf{p}(\mathbf{x}_i, \boldsymbol{\theta}_0) E(Y \mid \mathbf{x}_i)^2 - \mathbf{p}(\mathbf{x}_i, \boldsymbol{\theta}_0) E(Y^2 \mid \mathbf{x}_i) \big\} \\ & + O_p(n^{-1/2}M^{1/2}) \end{aligned}$$

$$
\begin{aligned}
&= n^{-1}\sum_{i=1}^{n} E([\mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{x}_i\}]\{Y - E(Y \mid \mathbf{x}_i)\}^2 \mid \mathbf{x}_i) + O_p(n^{-1/2}M^{1/2}) \\
&= E([\mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\}]\{Y - E(Y \mid \mathbf{X})\}^2) \\
&\quad + n^{-1}\sum_{i=1}^{n} E([\mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{x}_i\}]\{Y - E(Y \mid \mathbf{x}_i)\}^2 \mid \mathbf{x}_i) \\
&\quad - E([\mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\}]\{Y - E(Y \mid \mathbf{X})\}^2) \\
&\quad + O_p(n^{-1/2}M^{1/2}).
\end{aligned}
$$

The second equality holds by Conditions (D1) and (D2), and the third equality holds by Lemma B.13.1. Now let $\mathbf{W}_1 \equiv \mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\}$ and $W_2 \equiv \{Y - E(Y \mid \mathbf{X})\}^2$, then we can show $\|E(\mathbf{W}_1 W_2)\|_2 = O(M^{-1/2})$ since for any $\mathbf{u} \in \mathbb{R}^M$ such that $\|\mathbf{u}\|_2 = 1$,

$$
\begin{aligned}
\{E(\mathbf{W}_1 W_2)^{\mathrm{T}}\mathbf{u}\}^2 &= \mathbf{u}^{\mathrm{T}} E\{\mathrm{cov}(\mathbf{W}_1, W_2 \mid \mathbf{X})\} E\{\mathrm{cov}(W_2, \mathbf{W}_1 \mid \mathbf{X})\}\mathbf{u} \\
&\leqslant \mathbf{u}^{\mathrm{T}} E\{\mathrm{cov}(\mathbf{W}_1, W_2 \mid \mathbf{X})\mathrm{cov}(W_2, \mathbf{W}_1 \mid \mathbf{X})\}\mathbf{u} \\
&\leqslant \mathbf{u}^{\mathrm{T}} E\{\mathrm{var}(\mathbf{W}_1 \mid \mathbf{X})\mathrm{var}(W_2 \mid \mathbf{X})\}\mathbf{u} \\
&\leqslant C\mathbf{u}^{\mathrm{T}} E\{\mathrm{var}(\mathbf{W}_1 \mid \mathbf{X})\}\mathbf{u} \\
&= O(M^{-1})
\end{aligned}
$$

for some constant $C > 0$. The fourth argument holds since $\mathrm{var}(W_2 \mid \mathbf{x})$ is uniformly bounded by Condition (D1), and the last argument holds because $\|E\{\mathrm{var}(\mathbf{W}_1 \mid \mathbf{X})\}\|_2 = \|\boldsymbol{\Sigma}_{22}\|_2 \asymp M^{-1}$ by Remark 3.3.1. This leads to

$$
\|\mathbf{A}_2\|_2 = \|\boldsymbol{\beta}_0 E(\mathbf{W}_1 W_2)^{\mathrm{T}}\|_2 = \|\boldsymbol{\beta}_0\|_2 \|E(\mathbf{W}_1 W_2)\|_2 = O(M^{-1/2}). \tag{B.112}
$$

Similarly,

$$
\begin{aligned}
E\left[\left\{M^{1/2}E(\mathbf{W}_1^{\mathrm{T}}\mathbf{u}W_2 \mid \mathbf{X})\right\}^2\right] &= ME\left[\left\{\mathrm{cov}(\mathbf{W}_1^{\mathrm{T}}\mathbf{u}, W_2 \mid \mathbf{X})\right\}^2\right] \\
&\leqslant ME\left\{\mathrm{var}(\mathbf{W}_1^{\mathrm{T}}\mathbf{u} \mid \mathbf{X})\mathrm{var}(W_2 \mid \mathbf{X})\right\} \\
&\leqslant CM\mathbf{u}^{\mathrm{T}} E\left\{\mathrm{var}(\mathbf{W}_1 \mid \mathbf{X})\right\}\mathbf{u} \\
&= O(1),
\end{aligned}
$$

i.e. the second moment of $M^{1/2}E(\mathbf{W}_1^{\mathrm{T}}\mathbf{u}W_2 \mid \mathbf{X})$ is finite, then

$$
\left\| n^{-1}\sum_{i=1}^{n} E(\mathbf{W}_1 W_2 \mid \mathbf{x}_i) - E(\mathbf{W}_1 W_2) \right\|_2
$$

$$
\begin{aligned}
&= M^{-1/2} \sup_{\mathbf{u}\in\mathbf{R}^M:\|\mathbf{u}\|_2=1} \left| n^{-1}\sum_{i=1}^{n} M^{1/2} E(\mathbf{W}_1^{\mathrm{T}}\mathbf{u}W_2 \mid \mathbf{x}_i) - M^{1/2}E(\mathbf{W}_1^{\mathrm{T}}\mathbf{u}W_2)\right| \\
&= O_p(M^{-1/2}n^{-1/2}) \\
&= o_p(M^{-1/2}).
\end{aligned}
$$

Thus using $n^{-1/2}M^{1/2} = o(M^{-1/2})$ under Condition (D3), we get

$$
\begin{aligned}
&\left\| n^{-1}\sum_{i=1}^{n} \frac{\partial \mathrm{var}(Y \mid \mathbf{x}_i, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\gamma}} - E([\mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\}]\{Y - E(Y \mid \mathbf{X})\}^2)\right\|_2 \\
= &\left\| n^{-1}\sum_{i=1}^{n} E(\mathbf{W}_1 W_2 \mid \mathbf{x}_i) - E(\mathbf{W}_1 W_2)\right\|_2 + O_p(n^{-1/2}M^{1/2}) \\
= &\ o_p(M^{-1/2}). \qquad \text{(B.113)}
\end{aligned}
$$

Combining the results of Proposition 3.3.4, (B.109), (B.110), (B.111), (B.112), and (B.113), in terms of the 2-norm we get

$$
\begin{aligned}
&\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0 \\
= &\ \mathbf{A}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \mathbf{A}_2(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + \boldsymbol{\beta}_0 \left[n^{-1}\sum_{i=1}^{n} \mathrm{var}(Y \mid \mathbf{x}_i) - E\{\mathrm{var}(Y \mid \mathbf{X})\}\right] \\
&+O_p(n^{-1}M^{1/2}) + o_p(n^{-1/2}) \\
= &\ \mathbf{A}\boldsymbol{\Sigma}^{-1} n^{-1}\sum_{i=1}^{n} \begin{bmatrix} \mathbf{x}_i\{y_i - E(Y \mid \mathbf{x}_i)\} \\ \mathbf{D}(y_i) - E\{\mathbf{D}(Y) \mid \mathbf{x}_i\}\end{bmatrix} + \boldsymbol{\beta}_0 \left[n^{-1}\sum_{i=1}^{n} \mathrm{var}(Y \mid \mathbf{x}_i) - E\{\mathrm{var}(Y \mid \mathbf{X})\}\right] \\
&+\mathbf{A}_1\mathbf{r}_1 + \mathbf{A}_2\mathbf{r}_2 + o_p(n^{-1/2}) \\
= &\ \mathbf{A}\boldsymbol{\Sigma}^{-1} n^{-1}\sum_{i=1}^{n} \begin{bmatrix} \mathbf{x}_i\{y_i - E(Y \mid \mathbf{x}_i)\} \\ \mathbf{D}(y_i) - E\{\mathbf{D}(Y) \mid \mathbf{x}_i\}\end{bmatrix} + \boldsymbol{\beta}_0 \left[n^{-1}\sum_{i=1}^{n} \mathrm{var}(Y \mid \mathbf{x}_i) - E\{\mathrm{var}(Y \mid \mathbf{X})\}\right] \\
&+o_p(n^{-1/2})
\end{aligned}
$$

since $M = o(n^{1/3})$ by Condition (D3). Since

$$
\begin{aligned}
E\left(\begin{bmatrix} \mathbf{X}\{Y - E(Y \mid \mathbf{X})\} \\ \mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\}\end{bmatrix}^{\otimes 2}\right) &= \boldsymbol{\Sigma}, \\
\mathrm{cov}[\mathbf{X}\{Y - E(Y \mid \mathbf{X})\}, \mathrm{var}(Y \mid \mathbf{X})] &= \mathbf{0}, \\
\mathrm{cov}[\mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\}, \mathrm{var}(Y \mid \mathbf{X})] &= \mathbf{0},
\end{aligned}
$$

we can conclude $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0) \to N(\mathbf{0},\mathbf{I})$ in distribution as $n \to \infty$. □

B.16 Proof of Theorem 3.3.7

First note that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ and $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2 = O_p(n^{-1/2}M^{1/2})$ by Proposition 3.3.4. Also, $\|\boldsymbol{\Sigma}_{11}\|_2$ is of constant order by Conditions (D1), and since $\widehat{\boldsymbol{\Sigma}}_{11} = -n^{-1}\partial^2 l(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}$ we have $\|\widehat{\boldsymbol{\Sigma}}_{11}-\boldsymbol{\Sigma}_{11}\|_2 = o_p(1)$ by (B.101) through replacing $(\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*)$ by $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$. Further, $\|\boldsymbol{\Sigma}_{12}\|_2 = O(M^{-1/2})$ by (B.105), and

$$\widehat{\boldsymbol{\Sigma}}_{12} = \widehat{\boldsymbol{\Sigma}}_{21}^{\mathrm{T}} = -n^{-1}\partial^2 l(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})/\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^{\mathrm{T}}$$

then similarly $\|\widehat{\boldsymbol{\Sigma}}_{12}-\boldsymbol{\Sigma}_{12}\|_2 = o_p(M^{-1/2})$ by (B.106). Lastly, $\|\boldsymbol{\Sigma}_{22}\|_2 \asymp M^{-1}$ by Remark 3.3.1, and $\widehat{\boldsymbol{\Sigma}}_{22} = -n^{-1}\partial^2 l(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}$ which implies $\|\widehat{\boldsymbol{\Sigma}}_{22}-\boldsymbol{\Sigma}_{22}\|_2 = o_p(M^{-1})$ by (B.103). Hence, by Remark 3.3.1,

$$\|\widehat{\boldsymbol{\Sigma}}_{22}^{-1}-\boldsymbol{\Sigma}_{22}^{-1}\|_2 = \|\widehat{\boldsymbol{\Sigma}}_{22}^{-1}(\boldsymbol{\Sigma}_{22}-\widehat{\boldsymbol{\Sigma}}_{22})\boldsymbol{\Sigma}_{22}^{-1}\|_2 \asymp M \times o_p(M^{-1}) \times M = o_p(M).$$

Using the above results, we get $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{-1}-\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\|_2 \leqslant \|\widehat{\boldsymbol{\Sigma}}_{11}-\boldsymbol{\Sigma}_{11}\|_2 + \|\widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\widehat{\boldsymbol{\Sigma}}_{21} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\|_2 = o_p(1)$, hence $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}-\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\|_2 = o_p(1)$ by Condition (D5). In addition, note that $\boldsymbol{\Sigma}^{-1}$ can be expressed as

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} & -\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{22}^{-1}+\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix}. \tag{B.114}$$

We can easily show that $\|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\|_2 = O_p(M^{1/2})$, $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\|_2 = o_p(M^{1/2})$, $\|\boldsymbol{\Sigma}_{22}^{-1}+\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\|_2 = O_p(M)$, and $\|(\widehat{\boldsymbol{\Sigma}}_{22}^{-1}+\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\widehat{\boldsymbol{\Sigma}}_{21}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}) - (\boldsymbol{\Sigma}_{22}^{-1}+\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1})\|_2 = o_p(M)$. This implies each block component of $\widehat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1}$ has the 2-norm as

$$\widehat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} o_p(1) & o_p(M^{1/2}) \\ o_p(M^{1/2}) & o_p(M) \end{bmatrix}. \tag{B.115}$$

Now we will show $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}-\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\|_2 = o_p(1)$. First note that $\|\mathbf{A}_1\|_2$ is of constant order by Condition (D1) and $\|\widehat{\mathbf{A}}_1-\mathbf{A}_1\|_2 = o_p(1)$, because $\widehat{E}\{\widehat{\mathrm{var}}(Y|\mathbf{X})\} = E\{\mathrm{var}(Y|\mathbf{X})\} + o_p(1)$ by (B.110), $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by Proposition 3.3.4, and $\|\widehat{E}[\{Y-\widehat{E}(Y|\mathbf{X})\}^3\mathbf{X}] - E[\{Y-E(Y|\mathbf{X})\}^3\mathbf{X}]\|_2 = o_p(1)$ by (B.111). Also, $\|\mathbf{A}_2\|_2 = O(M^{-1/2})$ by (B.112) and $\|\widehat{\mathbf{A}}_2-\mathbf{A}_2\|_2 = o_p(M^{-1/2})$, since $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$ by Proposition 3.3.4 and

$\|\widehat{E}(\{Y-\widehat{E}(Y|\mathbf{X})\}^2[\mathbf{D}(Y)-\widehat{E}\{\mathbf{D}(Y)|\mathbf{X}\}])-E(\{Y-E(Y|\mathbf{X})\}^2[\mathbf{D}(Y)-E\{\mathbf{D}(Y)|\mathbf{X}\}])\|_2 = o_p(M^{-1/2})$ by (B.113). Further, since $\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{\text{var}}(Y|\mathbf{x})-\text{var}(Y|\mathbf{x})| = o_p(1)$ by Conditions (D1), (D2) and Proposition 3.3.4,

$$\begin{aligned}
& \widehat{\text{var}}\{\widehat{\text{var}}(Y|\mathbf{X})\} - \text{var}\{\text{var}(Y|\mathbf{X})\} \\
= \; & \widehat{E}[\{\widehat{\text{var}}(Y|\mathbf{X})\}^2 - \{\text{var}(Y|\mathbf{X})\}^2] + \widehat{E}[\{\text{var}(Y|\mathbf{X})\}^2] - E[\{\text{var}(Y|\mathbf{X})\}^2] \\
& -[\widehat{E}\{\widehat{\text{var}}(Y|\mathbf{X})\}]^2 + [\widehat{E}\{\text{var}(Y|\mathbf{X})\}]^2 - [\widehat{E}\{\text{var}(Y|\mathbf{X})\}]^2 + [E\{\text{var}(Y|\mathbf{X})\}]^2 \\
= \; & o_p(1).
\end{aligned}$$

Therefore, combining the above results with (B.114) and (B.115), we get $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}}\|_2 = o_p(1)$.

B.17 Proof of Proposition 3.3.5

The efficient score for $\boldsymbol{\beta}$ given in Appendix B.10 is $\mathbf{S}_{\text{eff}} = y\mathbf{x}-\mathbf{a}_0(y)-E\{Y\mathbf{x}-\mathbf{a}_0(Y) \mid \mathbf{x}\}$, where $\mathbf{a}_0(y)$ satisfies

$$\mathbf{a}_0(y) - E[E\{\mathbf{a}_0(Y) \mid \mathbf{X}\} \mid y] = E(y\mathbf{X} \mid y) - E\{E(Y\mathbf{X} \mid \mathbf{X}) \mid y\}. \tag{B.116}$$

Thus, the efficient variance is $\{E(\mathbf{S}_{\text{eff}}^{\otimes 2})\}^{-1}$.

To show that the MLE estimator $\widehat{\boldsymbol{\beta}}$ in Proposition 3.3.4 is efficient, we need to show $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} = E(\mathbf{S}_{\text{eff}}^{\otimes 2})$. $\mathbf{a}_0(y)$ must be of the general form $\Lambda\mathbf{D}(y)$, i.e. $\mathbf{a}_0(y) = \Lambda\mathbf{D}(y)$, where Λ is a $p \times M$ coefficient matrix. Then (B.116) implies that

$$\begin{aligned}
\Lambda E\{\mathbf{D}(Y)^{\otimes 2}\} - \Lambda E[E\{\mathbf{D}(Y) \mid \mathbf{X}\}^{\otimes 2}] &= E[E\{Y\mathbf{X} - E(Y\mathbf{X} \mid \mathbf{X})\mathbf{D}(Y)^{\text{T}} \mid \mathbf{X}\}] \\
&= E[\mathbf{X}\text{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{X}\}],
\end{aligned}$$

i.e.,

$$\mathbf{a}_0(y) = \Lambda\mathbf{D}(y) = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{D}(y).$$

Therefore,

$$E(\mathbf{S}_{\text{eff}}^{\otimes 2}) = E\left\{(Y\mathbf{X} - E(Y\mathbf{X} \mid \mathbf{X}) - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}[\mathbf{D}(Y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\}])^{\otimes 2}\right\} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}.$$

□

B.18 Proof of Theorem 3.3.8

Since the efficient score for $\boldsymbol{\xi}$ given in Section 3.2.1 leads to

$$\begin{aligned} E(\boldsymbol{\phi}_{\text{eff}}^{\otimes 2}) &= E\left([\boldsymbol{\beta} v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) - \boldsymbol{\beta} E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}]^{\otimes 2}\right) \\ &\quad + E\left([\boldsymbol{\beta} Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y - E\{\boldsymbol{\beta} Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y \mid \mathbf{X}\}]^{\otimes 2}\right), \end{aligned}$$

and the asymptotic variance of $\widehat{\boldsymbol{\xi}}$ in Theorem 3.3.6 is

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\beta}^{\otimes 2}\mathrm{var}\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\},$$

we only need to show

$$E\left([\boldsymbol{\beta} Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y - E\{\boldsymbol{\beta} Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y \mid \mathbf{X}\}]^{\otimes 2}\right) = \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}}.$$

Now we can always write $\boldsymbol{\beta} y^2 + \mathbf{a}(y) = \boldsymbol{\Lambda}\mathbf{D}(y)$, where $\boldsymbol{\Lambda} \in \mathbb{R}^{p \times M}$. Then (B.5) and (B.6) imply

$$\begin{aligned} & -\boldsymbol{\Lambda} E[\mathbf{D}(y) - E\{\mathbf{D}(Y) \mid \mathbf{X}\} \mid y] + \boldsymbol{\beta} E\{y^2 - E(Y^2 \mid \mathbf{X}) \mid y\} \\ = \;& 2\boldsymbol{\beta} E[yE(Y \mid \mathbf{X}) - E\{YE(Y \mid \mathbf{X}) \mid \mathbf{X}\} \mid y] + \mathbf{M} E[\mathbf{X}\{y - E(Y \mid \mathbf{X})\} \mid y] \\ = \;& 2\boldsymbol{\beta} E[yE(Y \mid \mathbf{X}) - E\{YE(Y \mid \mathbf{X}) \mid \mathbf{X}\} \mid y] \\ & + \left(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\left[2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}} Y v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \{\boldsymbol{\Lambda}\mathbf{D}(Y) - \boldsymbol{\beta} Y^2\}\{Y - E(Y \mid \mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right) \\ & \times \boldsymbol{\Sigma}_{11}^{-1} E[\mathbf{X}\{y - E(Y \mid \mathbf{X})\} \mid y]. \end{aligned}$$

Multiplying $\mathbf{D}^{\mathrm{T}}(y)$ on both sides and taking expectation lead to

$$\begin{aligned} & -\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{22} + \boldsymbol{\beta} E[\{Y^2 - E(Y^2 \mid \mathbf{X})\}\mathbf{D}^{\mathrm{T}}(Y)] \\ = \;& 2\boldsymbol{\beta} E[YE(Y \mid \mathbf{X})\mathbf{D}^{\mathrm{T}}(Y) - \{E(Y \mid \mathbf{X})\}^2\mathbf{D}^{\mathrm{T}}(Y)] \\ & + \left(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\left[2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}} Y v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \{\boldsymbol{\Lambda}\mathbf{D}(Y) - \boldsymbol{\beta} Y^2\}\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right) \\ & \times \boldsymbol{\Sigma}_{11}^{-1} E[\mathbf{X}\{Y - E(Y \mid \mathbf{X})\}\mathbf{D}^{\mathrm{T}}(Y)] \\ = \;& 2\boldsymbol{\beta} E[Y\mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{X}\}] + E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - E\left\{2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}} Y v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right\}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ & -\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} + \boldsymbol{\beta} E[Y^2\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}]\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ = \;& 2\boldsymbol{\beta} E[Y\mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{X}\}] + \mathbf{A}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}, \end{aligned}$$

where $\mathbf{A}_1, \mathbf{A}_2$ are defined in Theorem 3.3.6, hence

$$\begin{aligned} & -\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{22} + \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ = \; & 2\boldsymbol{\beta}E[Y\mathrm{cov}\{Y, \mathbf{D}(Y) \mid \mathbf{X}\}] + \mathbf{A}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \boldsymbol{\beta}E[\{Y^2 - E(Y^2 \mid \mathbf{X})\}\mathbf{D}^{\mathrm{T}}(Y)] \\ = \; & \mathbf{A}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \mathbf{A}_2, \end{aligned}$$

and

$$\boldsymbol{\beta}y^2 + \mathbf{a}(y) = (\mathbf{A}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \mathbf{A}_2)(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{22})^{-1}\mathbf{D}(y) = \mathbf{U}\mathbf{D}(y),$$

where, for notational brevity,

$$\mathbf{U} \;\equiv\; (\mathbf{A}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \mathbf{A}_2)(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{22})^{-1}.$$

Then

$$\begin{aligned} \mathbf{M} \;=\; & \left(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\{2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}}Yv(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} - \mathbf{U}E\left[\mathbf{D}(Y)\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right. \\ & \left. + \boldsymbol{\beta}E\left[Y^2\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right)\boldsymbol{\Sigma}_{11}^{-1} \\ = \; & \left(E\{v(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}\mathbf{I} - E\{2\boldsymbol{\beta}\mathbf{X}^{\mathrm{T}}Yv(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} - \mathbf{U}\boldsymbol{\Sigma}_{21} + \boldsymbol{\beta}E\left[Y^2\{Y - E(Y \mid \mathbf{X})\}\mathbf{X}^{\mathrm{T}}\right]\right)\boldsymbol{\Sigma}_{11}^{-1} \\ = \; & (\mathbf{A}_1 - \mathbf{U}\boldsymbol{\Sigma}_{21})\boldsymbol{\Sigma}_{11}^{-1}. \end{aligned}$$

Hence

$$\boldsymbol{\beta}y^2 + \mathbf{a}(y) - E\{\boldsymbol{\beta}Y^2 + \mathbf{a}(Y) \mid \mathbf{x}\} = \mathbf{U}[\mathbf{D}(y) - E\{\mathbf{D}(Y) \mid \mathbf{x}\}],$$

and

$$\begin{aligned} E\{\mathrm{var}(\mathbf{M}\mathbf{X}Y \mid \mathbf{X})\} &= \mathbf{M}\boldsymbol{\Sigma}_{11}\mathbf{M}^{\mathrm{T}}, \\ E[\mathrm{var}\{\boldsymbol{\beta}Y^2 + \mathbf{a}(Y) \mid \mathbf{X}\}] &= \mathbf{U}\boldsymbol{\Sigma}_{22}\mathbf{U}^{\mathrm{T}}, \\ E[\mathrm{cov}\{\mathbf{M}\mathbf{X}Y, \boldsymbol{\beta}Y^2 + \mathbf{a}(Y) \mid \mathbf{X}\}] &= \mathbf{M}\boldsymbol{\Sigma}_{12}\mathbf{U}^{\mathrm{T}}. \end{aligned}$$

Thus, noting that $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$,

$$\begin{aligned} & E\left([\boldsymbol{\beta}Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y - E\{\boldsymbol{\beta}Y^2 + \mathbf{a}(Y) + \mathbf{M}\mathbf{X}Y \mid \mathbf{X}\}]^{\otimes 2}\right) \\ = \; & \mathbf{M}\boldsymbol{\Sigma}_{11}\mathbf{M}^{\mathrm{T}} + \mathbf{U}\boldsymbol{\Sigma}_{22}\mathbf{U}^{\mathrm{T}} + \mathbf{M}\boldsymbol{\Sigma}_{12}\mathbf{U}^{\mathrm{T}} + \mathbf{U}\boldsymbol{\Sigma}_{21}\mathbf{M}^{\mathrm{T}} \\ = \; & (\mathbf{A}_1 - \mathbf{U}\boldsymbol{\Sigma}_{21})\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{A}_1^{\mathrm{T}} - \boldsymbol{\Sigma}_{12}\mathbf{U}^{\mathrm{T}}) + \mathbf{U}\boldsymbol{\Sigma}_{22}\mathbf{U}^{\mathrm{T}} + (\mathbf{A}_1 - \mathbf{U}\boldsymbol{\Sigma}_{21})\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{U}^{\mathrm{T}} \end{aligned}$$

$$
\begin{aligned}
& +\mathbf{U}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{A}_1^{\mathrm{T}} - \boldsymbol{\Sigma}_{12}\mathbf{U}^{\mathrm{T}}) \\
= \;& \mathbf{A}_1\boldsymbol{\Sigma}_{11}^{-1}\mathbf{A}_1^{\mathrm{T}} + \mathbf{U}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})\mathbf{U}^{\mathrm{T}} \\
= \;& \mathbf{A}_1\boldsymbol{\Sigma}_{11}^{-1}\mathbf{A}_1^{\mathrm{T}} + (\mathbf{A}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \mathbf{A}_2)(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{A}_1^{\mathrm{T}} - \mathbf{A}_2^{\mathrm{T}}),
\end{aligned}
$$

which equals to

$$
\mathbf{A}\begin{pmatrix}
\boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \\
-(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}
\end{pmatrix}\mathbf{A}^{\mathrm{T}},
$$

then this is equal to $\mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathrm{T}}$. □

B.19 Additional tables for simulation experiments

Table B.1: $\boldsymbol{\eta}_\tau$ estimation results under the truncated normal distribution.

τ		Method	|bias|	σ_{sim}	$\widehat{\sigma}_{\text{est}}$	C.I.
0.05	η_1	aMLE	0.022	0.028	0.029	0.954
		pMLE	0.026	0.029	-	-
		MLE	0.035	0.035	0.035	0.893
	η_2	aMLE	0.029	0.036	0.034	0.937
		pMLE	0.036	0.035	-	-
		MLE	0.053	0.041	0.036	0.681
	η_3	aMLE	0.033	0.041	0.042	0.957
		pMLE	0.047	0.040	-	-
		MLE	0.075	0.044	0.035	0.463
0.25	η_1	aMLE	0.023	0.029	0.029	0.950
		pMLE	0.023	0.029	-	-
		MLE	0.028	0.035	0.035	0.954
	η_2	aMLE	0.028	0.036	0.034	0.941
		pMLE	0.028	0.035	-	-
		MLE	0.035	0.041	0.036	0.898
	η_3	aMLE	0.032	0.040	0.040	0.953
		pMLE	0.032	0.039	-	-
		MLE	0.040	0.044	0.035	0.839
0.50	η_1	aMLE	0.023	0.029	0.029	0.945
		pMLE	0.023	0.030	-	-
		MLE	0.029	0.035	0.035	0.940
	η_2	aMLE	0.028	0.036	0.034	0.945
		pMLE	0.029	0.035	-	-
		MLE	0.039	0.041	0.036	0.841
	η_3	aMLE	0.032	0.040	0.040	0.957
		pMLE	0.033	0.039	-	-
		MLE	0.049	0.044	0.035	0.742
0.75	η_1	aMLE	0.023	0.029	0.029	0.941
		pMLE	0.023	0.029	-	-
		MLE	0.028	0.035	0.035	0.953
	η_2	aMLE	0.028	0.036	0.034	0.940
		pMLE	0.028	0.034	-	-
		MLE	0.035	0.041	0.036	0.898
	η_3	aMLE	0.033	0.040	0.040	0.959
		pMLE	0.032	0.039	-	-
		MLE	0.039	0.044	0.035	0.846
0.95	η_1	aMLE	0.023	0.028	0.029	0.940
		pMLE	0.025	0.029	-	-
		MLE	0.036	0.035	0.035	0.893
	η_2	aMLE	0.028	0.035	0.034	0.945
		pMLE	0.032	0.034	-	-
		MLE	0.053	0.041	0.036	0.678
	η_3	aMLE	0.033	0.041	0.042	0.964
		pMLE	0.041	0.038	-	-
		MLE	0.075	0.044	0.035	0.458

Table B.2: $\boldsymbol{\eta}_\tau$ estimation results under the normal distribution.

τ		Method	|bias|	σ_{sim}	$\widehat{\sigma}_{\text{est}}$	C.I.
0.05	η_1	aMLE	0.026	0.033	0.032	0.947
		pMLE	0.030	0.039	-	-
		MLE	0.026	0.032	0.032	0.954
	η_2	aMLE	0.026	0.032	0.032	0.948
		pMLE	0.031	0.039	-	-
		MLE	0.025	0.032	0.032	0.957
	η_3	aMLE	0.026	0.033	0.033	0.934
		pMLE	0.032	0.039	-	-
		MLE	0.025	0.032	0.032	0.941
0.25	η_1	aMLE	0.026	0.032	0.032	0.950
		pMLE	0.030	0.038	-	-
		MLE	0.026	0.032	0.032	0.954
	η_2	aMLE	0.025	0.032	0.032	0.950
		pMLE	0.030	0.039	-	-
		MLE	0.025	0.032	0.032	0.957
	η_3	aMLE	0.025	0.032	0.032	0.940
		pMLE	0.031	0.038	-	-
		MLE	0.025	0.032	0.032	0.941
0.50	η_1	aMLE	0.026	0.032	0.032	0.953
		pMLE	0.029	0.038	-	-
		MLE	0.026	0.032	0.032	0.954
	η_2	aMLE	0.025	0.032	0.032	0.955
		pMLE	0.030	0.039	-	-
		MLE	0.025	0.032	0.032	0.957
	η_3	aMLE	0.025	0.032	0.032	0.935
		pMLE	0.030	0.038	-	-
		MLE	0.025	0.032	0.032	0.941
0.75	η_1	aMLE	0.026	0.032	0.032	0.953
		pMLE	0.029	0.038	-	-
		MLE	0.026	0.032	0.032	0.954
	η_2	aMLE	0.025	0.032	0.032	0.951
		pMLE	0.030	0.038	-	-
		MLE	0.025	0.032	0.032	0.957
	η_3	aMLE	0.026	0.032	0.032	0.938
		pMLE	0.030	0.038	-	-
		MLE	0.025	0.032	0.032	0.941
0.95	η_1	aMLE	0.026	0.032	0.032	0.950
		pMLE	0.029	0.038	-	-
		MLE	0.026	0.032	0.032	0.954
	η_2	aMLE	0.026	0.032	0.032	0.951
		pMLE	0.031	0.039	-	-
		MLE	0.025	0.032	0.032	0.957
	η_3	aMLE	0.027	0.034	0.033	0.937
		pMLE	0.033	0.039	-	-
		MLE	0.025	0.032	0.032	0.941

Table B.3: $\boldsymbol{\eta}_\tau$ estimation results under the gamma distribution.

τ		Method	|bias|	σ_{sim}	$\widehat{\sigma}_{\text{est}}$	C.I.
0.05	η_1	aMLE	0.031	0.040	0.040	0.949
		pMLE	0.561	0.690	-	-
		MLE	0.021	0.026	0.025	0.950
	η_2	aMLE	0.038	0.049	0.049	0.939
		pMLE	0.840	0.588	-	-
		MLE	0.022	0.028	0.028	0.950
0.25	η_1	aMLE	0.051	0.066	0.064	0.942
		pMLE	1.070	0.953	-	-
		MLE	0.035	0.044	0.043	0.949
	η_2	aMLE	0.058	0.074	0.073	0.943
		pMLE	1.666	0.753	-	-
		MLE	0.037	0.046	0.046	0.953
0.50	η_1	aMLE	0.070	0.089	0.088	0.944
		pMLE	1.259	1.080	-	-
		MLE	0.048	0.060	0.059	0.948
	η_2	aMLE	0.077	0.096	0.096	0.950
		pMLE	1.945	0.844	-	-
		MLE	0.051	0.064	0.064	0.950
0.75	η_1	aMLE	0.093	0.117	0.116	0.944
		pMLE	1.246	1.410	-	-
		MLE	0.064	0.081	0.080	0.948
	η_2	aMLE	0.103	0.127	0.126	0.944
		pMLE	1.831	0.907	-	-
		MLE	0.069	0.086	0.086	0.951
0.95	η_1	aMLE	0.146	0.186	0.185	0.948
		pMLE	0.828	1.529	-	-
		MLE	0.094	0.118	0.116	0.949
	η_2	aMLE	0.178	0.227	0.232	0.957
		pMLE	0.991	1.025	-	-
		MLE	0.102	0.127	0.126	0.948

Appendix C
Supplement to Chapter 4

C.1 Proof of Theorem 4.2.1

Given an arbitrary sampling scheme $\mathbf{w}$ that satisfies $0 < w_i \leqslant 1$ for $i = 1, \ldots, N_0$ and $\sum_{i=1}^{N_0} w_i \leqslant N_0\pi_0$, let $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ satisfy (4.2). Solving (4.2) is equivalent to minimizing

$$
\begin{aligned}
& E[\{Y - p_0(\mathbf{X})\}^2] + E\{p_0(\mathbf{X})^2\} + N^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_T(\mathbf{x}_i, S_i)}\left\{p(\mathbf{x}_i, \boldsymbol{\beta})^2 - 2Y_i p(\mathbf{x}_i, \boldsymbol{\beta})\right\} \\
= \; & E[\{Y - p_0(\mathbf{X})\}^2] + E\{p_0(\mathbf{X})^2\} + E\left[\frac{R}{\pi_T(\mathbf{X}, S)}\{p(\mathbf{X}, \boldsymbol{\beta})^2 - 2Y p(\mathbf{X}, \boldsymbol{\beta})\}\right] + o_p(1) \\
= \; & E[\{Y - p_0(\mathbf{X})\}^2] + E\{p_0(\mathbf{X})^2\} + E\left\{p(\mathbf{X}, \boldsymbol{\beta})^2 - 2p_0(\mathbf{X})p(\mathbf{X}, \boldsymbol{\beta})\right\} + o_p(1) \\
= \; & E[\{Y - p_0(\mathbf{X})\}^2] + E[\{p_0(\mathbf{X}) - p(\mathbf{X}, \boldsymbol{\beta})\}^2] + o_p(1) \\
= \; & E[\{Y - p(\mathbf{X}, \boldsymbol{\beta})\}^2] + o_p(1).
\end{aligned}
$$

This means that $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$ is the minimizer of an estimator of the classification MSE. Now, by the definition of $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$, we have

$$
\begin{aligned}
\mathbf{0} \; = \; & N^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_T(\mathbf{x}_i, S_i)}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})\left\{p(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\mathbf{w}}) - Y_i\right\} \\
= \; & N^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_T(\mathbf{x}_i, S_i)}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*)\left\{p(\mathbf{x}_i, \boldsymbol{\beta}^*) - Y_i\right\} \\
& + E\left(\frac{R}{\pi_T(\mathbf{X}, S)}\left[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*)^{\otimes 2} + \mathbf{p}''_{\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}}(\mathbf{X}, \boldsymbol{\beta}^*)\left\{p(\mathbf{X}, \boldsymbol{\beta}^*) - Y\right\}\right]\right) N^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*) \\
& + o_p(1) \\
= \; & N^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_T(\mathbf{x}_i, S_i)}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*)\left\{p(\mathbf{x}_i, \boldsymbol{\beta}^*) - Y_i\right\} + \mathbf{A}N^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*) + o_p(1),
\end{aligned}
$$

where $\mathbf{A} \equiv E[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*)^{\otimes 2} + \mathbf{p}''_{\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}}(\mathbf{X}, \boldsymbol{\beta}^*)\{p(\mathbf{X}, \boldsymbol{\beta}^*) - p_0(\mathbf{X})\}]$. Note that Condition (C1) ensures $\boldsymbol{\beta}^*$ is the unique solution of

$$\mathbf{0} = E[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta})\{p(\mathbf{X}, \boldsymbol{\beta}) - p_0(\mathbf{X})\}] = E\left[\frac{R}{\pi_T(\mathbf{X}, S)}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta})\left\{p(\mathbf{X}, \boldsymbol{\beta}) - Y\right\}\right].$$

Hence under Condition (C2), we get

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*) \xrightarrow{\mathrm{D}} \mathrm{N}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}_{\pi}\mathbf{A}^{-1}), \tag{C.1}$$

where

$$\begin{aligned}
\mathbf{B}_{\pi} &\equiv \mathrm{var}\left[\frac{R}{\pi_T(\mathbf{X}, S)}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*)\left\{p(\mathbf{X}, \boldsymbol{\beta}^*) - Y\right\}\right] \\
&= E\left[\frac{R}{\pi_T^2(\mathbf{X}, S)}\mathbf{p}'^{\otimes 2}_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*)\left\{p^2(\mathbf{X}, \boldsymbol{\beta}^*) + Y - 2Yp(\mathbf{X}, \boldsymbol{\beta}^*)\right\}\right] \\
&= E\left[\frac{1}{\pi_T(\mathbf{X}, S)}\mathbf{p}'^{\otimes 2}_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*)\left\{p^2(\mathbf{X}, \boldsymbol{\beta}^*) + p_0(\mathbf{X}) - 2p_0(\mathbf{X})p(\mathbf{X}, \boldsymbol{\beta}^*)\right\}\right].
\end{aligned}$$

Now, the mean squared error of $p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})$ is

$$\begin{aligned}
& E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}^2] \\
= \; & E[\{Y^* - p_0(\mathbf{X}^*)\}^2] + E[\{p_0(\mathbf{X}^*) - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}^2] \\
= \; & E[\{Y^* - p_0(\mathbf{X}^*)\}^2] + E[\{p_0(\mathbf{X}^*) - p(\mathbf{X}^*, \boldsymbol{\beta}^*)\}^2] + E[\{p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}}) - p(\mathbf{X}^*, \boldsymbol{\beta}^*)\}^2] \\
& + 2E[\{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}E\{p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}}) - p(\mathbf{X}^*, \boldsymbol{\beta}^*)|\mathbf{X}^*\}].
\end{aligned} \tag{C.2}$$

By the Taylor expansion of $p(\mathbf{X}^*, \boldsymbol{\beta})$, we can write the last two components as

$$\begin{aligned}
& E[\{p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}}) - p(\mathbf{X}^*, \boldsymbol{\beta}^*)\}^2] \\
& + 2E[\{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}E\{p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}}) - p(\mathbf{X}^*, \boldsymbol{\beta}^*)|\mathbf{X}^*\}] \\
= \; & E\{(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}^*, \boldsymbol{\beta}^*)^{\otimes 2}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*)\} \\
& + 2E[\{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}\mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{X}^*, \boldsymbol{\beta}^*)E\{(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*)|\mathbf{X}^*\}] \\
& + E(\{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}E[(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*)^{\mathrm{T}}\{\partial^2 p(\mathbf{X}^*, \boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}\}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*)|\mathbf{X}^*]) \\
& + O\{E(\|\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*\|_2^3)\} \\
= \; & \mathrm{trace}(E[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}^*, \boldsymbol{\beta}^*)^{\otimes 2}E\{(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*)^{\otimes 2}|\mathbf{X}^*\}]) \\
& + E(\{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}\mathrm{trace}[\{\partial^2 p(\mathbf{X}^*, \boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}\}E\{(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*)^{\otimes 2}|\mathbf{X}^*\}]) \\
& + O\{E(\|\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*\|_2^3)\}
\end{aligned}$$

$$
\begin{aligned}
&= \operatorname{trace}[E\{\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}^*, \boldsymbol{\beta}^*)^{\otimes 2}\}\operatorname{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}})] \\
&\quad + \operatorname{trace}(E[\{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}\{\partial^2 p(\mathbf{X}^*, \boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}\}]\operatorname{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}})) \\
&\quad + O\{E(\|\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*\|_2^3)\} \\
&= \operatorname{trace}\{\mathbf{A}\operatorname{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}})\} + O\{E(\|\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*\|_2^3)\}. \qquad \text{(C.3)}
\end{aligned}
$$

The second equality in (C.3) holds because $\mathbf{X}^*$ is independent of $\widehat{\boldsymbol{\beta}}_{\mathbf{w}}$, thus

$$
\begin{aligned}
& E[\{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}\mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{X}^*, \boldsymbol{\beta}^*)E\{(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*)|\mathbf{X}^*\}] \\
= \; & E[\mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{X}^*, \boldsymbol{\beta}^*)\{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}]E(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*) \\
= \; & 0.
\end{aligned}
$$

by Condition (C1). Therefore, incorporating (C.1), we get

$$
NE[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}^2 - \{Y^* - p(\mathbf{X}^*, \boldsymbol{\beta}^*)\}^2] \to \operatorname{trace}(\mathbf{B}_\pi \mathbf{A}^{-1}) \qquad \text{(C.4)}
$$

as $N \to \infty$. We further have

$$
\begin{aligned}
& \operatorname{trace}(\mathbf{B}_\pi \mathbf{A}^{-1}) \qquad \text{(C.5)} \\
= \; & \operatorname{trace}\left(E\left[\frac{1}{\pi_T(\mathbf{X}, S)} \mathbf{p}'^{\otimes 2}_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*) \left\{ p^2(\mathbf{X}, \boldsymbol{\beta}^*) + p_0(\mathbf{X}) - 2p_0(\mathbf{X})p(\mathbf{X}, \boldsymbol{\beta}^*) \right\} \right] \mathbf{A}^{-1} \right) \\
= \; & E\left\{ \frac{p^2(\mathbf{X}, \boldsymbol{\beta}^*) + p_0(\mathbf{X}) - 2p_0(\mathbf{X})p(\mathbf{X}, \boldsymbol{\beta}^*)}{\pi_T(\mathbf{X}, S)} \mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{X}, \boldsymbol{\beta}^*)\mathbf{A}^{-1}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*) \right\} \\
= \; & N^{-1} \sum_{i=1}^{n} \frac{p^2(\mathbf{x}_i, \boldsymbol{\beta}^*) + p_0(\mathbf{x}_i) - 2p_0(\mathbf{x}_i)p(\mathbf{x}_i, \boldsymbol{\beta}^*)}{(1 - S_i)w_i + S_i} \mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}^*)\mathbf{A}_N^{*-1}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*) + o_p(1) \\
= \; & N^{-1} \sum_{i=1}^{N_0} \frac{p^2(\mathbf{x}_i, \boldsymbol{\beta}^*) + p_0(\mathbf{x}_i) - 2p_0(\mathbf{x}_i)p(\mathbf{x}_i, \boldsymbol{\beta}^*)}{w_i} \mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}^*)\mathbf{A}_N^{*-1}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*) \\
& + N^{-1} \sum_{i=N_0+1}^{N} \{p^2(\mathbf{x}_i, \boldsymbol{\beta}^*) + p_0(\mathbf{x}_i) - 2p_0(\mathbf{x}_i)p(\mathbf{x}_i, \boldsymbol{\beta}^*)\}\mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}^*)\mathbf{A}_N^{*-1}\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*) \\
& + o_p(1),
\end{aligned}
$$

where

$$
\mathbf{A}_N^* \equiv N^{-1} \sum_{i=1}^{n} \left[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*)^{\otimes 2} + \mathbf{p}''_{\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}}(\mathbf{x}_i, \boldsymbol{\beta}^*)\{p(\mathbf{x}_i, \boldsymbol{\beta}^*) - p_0(\mathbf{x}_i)\} \right].
$$

Note that the second term in (C.5) does not depend on $\mathbf{w}$. By definition, when $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$ and $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, $\mathbf{A}_N^* = \mathbf{A}_N$ and the first term in (C.5) is identical to the target function in (4.1), hence $\mathbf{w}_{\text{opt}}$ minimizes the leading term in (C.5) under $0 < w_i \leqslant 1$ for $i = 1, \ldots, N_0$ and $\sum_{i=1}^{N_0} w_i \leqslant N_0\pi_0$. Therefore, $c_\pi \equiv \text{trace}(\mathbf{B}_\pi \mathbf{A}^{-1}) - \text{trace}(\mathbf{B}_{\pi_{\text{opt}}} \mathbf{A}^{-1}) \geqslant 0$, and by (C.4),

$$N\left(E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}^2] - E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}_{\text{opt}}})\}^2]\right) \to c_\pi \geqslant 0$$

when $N \to \infty$. □

C.2 Lemma

Lemma C.2.1. *Assume the regularity conditions specified in Robins et al. (1994) and let $\widehat{\boldsymbol{\beta}}$ be the solution of (4.2). Then $\widehat{\boldsymbol{\beta}}$ is a realization of the locally efficient estimator of $\boldsymbol{\beta}^*$ and is efficient when $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$.*

Proof. Let

$$\mathbf{U}(\mathbf{x}, Y; \boldsymbol{\beta}) \equiv \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}, \boldsymbol{\beta}) \{p(\mathbf{x}, \boldsymbol{\beta}) - Y\},$$

then we have $E\{\mathbf{U}(\mathbf{X}, Y; \boldsymbol{\beta}^*)\} = E[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}^*) \{p(\mathbf{X}, \boldsymbol{\beta}^*) - p_0(\mathbf{X})\}] = \mathbf{0}$ by Condition (C1). When $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$, we further have $E\{\mathbf{U}(\mathbf{X}, Y; \boldsymbol{\beta}^*) \mid \mathbf{X}\} = \mathbf{0}$, so we can write (4.2) as

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{n} \frac{R_i}{\pi_T(\mathbf{x}_i, S_i)} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}) \{p(\mathbf{x}_i, \boldsymbol{\beta}) - Y_i\} \\ &= \sum_{i=1}^{n} \left[\frac{R_i}{\pi_T(\mathbf{x}_i, S_i)} \mathbf{U}(\mathbf{x}_i, Y_i; \boldsymbol{\beta}) - \frac{R_i - \pi_T(\mathbf{x}_i, S_i)}{\pi_T(\mathbf{x}_i, S_i)} E\{\mathbf{U}(\mathbf{x}_i, Y_i; \boldsymbol{\beta}) | \mathbf{x}_i\} \right], \end{aligned}$$

which is of the form of (2.4) in Qin et al. (2017). Hence, by Theorem 1 (iii) in Qin et al. (2017), $\widehat{\boldsymbol{\beta}}$ is an asymptotically semiparametric efficient estimator of $\boldsymbol{\beta}^*$. □

C.3 Proof of Theorem 4.2.2

Let $\mathbf{w}$ be an arbitrary sampling scheme that satisfies $0 < w_i \leqslant 1$ for $i = 1, \ldots, N_0$ and $\sum_{i=1}^{N_0} w_i \leqslant N_0\pi_0$. If $\breve{\boldsymbol{\beta}}_{\mathbf{w}}$ converges to $\boldsymbol{\beta}^\dagger \neq \boldsymbol{\beta}^*$, then by (C.2) and (C.3), we can write the

classification mean squared error of $p(\mathbf{X}^*, \breve{\boldsymbol{\beta}}_{\mathbf{w}})$ as

$$E[\{Y^* - p(\mathbf{X}^*, \breve{\boldsymbol{\beta}}_{\mathbf{w}})\}^2] = E[\{Y^* - p_0(\mathbf{X}^*)\}^2] + E[\{p_0(\mathbf{X}^*) - p(\mathbf{X}^*, \boldsymbol{\beta}^\dagger)\}^2] + o(1).$$

By Condition (C1),

$$\begin{aligned} & E[\{Y^* - p(\mathbf{X}^*, \breve{\boldsymbol{\beta}}_{\mathbf{w}})\}^2] - E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}^2] \\ = \; & E[\{p_0(\mathbf{X}^*) - p(\mathbf{X}^*, \boldsymbol{\beta}^\dagger)\}^2] - E[\{p_0(\mathbf{X}^*) - p(\mathbf{X}^*, \boldsymbol{\beta}^*)\}^2] + o(1) > 0 \end{aligned}$$

when $N \to \infty$, hence the result holds.

Now, consider the case that $\breve{\boldsymbol{\beta}}_{\mathbf{w}}$ is a $\sqrt{N}$-consistent estimator of $\boldsymbol{\beta}^*$. Recall that

$$\mathbf{A} \equiv E[\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}^*, \boldsymbol{\beta}^*)^{\otimes 2} + \mathbf{p}''_{\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}}(\mathbf{X}^*, \boldsymbol{\beta}^*)\{p(\mathbf{X}^*, \boldsymbol{\beta}^*) - p_0(\mathbf{X}^*)\}] = E\{\mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{X}^*, \boldsymbol{\beta}^*)^{\otimes 2}\}$$

is positive definite under Condition (C2) and $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$. In addition, $N\{\mathrm{var}(\breve{\boldsymbol{\beta}}_{\mathbf{w}}) - \mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}$ is positive definite by Lemma C.2.1, hence $\mathrm{trace}[\mathbf{A}N\{\mathrm{var}(\breve{\boldsymbol{\beta}}_{\mathbf{w}}) - \mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}] > 0$. By (C.2) and (C.3), we obtain

$$\begin{aligned} & N(E[\{Y^* - p(\mathbf{X}^*, \breve{\boldsymbol{\beta}}_{\mathbf{w}})\}^2] - E[\{Y^* - p(\mathbf{X}^*, \widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}^2]) \\ = \; & N\mathrm{trace}\{\mathbf{A}\mathrm{var}(\breve{\boldsymbol{\beta}}_{\mathbf{w}})\} - N\mathrm{trace}\{\mathbf{A}\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}})\} + O(N^{-1/2}) \\ = \; & \mathrm{trace}[\mathbf{A}N\{\mathrm{var}(\breve{\boldsymbol{\beta}}_{\mathbf{w}}) - \mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{w}})\}] + O(N^{-1/2}) \\ \geqslant \; & 0 \end{aligned}$$

as $N \to \infty$, which gives the result. □

C.4 Proof of Corollary 4.2.2

We will prove the result by showing that Theorem 4.2.1 still holds when we replace the condition $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ by that $\widetilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}^*$. In fact, note that in the proof of Theorem 4.2.1, the assumption $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ is only used to show that $\mathbf{w}_{\mathrm{opt}}$ minimizes the leading term of $\mathrm{trace}(\mathbf{B}_\pi \mathbf{A}^{-1})$ in (C.5). Thus it suffices to show that the leading term of $\mathrm{trace}(\mathbf{B}_\pi \mathbf{A}^{-1})$ remains unchanged when $\boldsymbol{\beta}^*$ is replaced by $\widetilde{\boldsymbol{\beta}}$. Indeed, when $p(\mathbf{x}, \boldsymbol{\beta}^*) = p_0(\mathbf{x})$ and $\widetilde{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}^*$ in probability, we get from (C.5) that

$$\mathrm{trace}(\mathbf{B}_\pi \mathbf{A}^{-1})$$

$$
\begin{aligned}
& -N^{-1} \sum_{i=N_0+1}^{N} \{p^2(\mathbf{x}_i, \boldsymbol{\beta}^*) + p_0(\mathbf{x}_i) - 2p_0(\mathbf{x}_i)p(\mathbf{x}_i, \boldsymbol{\beta}^*)\} \mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}^*) \mathbf{A}_N^{*-1} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*) \\
= \; & N^{-1} \sum_{i=1}^{N_0} \frac{p^2(\mathbf{x}_i, \boldsymbol{\beta}^*) + p_0(\mathbf{x}_i) - 2p_0(\mathbf{x}_i)p(\mathbf{x}_i, \boldsymbol{\beta}^*)}{w_i} \mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}^*) \mathbf{A}_N^{*-1} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*) + o_p(1) \\
= \; & N^{-1} \sum_{i=1}^{N_0} \frac{p(\mathbf{x}_i, \boldsymbol{\beta}^*)\{1 - p(\mathbf{x}_i, \boldsymbol{\beta}^*)\}}{w_i} \mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}^*) \mathbf{A}_N^{-1} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*) + o_p(1) \\
= \; & N^{-1} \sum_{i=1}^{N_0} \frac{p(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})\{1 - p(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})\}}{w_i} \mathbf{p}'_{\boldsymbol{\beta}}{}^{\mathrm{T}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}}) \widetilde{\mathbf{A}}_N^{-1} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}}) + o_p(1),
\end{aligned}
$$

where $\widetilde{\mathbf{A}}_N \equiv N^{-1} \sum_{i=1}^{n} \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})^{\otimes 2}$. The last equality holds by

$$
\begin{aligned}
\widetilde{\mathbf{A}}_N &= N^{-1} \sum_{i=1}^{n} \left\{ \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \widetilde{\boldsymbol{\beta}})^{\otimes 2} \right\} \\
&= N^{-1} \sum_{i=1}^{n} \left\{ \mathbf{p}'_{\boldsymbol{\beta}}(\mathbf{x}_i, \boldsymbol{\beta}^*)^{\otimes 2} \right\} + o_p(1) \\
&= \mathbf{A}_N + o_p(1),
\end{aligned}
$$

therefore $\widetilde{\mathbf{A}}_N^{-1} = \mathbf{A}_N^{-1} + o_p(1)$ by Condition (C2). Hence the result holds. □

Appendix D
Supplement to Chapter 5

D.1 Proof of Lemma 5.2.1

Suppose we have two different sets of models: $g(\mathbf{x}, y)$, $p_Y(y)$, $\rho(y)$ and $\widetilde{g}(\mathbf{x}, y)$, $\widetilde{p}_Y(y)$, $\widetilde{\rho}(y)$. From the likelihood (5.1), it is obvious that

$$\begin{aligned} g(\mathbf{x}, y)p_Y(y) &= \widetilde{g}(\mathbf{x}, y)\widetilde{p}_Y(y), \\ \int g(\mathbf{x}, y)p_Y(y)\rho(y)dy &= \int \widetilde{g}(\mathbf{x}, y)\widetilde{p}_Y(y)\widetilde{\rho}(y)dy. \end{aligned}$$

Integrate with respect to $\mathbf{x}$ on both sides of the first equation above, it is apparent that $p_Y(y) = \widetilde{p}_Y(y)$ and $g(\mathbf{x}, y) = \widetilde{g}(\mathbf{x}, y)$. Therefore, the second equation above becomes

$$\begin{aligned} &\int g(\mathbf{x}, y)p_Y(y)\rho(y)dy = \int g(\mathbf{x}, y)p_Y(y)\widetilde{\rho}(y)dy \\ = &\int p_{Y|\mathbf{X}}(y \mid \mathbf{x})p_{\mathbf{X}}(\mathbf{x})\rho(y)dy = \int p_{Y|\mathbf{X}}(y \mid \mathbf{x})p_{\mathbf{X}}(\mathbf{x})\widetilde{\rho}(y)dy. \end{aligned}$$

Hence, $\int p_{Y|\mathbf{X}}(y \mid \mathbf{x})\rho(y)dy = \int p_{Y|\mathbf{X}}(y \mid \mathbf{x})\widetilde{\rho}(y)dy$. Using the completeness condition, clearly $\rho(y) = \widetilde{\rho}(y)$. □

D.2 Derivation of influence functions

We first state and prove a result regarding the tangant space.

Proposition D.2.1. *The tangent space of (5.1) is* $\mathcal{T} \equiv \mathcal{T}_{\boldsymbol{\alpha}} \oplus (\mathcal{T}_{\boldsymbol{\beta}} + \mathcal{T}_{\boldsymbol{\gamma}})$*, where*

$$\mathcal{T}_{\boldsymbol{\alpha}} = \left[r\mathbf{a}_1(y) : \mathrm{E}_p\{\mathbf{a}_1(Y)\} = \mathbf{0}\right],$$

$$
\begin{aligned}
\mathcal{T}_{\boldsymbol{\beta}} &= [r\mathbf{a}_2(\mathbf{x},y) + (1-r)\mathrm{E}_q\{\mathbf{a}_2(\mathbf{x},Y) \mid \mathbf{x}\} : E\{\mathbf{a}_2(\mathbf{X},y) \mid y\} = \mathbf{0}], \\
\mathcal{T}_{\boldsymbol{\gamma}} &= [(1-r)\mathrm{E}_q\{\mathbf{a}_3(Y) \mid \mathbf{x}\} : \mathrm{E}_q\{\mathbf{a}_3(Y)\} = \mathbf{0}].
\end{aligned}
$$

Proof. Consider a parametric submodel of (5.1),

$$
f_{\mathbf{X},R,RY}(\mathbf{x},r,ry,\boldsymbol{\delta}) = \pi^r(1-\pi)^{1-r}\{g(\mathbf{x},y,\boldsymbol{\beta})p_Y(y,\boldsymbol{\alpha})\}^r \left\{\int g(\mathbf{x},y,\boldsymbol{\beta})q_Y(y,\boldsymbol{\gamma})dy\right\}^{1-r} \quad \text{(D.1)}
$$

where $\boldsymbol{\delta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. We can derive that the score function associated with an arbitrary $\boldsymbol{\delta}$ is $\mathbf{S}_{\boldsymbol{\delta}} \equiv (\mathbf{S}_{\boldsymbol{\alpha}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, where

$$
\begin{aligned}
\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x},r,ry) &\equiv r\mathbf{a}_{\boldsymbol{\alpha}}(y), \\
\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x},r,ry) &\equiv r\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{x},y) + (1-r)\mathrm{E}_q\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{x},Y) \mid \mathbf{x}\}, \\
\mathbf{S}_{\boldsymbol{\gamma}}(\mathbf{x},r,ry) &\equiv (1-r)\mathrm{E}_q\{\mathbf{a}_{\boldsymbol{\gamma}}(Y) \mid \mathbf{x}\},
\end{aligned}
$$

$\mathrm{E}_p\{\mathbf{a}_{\boldsymbol{\alpha}}(Y)\} = \mathbf{0}$, $\mathrm{E}_p\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X},y)|y\} = \mathrm{E}_q\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X},y)|y\} = \mathrm{E}\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X},y)|y\} = \mathbf{0}$, and $\mathrm{E}_q\{\mathbf{a}_{\boldsymbol{\gamma}}(Y)\} = \mathbf{0}$. The above derivation directly leads to Proposition D.2.1. □

We are now ready to establish the result regarding $\mathcal{F}$, which we explicitly write out as Proposition D.2.2.

Proposition D.2.2. *The set of the influence functions for θ is*

$$
\mathcal{F} \equiv \left[\frac{r}{\pi}\{y\rho(y) - b(\mathbf{x})\rho(y) + c\} + \frac{1-r}{1-\pi}\{b(\mathbf{x}) - \theta - c\} : E\{b(\mathbf{X}) \mid y\} = y, \forall c\right]. \quad \text{(D.2)}
$$

Proof. Note that $\theta = \mathrm{E}_q(Y)$, hence

$$
\begin{aligned}
\frac{\partial\theta}{\partial\boldsymbol{\alpha}^{\mathrm{T}}} &= \mathbf{0}^{\mathrm{T}}, \\
\frac{\partial\theta}{\partial\boldsymbol{\beta}^{\mathrm{T}}} &= \mathbf{0}^{\mathrm{T}}, \\
\frac{\partial\theta}{\partial\boldsymbol{\gamma}^{\mathrm{T}}} &= \mathrm{E}_q\{Y\mathbf{a}_{\boldsymbol{\gamma}}^{\mathrm{T}}(Y)\}.
\end{aligned}
$$

Let $\phi(\mathbf{x},r,ry)$ be

$$
\phi(\mathbf{x},r,ry) \equiv \frac{r}{\pi}\phi_1(\mathbf{x},y) + \frac{1-r}{1-\pi}\phi_2(\mathbf{x}).
$$

For $\phi(\mathbf{x}, r, ry)$ to be an influence function, it must satisfy

$$\mathrm{E}(\phi) = \mathrm{E}_p\{\phi_1(\mathbf{X}, Y)\} + \mathrm{E}_q\{\phi_2(\mathbf{X})\} = 0 \tag{D.3}$$

and $\mathrm{E}(\phi \mathbf{S}_{\boldsymbol{\delta}}^{\mathrm{T}}) = \partial\theta/\partial\boldsymbol{\delta}^{\mathrm{T}}$. $\mathrm{E}(\phi \mathbf{S}_{\boldsymbol{\alpha}}^{\mathrm{T}}) = \partial\theta/\partial\boldsymbol{\alpha}^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$ is equivalent to

$$\mathrm{E}\{\phi_1(\mathbf{X}, y) \mid y\} = c \tag{D.4}$$

for some constant c. In addition, since

$$\begin{aligned} \mathrm{E}(\phi \mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}) &= \mathrm{E}_p\{\phi_1(\mathbf{X}, Y)\mathbf{a}_{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{X}, Y)\} + \mathrm{E}_q\{\phi_2(\mathbf{X})\mathbf{a}_{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{X}, Y)\} \\ &= \mathrm{E}_p[\{\phi_1(\mathbf{X}, Y) + \phi_2(\mathbf{X})\rho(Y)\}\mathbf{a}_{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{X}, Y)] \end{aligned}$$

and $\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{x}, y)$ is an arbitrary function which satisfies $\mathrm{E}\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, y)|y\} = \mathbf{0}$, $\mathrm{E}(\phi \mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}) = \partial\theta/\partial\boldsymbol{\beta}^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$ implies $\phi_1(\mathbf{x}, y) + \phi_2(\mathbf{x})\rho(y) = a(y)$ for some function $a(y)$. Then (D.4) yields

$$\phi_1(\mathbf{x}, y) = [\mathrm{E}\{\phi_2(\mathbf{X}) \mid y\} - \phi_2(\mathbf{x})]\rho(y) + c.$$

Also, noting that $\mathrm{E}(\phi \mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}}) = \partial\theta/\partial\boldsymbol{\gamma}^{\mathrm{T}} = \mathrm{E}_q\{Y\mathbf{a}_{\boldsymbol{\gamma}}^{\mathrm{T}}(Y)\}$ is equivalent to $\mathrm{E}\{\phi_2(\mathbf{X}) \mid y\} = y + c^*$ for some constant c^*, we have from (D.3) and (D.4) that

$$\mathrm{E}\{\phi_2(\mathbf{X}) \mid y\} = y - \mathrm{E}_q(Y) - c.$$

Therefore, defining $b(\mathbf{x}) \equiv \phi_2(\mathbf{x}) + \mathrm{E}_q(Y) + c$, the summary description of the influence function is

$$\phi(\mathbf{x}, r, ry) = \frac{r}{\pi}\{y\rho(y) - b(\mathbf{x})\rho(y) + c\} + \frac{1-r}{1-\pi}\{b(\mathbf{x}) - \mathrm{E}_q(Y) - c\},$$

where $b(\mathbf{x})$ satisfies $\mathrm{E}\{b(\mathbf{X}) \mid y\} = y$ and c is a constant. Hence, we get the result (D.2). □

D.3 Proof of Proposition 5.3.1

Note that

$$\frac{\mathrm{E}_p\{a(\mathbf{x}, Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho(Y) \mid \mathbf{x}\}} = \frac{\int a(\mathbf{x}, y)\rho(y)g(\mathbf{x}, y)p_Y(y)dy}{\int \rho(y)g(\mathbf{x}, y)p_Y(y)dy} = \frac{\int a(\mathbf{x}, y)g(\mathbf{x}, y)q_Y(y)dy}{\int g(\mathbf{x}, y)q_Y(y)dy}$$

$$= \quad \mathrm{E}_q\{a(\mathbf{x}, Y) \mid \mathbf{x}\},$$

then $\phi_{\text{eff}}(\mathbf{x}, r, ry)$ can be alternatively written as

$$\begin{aligned} & \phi_{\text{eff}}(\mathbf{x}, r, ry) \\ = \quad & \frac{r}{\pi}\rho(y)\left[y - \frac{\mathrm{E}_q\{a(Y) \mid \mathbf{x}\}}{\mathrm{E}_q\{\rho(Y) \mid \mathbf{x}\} + \pi/(1-\pi)}\right] + \frac{1-r}{1-\pi}\left[\frac{\mathrm{E}_q\{a(Y) \mid \mathbf{x}\}}{\mathrm{E}_q\{\rho(Y) \mid \mathbf{x}\} + \pi/(1-\pi)} - \theta\right], \end{aligned}$$

where $a(y)$ satisfies

$$\mathrm{E}\left[\frac{\mathrm{E}_q\{a(Y) \mid \mathbf{X}\}}{\mathrm{E}_q\{\rho(Y) \mid \mathbf{X}\} + \pi/(1-\pi)} \mid y\right] = y.$$

First, it is immediate that $\phi_{\text{eff}}(\mathbf{x}, r, ry)$ is an influence function for θ, i.e., belongs to $\mathcal{F}$ given in (D.2) from letting

$$\begin{aligned} b(\mathbf{x}) \quad &\equiv \quad \frac{\mathrm{E}_q\{a(Y) \mid \mathbf{x}\}}{\mathrm{E}_q\{\rho(Y) \mid \mathbf{x}\} + \pi/(1-\pi)}, \\ c \quad &\equiv \quad 0. \end{aligned}$$

Next, we show that $\phi_{\text{eff}}(\mathbf{x}, r, ry)$ is in the tangent space $\mathcal{T}$ of (D.1). We decompose $\phi_{\text{eff}}(\mathbf{x}, r, ry)$ into

$$\begin{aligned} \phi_{\text{eff}}(\mathbf{x}, r, ry) \quad &= \quad \frac{r}{\pi}\{y\rho(y) - b(\mathbf{x})\rho(y)\} + \frac{1-r}{1-\pi}\{b(\mathbf{x}) - \theta\} \\ &= \quad r\{a_1(y) + a_2(\mathbf{x}, y)\} + (1-r)[\mathrm{E}_q\{a_2(\mathbf{x}, Y) \mid \mathbf{x}\} + \mathrm{E}_q\{a_3(Y) \mid \mathbf{x}\}], \end{aligned}$$

where

$$\begin{aligned} a_1(y) \quad &\equiv \quad 0, \\ a_2(\mathbf{x}, y) \quad &\equiv \quad \frac{1}{\pi}\{y\rho(y) - b(\mathbf{x})\rho(y)\}, \\ a_3(y) \quad &\equiv \quad \frac{1}{\pi}\{a(y) - y\rho(y)\} - \frac{1}{1-\pi}\theta. \end{aligned}$$

Then it is easy to show $ra_1(y) \in \mathcal{T}_{\boldsymbol{\alpha}}$, and $ra_2(\mathbf{x}, y) + (1-r)\mathrm{E}_q\{a_2(\mathbf{x}, Y) \mid \mathbf{x}\} \in \mathcal{T}_{\boldsymbol{\beta}}$ because $\mathrm{E}\{b(\mathbf{X}) \mid y\} = y$. Further, $(1-r)\mathrm{E}_q\{a_3(Y) \mid \mathbf{x}\} \in \mathcal{T}_{\boldsymbol{\gamma}}$ since

$$\begin{aligned} \mathrm{E}_q\{a_3(Y)\} \quad &= \quad \frac{1}{\pi}\mathrm{E}_q[\mathrm{E}_q\{a(Y) \mid \mathbf{X}\}] - \frac{1}{\pi}\mathrm{E}_q\{Y\rho(Y)\} - \frac{1}{1-\pi}\theta \\ &= \quad \frac{1}{\pi}\mathrm{E}_q[b(\mathbf{X})\mathrm{E}_q\{\rho(Y) \mid \mathbf{X}\}] + \frac{1}{1-\pi}\mathrm{E}_q\{b(\mathbf{X})\} - \frac{1}{\pi}\mathrm{E}_q\{Y\rho(Y)\} - \frac{1}{1-\pi}\theta \end{aligned}$$

$$= \quad 0.$$

Hence $\phi_{\text{eff}}(\mathbf{x}, r, ry)$ belongs to $\mathcal{T}$, which proves the result. □

D.4 Proof of Proposition 5.3.2

The definition of $b^{**}(\mathbf{x})$ immediately leads to $\mathrm{E}\{b^{**}(\mathbf{X}) \mid y\} = y$. Therefore,

$$\begin{aligned}
\widehat{\theta}_t &\xrightarrow{P} \mathrm{E}\left[\frac{R}{\pi}\rho^*(Y)\{Y - b^{**}(\mathbf{X})\} + \frac{1-R}{1-\pi}b^{**}(\mathbf{X})\right] \\
&= \mathrm{E}_p[\rho^*(Y)\{Y - b^{**}(\mathbf{X})\}] + \mathrm{E}_q\{b^{**}(\mathbf{X})\} \\
&= \mathrm{E}_p[\rho^*(Y)[Y - \mathrm{E}\{b^{**}(\mathbf{X}) \mid Y\}]) + \mathrm{E}_q[\mathrm{E}\{b^{**}(\mathbf{X}) \mid Y\}] \\
&= \mathrm{E}_q(Y).
\end{aligned}$$

□

D.5 Algorithms for solving equations (5.8) and (5.11)

We first illustrate how to solve (5.8). Let

$$\phi^{**}(y,t) \equiv \rho^*(t)\sum_{i=1}^{n} \frac{p^{\star}_{Y|\mathbf{X}}(t, \mathbf{x}_i)}{\mathrm{E}^{\star}_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}^{\star}_p\{\rho^*(Y) \mid \mathbf{x}_i\}} \frac{r_i K_h(y - y_i)}{\sum_{j=1}^{n} r_j K_h(y - y_j)}, \quad \text{(D.5)}$$

then (5.8) is equivalently written as $y = \int \phi^{**}(y,t)a^{**}(t)dt$, which is a Fredholm integral equation of the first type. We can find the solution $\widehat{a}^{**}$ through Landweber's iterative method (Landweber 1951). Choose a starting point a_0, then iterate the formula

$$a_{k+1}(y) \leftarrow a_k(y) + \int \phi^{**}(t,y)tdt - \int \left\{ \int \phi^{**}(z,y)\phi^{**}(z,t)dz \right\} a_k(t)dt$$

while $\int\{a_{k+1}(t) - a_k(t)\}^2 dt / \int\{a_k(t)\}^2 dt$ is greater than a chosen tolerance.

For practical implementation, we can approximate the integration in (5.8) using quadrature method. For a given quadrature rule, let the quadrature nodes be $\mathbf{t} \equiv (t_1, \ldots, t_m)^{\mathrm{T}}$ and weights be $\mathbf{w} \equiv (w_1, \ldots, w_m)^{\mathrm{T}}$. Also, let $\widetilde{\mathbf{y}} \equiv (\widetilde{y}_1, \ldots, \widetilde{y}_l)^{\mathrm{T}}, \mathbf{a}^{**} \equiv \{a^{**}(t_1), \ldots, a^{**}(t_m)\}^{\mathrm{T}}$, and $\mathbf{\Phi}^{**}$ be an $l \times m$ matrix whose (i,j) component is $\phi^{**}(\widetilde{y}_i, t_j)$. Then (5.8) can be discretized as $\widetilde{\mathbf{y}} = \mathbf{\Phi}^{**}(\mathbf{a}^{**} \cdot \mathbf{w})$ where $\mathbf{a} \cdot \mathbf{b} \equiv (a_1 b_1, \ldots, a_m b_m)^{\mathrm{T}}$ is the

Hadamard product. Also, the above iterative formula can be approximated as

$$\mathbf{a}_{k+1} \leftarrow \mathbf{a}_k + l^{-1}\mathbf{\Phi}^{\star\star\mathrm{T}}\widetilde{\mathbf{y}} - l^{-1}\mathbf{\Phi}^{\star\star\mathrm{T}}\mathbf{\Phi}^{\star\star}(\mathbf{a}_k \cdot \mathbf{w}).$$

We iterately use this formula with the starting point $\mathbf{a}_0$ as long as$\|\mathbf{a}_{k+1} - \mathbf{a}_k\|_2^2/\|\mathbf{a}_k\|_2^2$ is greater than a chosen tolerance. We summarize the algorithm in Algorithm 3.

Algorithm 3 Solving the integral equation (5.8)

Input: a function $\phi^{\star\star}(y,t)$ in (D.5) and a tolerance Δ.

do

(a) adopt a set of evaluation points $\widetilde{\mathbf{y}} = (\widetilde{y}_1, \ldots, \widetilde{y}_l)^{\mathrm{T}}$;

(b) adopt a quadrature rule $(\mathbf{t}, \mathbf{w})$, where nodes $\mathbf{t} = (t_1, \ldots, t_m)^{\mathrm{T}}$ and weights $\mathbf{w} = (w_1, \ldots, w_m)^{\mathrm{T}}$;

(c) compute a matrix $\mathbf{\Phi}^{\star\star}$ where $\mathbf{\Phi}^{\star\star}_{(i,j)} = \phi^{\star\star}(\widetilde{y}_i, t_j), i = 1, \ldots, l, j = 1, \ldots, m$;

(d) Declare a starting point $\mathbf{a}_0 = (a_{01}, \ldots, a_{0m})^{\mathrm{T}}$;

while $\|\mathbf{a}_{k+1} - \mathbf{a}_k\|_2^2/\|\mathbf{a}_k\|_2^2 > \Delta$ **do**

| $\mathbf{a}_{k+1} \leftarrow \mathbf{a}_k + l^{-1}\mathbf{\Phi}^{\star\star\mathrm{T}}\widetilde{\mathbf{y}} - l^{-1}\mathbf{\Phi}^{\star\star\mathrm{T}}\mathbf{\Phi}^{\star\star}(\mathbf{a}_k \cdot \mathbf{w})$;

end

Output: $\widehat{\mathbf{a}}^{\star\star} \leftarrow \mathbf{a}_{k+1}$.

We now describe how to solve (5.11). If $\widehat{\mathrm{E}}_p\{a(Y) \mid \mathbf{x}\} = \int a(t)\widehat{p}_{Y|\mathbf{X}}(t, \mathbf{x}_i)dt$ for some $\widehat{p}_{Y|\mathbf{X}}$, then we can follow the above procedure to obtain the solution while replacing $p^{\star}_{Y|\mathbf{X}}(t, \mathbf{x}_i)$ in (D.5) by $\widehat{p}_{Y|\mathbf{X}}(t, \mathbf{x}_i)$. Now, suppose $\widehat{\mathrm{E}}_p\{a(Y) \mid \mathbf{x}\} = \sum_{i=1}^{n_1} a(y_i)w_i(\mathbf{x})$ for some $w_i, i = 1, \ldots, n_1$. This form includes a general class of nonparametric regression estimators, for instance, the Nadaraya-Watson estimator is of this form with $w_i(\mathbf{x}) = r_i K_h(\mathbf{x} - \mathbf{x}_i)/\sum_{j=1}^{n} r_j K_h(\mathbf{x} - \mathbf{x}_j)$. Then (5.11) is equivalent to

$$\begin{aligned} y &= \sum_{i=1}^{n} \frac{\sum_{k=1}^{n_1} a^*(y_k)\rho^*(y_k)w_k(\mathbf{x}_i)}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y) \mid \mathbf{x}_i\}} \frac{r_i K_h(y - y_i)}{\sum_{j=1}^{n} r_j K_h(y - y_j)} \\ &= \sum_{k=1}^{n_1} \phi^*(y, y_k, w_k)a^*(y_k), \end{aligned}$$

where

$$\phi^*(y, y_k, w_k) \equiv \sum_{i=1}^{n} \frac{\rho^*(y_k)w_k(\mathbf{x}_i)}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y) \mid \mathbf{x}_i\}} \frac{r_i K_h(y - y_i)}{\sum_{j=1}^{n} r_j K_h(y - y_j)}. \quad \text{(D.6)}$$

Now, let $\widetilde{\mathbf{y}} \equiv (\widetilde{y}_1, \ldots, \widetilde{y}_l)^{\mathrm{T}}, \mathbf{a}^* \equiv \{a^*(y_1), \ldots, a^*(y_{n_1})\}^{\mathrm{T}}$, and $\mathbf{\Phi}^*$ be a $l \times n_1$ matrix whose

(i,j)th component is $\phi^*(\widetilde{y}_i, y_j, w_j)$. Then we can find the approximate solution $\widehat{\mathbf{a}}^*$ via iteratively using the formula

$$\mathbf{a}_{k+1} \leftarrow \mathbf{a}_k + l^{-1}\mathbf{\Phi}^{*\mathrm{T}}\widetilde{\mathbf{y}} - l^{-1}\mathbf{\Phi}^{*\mathrm{T}}\mathbf{\Phi}^*\mathbf{a}_k$$

with a starting point $\mathbf{a}_0$, as long as $\|\mathbf{a}_{k+1} - \mathbf{a}_k\|_2^2/\|\mathbf{a}_k\|_2^2$ is greater than a chosen tolerance. We summarize the algorithm in Algorithm 4.

Algorithm 4 Solving the integral equation (5.11)

Input: a function $\phi^*(y, y_k, w_k)$ in (D.6) and a tolerance Δ.

do

(a) adopt a set of evaluation points $\widetilde{\mathbf{y}} = (\widetilde{y}_1, \ldots, \widetilde{y}_l)^{\mathrm{T}}$;

(b) compute a matrix $\mathbf{\Phi}^*$ where $\mathbf{\Phi}^*_{(i,j)} = \phi^*(\widetilde{y}_i, y_j, w_j), i = 1, \ldots, l, j = 1, \ldots, n_1$;

(c) Declare a starting point $\mathbf{a}_0 = (a_{01}, \ldots, a_{0m})^{\mathrm{T}}$;

while $\|\mathbf{a}_{k+1} - \mathbf{a}_k\|_2^2/\|\mathbf{a}_k\|_2^2 > \Delta$ **do**

| $\mathbf{a}_{k+1} \leftarrow \mathbf{a}_k + l^{-1}\mathbf{\Phi}^{*\mathrm{T}}\widetilde{\mathbf{y}} - l^{-1}\mathbf{\Phi}^{*\mathrm{T}}\mathbf{\Phi}^*\mathbf{a}_k$;

end

Output: $\widehat{\mathbf{a}}^* \leftarrow \mathbf{a}_{k+1}$.

D.6 Proof of Lemma 5.3.1

From the definition of $\mathcal{L}^{**}(a)$, Conditions (A3), and (A4), it is immediate that there exists a constant $0 < c_2 < \infty$ such that $\|\mathcal{L}^{**}(a)\|_\infty \leqslant c_2\|a\|_\infty$. Now we show there is a constant $0 < c_1 < \infty$ such that $c_1\|a\|_\infty \leqslant \|\mathcal{L}^{**}(a)\|_\infty$. Note that if $\mathcal{L}^{**}$ is invertible, by the bounded inverse theorem we have $\|\mathcal{L}^{**-1}(v)\|_\infty \leqslant c_1^{-1}\|v\|_\infty$ for some constant $0 < c_1 < \infty$, i.e., $c_1\|a\|_\infty \leqslant \|\mathcal{L}^{**}(a)\|_\infty$. Hence it suffices to show that $\mathcal{L}^{**}$ is invertible. We prove this by contradiction. Suppose there are $a_1(y)$ and $a_2(y)$ such that $\mathcal{L}^{**}(a_1)(y) = \mathcal{L}^{**}(a_2)(y) = v(y)$ and $a_1 \neq a_2$. Then by Conditions (A1) and (A2), we have

$$\mathrm{E}_p^\star\{a_1(Y)\rho^*(Y) \mid \mathbf{x}\} \neq \mathrm{E}_p^\star\{a_2(Y)\rho^*(Y) \mid \mathbf{x}\}.$$

Now, the efficient score calculated under the posited models is

$$\begin{aligned}\phi^{**}_{\mathrm{eff}}(\mathbf{x}, r, ry) &= \frac{r}{\pi}\rho^*(y)\left[y - \frac{\mathrm{E}_p^\star\{a(Y)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}\}}\right] \\ &\quad + \frac{1-r}{1-\pi}\left[\frac{\mathrm{E}_p^\star\{a(Y)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}\}} - \theta\right],\end{aligned}$$

where $a(y)$ satisfies $\mathcal{L}^{**}(a)(y) = v(y)$. Then letting $a = a_1$ and $a = a_2$ gives two distinct efficient scores, which contradicts the uniqueness of the efficient score. Therefore, there is a unique solution $a^{**}(y)$ for $\mathcal{L}^{**}(a^{**})(y) = v(y)$, hence $\mathcal{L}^{**}$ is invertible. □

D.7 Proof of Theorem 5.3.1

We define

$$b^{**}(\mathbf{x}, a, \boldsymbol{\zeta}) \equiv \frac{\mathrm{E}_p^{\star}\{a(Y)\rho^*(Y) \mid \mathbf{x}, \boldsymbol{\zeta}\}}{\mathrm{E}_p^{\star}\{\rho^{*2}(Y) \mid \mathbf{x}, \boldsymbol{\zeta}\} + \pi/(1-\pi)\mathrm{E}_p^{\star}\{\rho^*(Y) \mid \mathbf{x}, \boldsymbol{\zeta}\}},$$

and analyze $b^{**}(\mathbf{x}, \widehat{a}^{**}, \widehat{\boldsymbol{\zeta}})$. First, under Conditions (A3)-(A6) and $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$, we have

$$\begin{aligned}
& \widehat{\mathcal{L}}^{**}(a^{**})(y) \\
= \; & n_1^{-1} \sum_{i=1}^{n} r_i K_h(y - y_i) b^{**}(\mathbf{x}_i, a^{**}, \widehat{\boldsymbol{\zeta}}) \\
= \; & n_1^{-1} \sum_{i=1}^{n} r_i K_h(y - y_i) \left\{ b^{**}(\mathbf{x}_i, a^{**}, \boldsymbol{\zeta}) + \frac{\partial b^{**}(\mathbf{x}_i, a^{**}, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2}) \right\} \\
= \; & n_1^{-1} \sum_{i=1}^{n} \mathcal{L}_{i,h}^{**}(a^{**})(y) + \left[p_Y(y) \frac{\partial E\{b^{**}(\mathbf{X}, a^{**}, \boldsymbol{\zeta}) \mid y\}}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} + o_p(1) \right] (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2}) \\
= \; & n_1^{-1} \sum_{i=1}^{n} \mathcal{L}_{i,h}^{**}(a^{**})(y) + o_p(n_1^{-1/2}) \qquad \text{(D.7)}
\end{aligned}$$

uniformly in y since $E\{b^{**}(\mathbf{X}, a^{**}, \boldsymbol{\zeta}) \mid y\} = y$. Here, we define

$$\mathcal{L}_{i,h}^{**}(a)(y) \equiv r_i K_h(y - y_i) \frac{\mathrm{E}_p^{\star}\{a(Y)\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^{\star}\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}_p^{\star}\{\rho^*(Y) \mid \mathbf{x}_i\}}.$$

We also define $g_{i,h}(y) \equiv \mathcal{L}^{**-1}\{v_{i,h} - \mathcal{L}_{i,h}^{**}(a^{**})\}(y)$. Note that $\|E(V_{i,h}) - v\|_\infty = O(h^2)$ by Conditions (A4),(A5), then since $\mathcal{L}^{**-1}$ is a bounded linear operator by Lemma 5.3.1,

$$\begin{aligned}
\|E\{\mathcal{L}^{**-1}(V_{i,h})\} - \mathcal{L}^{**-1}(v)\|_\infty &= \|\mathcal{L}^{**-1}\{E(V_{i,h})\} - \mathcal{L}^{**-1}(v)\|_\infty \\
&= \|\mathcal{L}^{**-1}\{E(V_{i,h}) - v\}\|_\infty \\
&= O(h^2).
\end{aligned}$$

Similarly, $\|E\{(\mathcal{L}^{**-1}\mathcal{L}_{i,h}^{**})(a^{**})\} - a^{**}\|_\infty = O(h^2)$. Then using $\mathcal{L}^{**-1}(v)(y) = a^{**}(y)$, we get

$$\begin{aligned}
\|E(G_{i,h})\|_\infty &= \|E\{\mathcal{L}^{**-1}(V_{i,h})\} - E\{(\mathcal{L}^{**-1}\mathcal{L}_{i,h}^{**})(a^{**})\} - \{\mathcal{L}^{**-1}(v) - a^{**}\}\|_\infty \\
&\leqslant \|E\{\mathcal{L}^{**-1}(V_{i,h})\} - \mathcal{L}^{**-1}(v)\|_\infty + \|E\{(\mathcal{L}^{**-1}\mathcal{L}_{i,h})(a^{**})\} - a^{**}\|_\infty \\
&= O(h^2). \qquad \text{(D.8)}
\end{aligned}$$

Now, for any bounded function $a(y)$, $\|(\widehat{\mathcal{L}}^{**} - \mathcal{L}^{**})(a)\|_\infty = O_p\left\{(n_1 h)^{-1/2}\log n_1 + h^2\right\} = o_p(n_1^{-1/4})$ under Conditions (A3)-(A6) and the assumption $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$. Similarly, $\|\widehat{v} - v\|_\infty = o_p(n_1^{-1/4})$. Then using that $\mathcal{L}^{**-1}$ is a bounded linear operator by Lemma 5.3.1, $\widehat{a}^{**}(y)$ can be expressed as

$$\begin{aligned}
\widehat{a}^{**}(y) &= \widehat{\mathcal{L}}^{**-1}(\widehat{v})(y) \\
&= \{\mathcal{L}^{**} + (\widehat{\mathcal{L}}^{**} - \mathcal{L}^{**})\}^{-1}(\widehat{v})(y) \\
&= \{\mathcal{L}^{**-1} - \mathcal{L}^{**-1}(\widehat{\mathcal{L}}^{**} - \mathcal{L}^{**})\mathcal{L}^{**-1}\}\{v + (\widehat{v} - v)\}(y) + o_p(n_1^{-1/2}) \\
&= a^{**}(y) + \mathcal{L}^{**-1}(\widehat{v} - v)(y) - \{\mathcal{L}^{**-1}(\widehat{\mathcal{L}}^{**} - \mathcal{L}^{**})\}(a^{**})(y) + o_p(n_1^{-1/2}) \\
&= a^{**}(y) + \mathcal{L}^{**-1}\{\widehat{v} - \widehat{\mathcal{L}}^{**}(a^{**})\}(y) + o_p(n_1^{-1/2}) \\
&= a^{**}(y) + n_1^{-1}\sum_{i=1}^{n} g_{i,h}(y) + o_p(n_1^{-1/2})
\end{aligned}$$

uniformly in y, where the last step is by (D.7) and the definition of $g_{i,h}(y)$. Then we further get

$$\begin{aligned}
& \left|\frac{\partial b^{**}(\mathbf{x}, \widehat{a}^{**}, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) - \frac{\partial b^{**}(\mathbf{x}, a^{**}, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})\right| \\
= & \left|\frac{\partial}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}\left[\frac{\mathrm{E}_p^\star\{(\widehat{a}^{**} - a^{**})(Y)\rho^*(Y) \mid \mathbf{x}, \boldsymbol{\zeta}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}, \boldsymbol{\zeta}\}}\right](\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})\right| \\
= & \left|\frac{\mathrm{E}_p^\star\{(\widehat{a}^{**} - a^{**})(Y)\rho^*(Y)\mathbf{S}_{\boldsymbol{\zeta}}^{\star\mathrm{T}}(Y, \mathbf{x}, \boldsymbol{\zeta}) \mid \mathbf{x}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})\right. \\
& \left. -\mathrm{E}_p^\star\{(\widehat{a}^{**} - a^{**})(Y)\rho^*(Y) \mid \mathbf{x}\}\frac{\mathrm{E}_p[\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y)\}\mathbf{S}_{\boldsymbol{\zeta}}^{\star\mathrm{T}}(Y, \mathbf{x}, \boldsymbol{\zeta}) \mid \mathbf{x}]}{[\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}]^2}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})\right| \\
\leqslant & \|\widehat{a}^{**} - a^{**}\|_\infty \frac{\mathrm{E}_p^\star\{\rho^*(Y)\|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y, \mathbf{x}, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}\}\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}} \\
& + \|\widehat{a}^{**} - a^{**}\|_\infty \mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}\}\frac{\mathrm{E}_p^\star[\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y)\}\|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y, \mathbf{x}, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}]\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2}{[\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}]^2}
\end{aligned}$$

$$= \quad o_p(n_1^{-1/4})O_p(n_1^{-1/2}) = o_p(n_1^{-1/2}),$$

because $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$, $\mathrm{E}_p^\star\{\|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y, \mathbf{x}, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}\}$ is bounded, and $\|\widehat{a}^{\star\star} - a^{\star\star}\|_\infty = O_p\left\{(n_1 h)^{-1/2} \log n_1 + h^2\right\} = o_p(n^{-1/4})$ by (D.8) and Condition (A6). Hence, noting that $b^{\star\star}(\mathbf{x}, a, \boldsymbol{\zeta})$ is linear with respect to a, we get

$$\begin{aligned} b^{\star\star}(\mathbf{x}, \widehat{a}^{\star\star}, \widehat{\boldsymbol{\zeta}}) &= b^{\star\star}(\mathbf{x}, \widehat{a}^{\star\star}, \boldsymbol{\zeta}) + \frac{\partial b^{\star\star}(\mathbf{x}, \widehat{a}^{\star\star}, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2}) \qquad \text{(D.9)} \\ &= b^{\star\star}(\mathbf{x}, a^{\star\star}, \boldsymbol{\zeta}) + b^{\star\star}(\mathbf{x}, \widehat{a}^{\star\star} - a^{\star\star}, \boldsymbol{\zeta}) + \frac{\partial b^{\star\star}(\mathbf{x}, a^{\star\star}, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2}) \\ &= b^{\star\star}(\mathbf{x}, a^{\star\star}, \boldsymbol{\zeta}) + n_1^{-1}\sum_{i=1}^{n} b^{\star\star}(\mathbf{x}, g_{i,h}, \boldsymbol{\zeta}) + \frac{\partial b^{\star\star}(\mathbf{x}, a^{\star\star}, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2}) \end{aligned}$$

uniformly in $\mathbf{x}$ by Condition (A4). Now we analyze $\widehat{\theta}$. By the definition of $\widehat{\theta}$,

$$\begin{aligned} & \sqrt{n_1}(\widehat{\theta} - \theta) \\ = \quad & \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \left[\frac{r_i}{\pi} \rho^*(y_i)\{y_i - b^{\star\star}(\mathbf{x}_i, \widehat{a}^{\star\star}, \widehat{\boldsymbol{\zeta}})\} + \frac{1 - r_i}{1 - \pi}\{b^{\star\star}(\mathbf{x}_i, \widehat{a}^{\star\star}, \widehat{\boldsymbol{\zeta}}) - \theta\} \right] \\ = \quad & \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \phi_{\mathrm{eff}}^{\star\star}(\mathbf{x}_i, r_i, r_i y_i) \\ & + \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \left\{ \frac{r_i}{\pi} \rho^*(y_i) - \frac{1 - r_i}{1 - \pi} \right\} \{b^{\star\star}(\mathbf{x}_i, a^{\star\star}, \boldsymbol{\zeta}) - b^{\star\star}(\mathbf{x}_i, \widehat{a}^{\star\star}, \widehat{\boldsymbol{\zeta}})\} \\ = \quad & \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \phi_{\mathrm{eff}}^{\star\star}(\mathbf{x}_i, r_i, r_i y_i) - T_1 - T_2 + o_p(1), \qquad \text{(D.10)} \end{aligned}$$

where

$$\begin{aligned} T_1 &\equiv n_1^{-1/2} n^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} \left\{ \frac{r_i}{\pi} \rho^*(y_i) - \frac{1 - r_i}{1 - \pi} \right\} b^{\star\star}(\mathbf{x}_i, g_{j,h}, \boldsymbol{\zeta}), \\ T_2 &\equiv n^{-1} \sum_{i=1}^{n} \left\{ \frac{r_i}{\pi} \rho^*(y_i) - \frac{1 - r_i}{1 - \pi} \right\} \frac{\partial b^{\star\star}(\mathbf{x}_i, a^{\star\star}, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} \sqrt{n_1}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}). \end{aligned}$$

Note that since $\|E(G_{j,h})\|_\infty = O(h^2)$ by (D.8) and $b^{\star\star}(\mathbf{x}, a, \boldsymbol{\zeta})$ is linear with respect to a,

$$\|E\{b^{\star\star}(\mathbf{x}, G_{j,h}, \boldsymbol{\zeta})\}\|_\infty = \|b^{\star\star}\{\mathbf{x}, E(G_{j,h}), \boldsymbol{\zeta}\}\|_\infty = O(h^2). \qquad \text{(D.11)}$$

Hence using the property of the U-statistic and Condition (A6), we can rewrite T_1 as

$$
\begin{aligned}
& T_1 \\
= \; & n_1^{-1/2} \sum_{i=1}^{n} \left\{ \frac{r_i}{\pi} \rho^*(y_i) - \frac{1-r_i}{1-\pi} \right\} E\{ b^{**}(\mathbf{x}_i, G_{j,h}, \boldsymbol{\zeta}) \mid \mathbf{x}_i, r_i, r_i y_i \} \\
& + n_1^{-1/2} \sum_{j=1}^{n} \mathrm{E} \left[\left\{ \frac{R_i}{\pi} \rho^*(Y_i) - \frac{1-R_i}{1-\pi} \right\} b^{**}(\mathbf{X}_i, g_{j,h}, \boldsymbol{\zeta}) \mid \mathbf{x}_j, r_j, r_j y_j \right] \\
& - n^{1/2} n_1^{-1/2} \mathrm{E} \left[\left\{ \frac{R_i}{\pi} \rho^*(Y_i) - \frac{1-R_i}{1-\pi} \right\} b^{**}(\mathbf{X}_i, G_{j,h}, \boldsymbol{\zeta}) \right] + O_p(n_1^{-1/2}) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} \mathrm{E} \left[\left\{ \frac{R_i}{\pi} \rho^*(Y_i) - \frac{1-R_i}{1-\pi} \right\} b^{**}(\mathbf{X}_i, g_{j,h}, \boldsymbol{\zeta}) \mid \mathbf{x}_j, r_j, r_j y_j \right] + O_p(n_1^{-1/2} + n n_1^{-1/2} h^2) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} \int [\rho^*(y) \mathrm{E}\{ b^{**}(\mathbf{X}, g_{j,h}, \boldsymbol{\zeta}) \mid y \} p_Y(y) - \mathrm{E}\{ b^{**}(\mathbf{X}, g_{j,h}, \boldsymbol{\zeta}) \mid y \} q_Y(y)] \, dy + o_p(1) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} \int \mathcal{L}^{**}(g_{j,h})(y) \{ \rho^*(y) - \rho(y) \} \, dy + o_p(1) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} \int \{ v_{j,h}(y) - \mathcal{L}_{j,h}^{**}(a^{**})(y) \} \{ \rho^*(y) - \rho(y) \} \, dy + o_p(1) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} r_j \int K_h(y - y_j) \{ y - b^{**}(\mathbf{x}_j, a^{**}, \boldsymbol{\zeta}) \} \{ \rho^*(y) - \rho(y) \} \, dy + o_p(1).
\end{aligned}
$$

On the other hand, using $\mathrm{E}\{ b^{**}(\mathbf{X}, a^{**}, \boldsymbol{\zeta}) \mid y \} = y$, T_2 can be written as

$$
\begin{aligned}
T_2 &= \mathrm{E} \left[\left\{ \frac{R}{\pi} \rho^*(Y) - \frac{1-R}{1-\pi} \right\} \frac{\partial b^{**}(\mathbf{X}, a^{**}, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} \right] \sqrt{n_1} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(1) \\
&= \frac{\partial}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} \mathrm{E} \left[\left\{ \frac{R}{\pi} \rho^*(Y) - \frac{1-R}{1-\pi} \right\} Y \right] \sqrt{n_1} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(1) \\
&= o_p(1).
\end{aligned}
$$

Therefore, (D.10) leads to

$$
\begin{aligned}
\sqrt{n_1}(\widehat{\theta} - \theta) &= \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \phi_{\mathrm{eff}}^{**}(\mathbf{x}_i, r_i, r_i y_i) + n_1^{-1/2} \sum_{i=1}^{n} r_i h(\mathbf{x}_i, y_i) + o_p(1) \\
&= n^{-1/2} \sum_{i=1}^{n} \left\{ \sqrt{\pi} \phi_{\mathrm{eff}}^{**}(\mathbf{x}_i, r_i, r_i y_i) + \frac{r_i}{\sqrt{\pi}} h(\mathbf{x}_i, y_i) \right\} + o_p(1),
\end{aligned}
$$

where

$$\begin{aligned} h(\mathbf{x}_i, y_i) &\equiv \int K_h(y - y_i)\{b^{**}(\mathbf{x}_i, a^{**}, \boldsymbol{\zeta}) - y\}\{\rho^*(y) - \rho(y)\}dy \\ &= \int K(t)\{b^{**}(\mathbf{x}_i, a^{**}, \boldsymbol{\zeta}) - y_i - ht\}\{\rho^*(y_i + ht) - \rho(y_i + ht)\}dt \\ &= \{b^{**}(\mathbf{x}_i, a^{**}, \boldsymbol{\zeta}) - y_i\}\{\rho^*(y_i) - \rho(y_i)\} \\ &\quad +[\{b^{**}(\mathbf{x}_i, a^{**}, \boldsymbol{\zeta}) - y_i\}\{\rho^{*\prime\prime}(y_i) - \rho''(y_i)\} - 2\{\rho^{*\prime}(y_i) - \rho'(y_i)\}] \\ &\quad \times \frac{h^2}{2}\int t^2 K(t)dt + O(h^4) \end{aligned}$$

by Conditions (A2) and (A5). Hence, by Condition (A6), we have

$$\begin{aligned} &\sqrt{n_1}(\widehat{\theta} - \theta) \\ = \; & n^{-1/2}\sum_{i=1}^{n}\left[\sqrt{\pi}\phi_{\text{eff}}^{**}(\mathbf{x}_i, r_i, r_i y_i) + \frac{r_i}{\sqrt{\pi}}\{b^{**}(\mathbf{x}_i, a^{**}, \boldsymbol{\zeta}) - y_i\}\{\rho^*(y_i) - \rho(y_i)\}\right] + o_p(1). \end{aligned}$$

□

D.8 Proof of Theorem 5.4.1

First let

$$b^*(\mathbf{x}, a, \mathrm{E}_p) \equiv \frac{\mathrm{E}_p\{a(Y)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}},$$

and for any function $g(\cdot, \mathrm{E}_p)$, its kth Gateaux derivative with respect to E_p at μ_1 in the direction μ_2 as

$$\frac{\partial^k g(\cdot, \mu_1)}{\partial {\mathrm{E}_p}^k}(\mu_2) \equiv \left.\frac{\partial^k g(\cdot, \mu_1 + h\mu_2)}{\partial h^k}\right|_{h=0}.$$

We have

$$\begin{aligned} &\frac{\partial b^*(\mathbf{x}, a, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\ = \; & \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y)\rho^*(Y) \mid \mathbf{x}\}\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}{[\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}]^2} \\ & -\frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}\mathrm{E}_p\{a(Y)\rho^*(Y) \mid \mathbf{x}\}}{[\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}]^2} \end{aligned}$$

$$
\begin{aligned}
&= b^*(\mathbf{x}, a, \mathrm{E}_p)\left[\frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{a(Y)\rho^*(Y) \mid \mathbf{x}\}} - \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}\right] \\
&= b^*(\mathbf{x}, a, \mathrm{E}_p)o_p(n_1^{-1/4}), \\
&\quad \frac{\partial^2 b^*(\mathbf{x}, a, \mu)}{\partial {\mathrm{E}_p}^2}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
&= \frac{-2(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}{[\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}]^3} \\
&\quad \times \Big[(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{a(Y)\rho^*(Y) \mid \mathbf{x}\}\mu\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\} \\
&\quad -(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}\mu\{a(Y)\rho^*(Y) \mid \mathbf{x}\}\Big] \\
&= o_p(n_1^{-1/2})
\end{aligned}
$$

for any bounded $a(y)$ and $\mu(\cdot \mid \mathbf{x})$ since $|(\widehat{\mathrm{E}}_p - \mathrm{E}_p)(\cdot \mid \mathbf{x})| = o_p(n_1^{-1/4})$ by the assumption, and these hold uniformly with respect to $\mathbf{x}$ by Condition (A4). Then by the Taylor expansion and mean value theorem, for any bounded $a(y)$ and some $\alpha \in (0,1)$,

$$
\begin{aligned}
& b^*(\mathbf{x}, a, \widehat{\mathrm{E}}_p) \\
&= b^*(\mathbf{x}, a, \mathrm{E}_p) + \frac{\partial b^*(\mathbf{x}, a, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + \frac{1}{2}\frac{\partial^2 b^*\{\mathbf{x}, a, \mathrm{E}_p + \alpha(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\}}{\partial {\mathrm{E}_p}^2}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
&= b^*(\mathbf{x}, a, \mathrm{E}_p) + \frac{\partial b^*(\mathbf{x}, a, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n_1^{-1/2}). \qquad \text{(D.12)}
\end{aligned}
$$

Noting that $a^*(y)$ is bounded under Condition (A7), we further get

$$
\begin{aligned}
\widehat{\mathcal{L}}^*(a^*)(y) &= n_1^{-1}\sum_{i=1}^n r_i K_h(y - y_i) b^*(\mathbf{x}_i, a^*, \widehat{\mathrm{E}}_p) \\
&= n_1^{-1}\sum_{i=1}^n r_i K_h(y - y_i)\left\{b^*(\mathbf{x}_i, a^*, \mathrm{E}_p) + \frac{\partial b^*(\mathbf{x}_i, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n_1^{-1/2})\right\} \\
&= n_1^{-1}\sum_{i=1}^n \mathcal{L}^*_{i,h}(a^*)(y) + o_p(n_1^{-1/2})
\end{aligned}
$$

uniformly in y by Condition (A4), where we define

$$
\mathcal{L}^*_{i,h}(a)(y) \equiv r_i K_h(y - y_i)\frac{\mathrm{E}_p\{a(Y)\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{x}_i\}}.
$$

The last equality above is because $E\{b^*(\mathbf{X}, a^*, \mathrm{E}_p) \mid y\} = y$ from the definition of a^*, hence

$$\begin{aligned}
& n_1^{-1}\sum_{i=1}^{n} r_i K_h(y-y_i)\frac{\partial b^*(\mathbf{x}_i, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= \; & n_1^{-1}\sum_{i=1}^{n} r_i K_h(y-y_i)\frac{\partial b^*(\mathbf{x}_i, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) - \frac{\partial[p_Y(y)E\{b^*(\mathbf{X}, a^*, \mathrm{E}_p) \mid y\}]}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= \; & n_1^{-1/4}\left[n_1^{-1}\sum_{i=1}^{n} r_i K_h(y-y_i) n_1^{1/4}\frac{\partial b^*(\mathbf{x}_i, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \right. \\
& \left. -p_Y(y)E\left\{ n_1^{1/4}\frac{\partial b^*(\mathbf{X}, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \mid y \right\} \right] \\
= \; & n_1^{-1/4} O_p\left\{(n_1 h)^{-1/2}\log n_1 + h^2\right\} \\
= \; & o_p(n_1^{-1/2})
\end{aligned}$$

under Conditions (A4)-(A7). In addition, for any bounded function $a(y)$,

$$\|(\widehat{\mathcal{L}}^* - \mathcal{L}^*)(a)\|_\infty = O_p\left\{(n_1 h)^{-1/2}\log n_1 + h^2\right\} + o_p(n_1^{-1/4}) = o_p(n_1^{-1/4})$$

under Conditions (A4)-(A7), and the assumption $|\widehat{\mathrm{E}}_p\{a(Y) \mid \mathbf{x}\} - \mathrm{E}_p\{a(Y) \mid \mathbf{x}\}| = o_p(n_1^{-1/4})$. Similarly, $\|\widehat{v} - v\|_\infty = o_p(n_1^{-1/4})$. Then using that $\mathcal{L}^{*-1}$ is a bounded linear operator by Lemma 5.3.1, $\widehat{a}^*(y)$ can be expressed as

$$\begin{aligned}
\widehat{a}^*(y) &= \{\mathcal{L}^* + (\widehat{\mathcal{L}}^* - \mathcal{L}^*)\}^{-1}(\widehat{v})(y) \\
&= \{\mathcal{L}^{*-1} - \mathcal{L}^{*-1}(\widehat{\mathcal{L}}^* - \mathcal{L}^*)\mathcal{L}^{*-1}\}\{v + (\widehat{v} - v)\}(y) + o_p(n_1^{-1/2}) \\
&= a^*(y) + \mathcal{L}^{*-1}(\widehat{v} - v)(y) - \{\mathcal{L}^{*-1}(\widehat{\mathcal{L}}^* - \mathcal{L}^*)\}(a^*)(y) + o_p(n_1^{-1/2}) \\
&= a^*(y) + \mathcal{L}^{*-1}\{\widehat{v} - \widehat{\mathcal{L}}^*(a^*)\}(y) + o_p(n_1^{-1/2}) \\
&= a^*(y) + n_1^{-1}\sum_{i=1}^{n} g_{i,h}(y) + o_p(n_1^{-1/2})
\end{aligned}$$

uniformly in y where $g_{i,h}(y) \equiv \mathcal{L}^{*-1}\{v_{i,h} - \mathcal{L}^*_{i,h}(a^*)\}(y)$. Hence, noting that $b^*(\mathbf{x}, a, \mathrm{E}_p)$ is linear with respect to a and using (D.12), we get

$$\begin{aligned}
b^*(\mathbf{x}, \widehat{a}^*, \widehat{\mathrm{E}}_p) &= b^*(\mathbf{x}, a^*, \mathrm{E}_p) + b^*(\mathbf{x}, \widehat{a}^* - a^*, \mathrm{E}_p) + \frac{\partial b^*(\mathbf{x}, \widehat{a}^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n_1^{-1/2}) \\
&= b^*(\mathbf{x}, a^*, \mathrm{E}_p) + n_1^{-1}\sum_{i=1}^{n} b^*(\mathbf{x}, g_{i,h}, \mathrm{E}_p) + \frac{\partial b^*(\mathbf{x}, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n_1^{-1/2})
\end{aligned}$$

uniformly in $\mathbf{x}$ by Condition (A4). The last equality holds because for any bounded $a(y)$,

$$|b^*(\mathbf{x}, a, \mathrm{E}_p)| \leqslant \|a\|_\infty \frac{\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}} = O(\|a\|_\infty),$$

and

$$\begin{aligned}
& \frac{\partial b^*(\mathbf{x}, \widehat{a}^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) - \frac{\partial b^*(\mathbf{x}, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= & \frac{\partial b^*(\mathbf{x}, \widehat{a}^* - a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= & b^*(\mathbf{x}, \widehat{a}^* - a^*, \mathrm{E}_p) o_p(n_1^{-1/4}) \\
= & \left[O_p(h^2) + O_p\{(n_1 h)^{-1/2} \log(n_1)\}\right] o_p(n_1^{-1/4}) \\
= & o_p(n_1^{-1/2}),
\end{aligned}$$

where the third equality is due to $\|E(G_{i,h})\|_\infty = O(h^2)$ by (D.8), and the last equality holds because $h^2 = o_p(n_1^{-1/2})$ and $n_1\{\log(n_1)\}^{-4}h^2 \to \infty$ by Condition (A6). Now, by the definition of $\widetilde{\theta}$,

$$\begin{aligned}
& \sqrt{n_1}(\widetilde{\theta} - \theta) \\
= & \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \left[\frac{r_i}{\pi} \rho^*(y_i)\{y_i - b^*(\mathbf{x}_i, \widehat{a}^*, \widehat{\mathrm{E}}_p)\} + \frac{1 - r_i}{1 - \pi}\{b^*(\mathbf{x}_i, \widehat{a}^*, \widehat{\mathrm{E}}_p) - \theta\} \right] \\
= & \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \phi^*_{\mathrm{eff}}(\mathbf{x}_i, r_i, r_i y_i) \\
& + \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \left\{ \frac{r_i}{\pi} \rho^*(y_i) - \frac{1 - r_i}{1 - \pi} \right\} \left\{ b^*(\mathbf{x}_i, a^*, \mathrm{E}_p) - b^*(\mathbf{x}_i, \widehat{a}^*, \widehat{\mathrm{E}}_p) \right\} \\
= & \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \phi^*_{\mathrm{eff}}(\mathbf{x}_i, r_i, r_i y_i) - T_1 - T_2 + o_p(1), \qquad \text{(D.13)}
\end{aligned}$$

where

$$\begin{aligned}
T_1 & \equiv n_1^{-1/2} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \frac{r_i}{\pi} \rho^*(y_i) - \frac{1 - r_i}{1 - \pi} \right\} b^*(\mathbf{x}_i, g_{j,h}, \mathrm{E}_p), \\
T_2 & \equiv \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \left\{ \frac{r_i}{\pi} \rho^*(y_i) - \frac{1 - r_i}{1 - \pi} \right\} \frac{\partial b^*(\mathbf{x}_i, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p).
\end{aligned}$$

Note that $E\{b^*(\mathbf{x}_i, G_{j,h}, \mathrm{E}_p) \mid \mathbf{x}_i\} = b^*\{\mathbf{x}_i, E(G_{j,h}), \mathrm{E}_p\} = O(h^2)$ by (D.8). Then using the property of the U-statistic and Condition (A6), we can express T_1 as

$$
\begin{aligned}
& T_1 \\
= \; & n_1^{-1/2} \sum_{i=1}^{n} \left\{ \frac{r_i}{\pi} \rho^*(y_i) - \frac{1-r_i}{1-\pi} \right\} E\{b^*(\mathbf{x}_i, G_{j,h}, \mathrm{E}_p) \mid \mathbf{x}_i, r_i, r_i y_i\} \\
& + n_1^{-1/2} \sum_{j=1}^{n} E\left[\left\{ \frac{R_i}{\pi} \rho^*(Y_i) - \frac{1-R_i}{1-\pi} \right\} b^*(\mathbf{X}_i, g_{j,h}, \mathrm{E}_p) \mid \mathbf{x}_j, r_j, r_j y_j \right] \\
& - n^{1/2} n_1^{-1/2} E\left[\left\{ \frac{R_i}{\pi} \rho^*(Y_i) - \frac{1-R_i}{1-\pi} \right\} b^*(\mathbf{X}_i, G_{j,h}, \mathrm{E}_p) \right] + O_p(n_1^{-1/2}) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} E\left[\left\{ \frac{R_i}{\pi} \rho^*(Y_i) - \frac{1-R_i}{1-\pi} \right\} b^*(\mathbf{X}_i, g_{j,h}, \mathrm{E}_p) \mid \mathbf{x}_j, r_j, r_j y_j \right] + O_p(n n_1^{-1/2} h^2 + n_1^{-1/2}) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} \int \left[\rho^*(y) E\{b^*(\mathbf{X}, g_{j,h}, \mathrm{E}_p) \mid y\} p_Y(y) - E\{b^*(\mathbf{X}, g_{j,h}, \mathrm{E}_p) \mid y\} q_Y(y) \right] dy + o_p(1) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} \int \mathcal{L}^*(g_{j,h})(y) \left\{ \rho^*(y) - \rho(y) \right\} dy + o_p(1) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} \int \{v_{j,h}(y) - \mathcal{L}^*_{j,h}(a^*)(y)\} \left\{ \rho^*(y) - \rho(y) \right\} dy + o_p(1) \\
= \; & n_1^{-1/2} \sum_{j=1}^{n} r_j \int K_h(y - y_j) \{ y - b^*(\mathbf{x}_j, a^*, \mathrm{E}_p) \} \left\{ \rho^*(y) - \rho(y) \right\} dy + o_p(1),
\end{aligned}
$$

where

$$
\begin{aligned}
& \int K_h(y - y_j) \{ y - b^*(\mathbf{x}_j, a^*, \mathrm{E}_p) \} \left\{ \rho^*(y) - \rho(y) \right\} dy \\
= \; & \int K(t) \{ y_j + ht - b^*(\mathbf{x}_j, a^*, \mathrm{E}_p) \} \{ \rho^*(y_j + ht) - \rho(y_j + ht) \} dt \\
= \; & \{ y_j - b^*(\mathbf{x}_j, a^*, \mathrm{E}_p) \} \{ \rho^*(y_j) - \rho(y_j) \} \\
& + \left[\{ y_j - b^*(\mathbf{x}_j, a^*, \mathrm{E}_p) \} \{ \rho^{*\prime\prime}(y_j) - \rho''(y_j) \} + 2 \{ \rho^{*\prime}(y_j) - \rho'(y_j) \} \right] \frac{h^2}{2} \int t^2 K(t) dt + O(h^4) \\
= \; & \{ y_j - b^*(\mathbf{x}_j, a^*, \mathrm{E}_p) \} \{ \rho^*(y_j) - \rho(y_j) \} + o(n_1^{-1/2})
\end{aligned}
$$

under Conditions (A2), (A5), and (A6). On the other hand, using $E\{b^*(\mathbf{X}, a^*, \mathrm{E}_p) \mid y\} = y$, T_2 can be written as

$$
T_2 \;=\; \sqrt{n_1} n^{-1} \sum_{i=1}^{n} \left\{ \frac{r_i}{\pi} \rho^*(y_i) - \frac{1-r_i}{1-\pi} \right\} \frac{\partial b^*(\mathbf{x}_i, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p} (\widehat{\mathrm{E}}_p - \mathrm{E}_p)
$$

$$
\begin{aligned}
& -\sqrt{n_1}\frac{\partial}{\partial \mathrm{E}_p} E\left[\left\{\frac{R}{\pi}\rho^*(Y)-\frac{1-R}{1-\pi}\right\} b^*(\mathbf{X}, a^*, \mathrm{E}_p)\right](\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= \; & n_1^{1/4}\left(n^{-1}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i)-\frac{1-r_i}{1-\pi}\right\} n_1^{1/4}\frac{\partial b^*(\mathbf{x}_i, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\right. \\
& \left. -E\left[\left\{\frac{R}{\pi}\rho^*(Y)-\frac{1-R}{1-\pi}\right\} n_1^{1/4}\frac{\partial b^*(\mathbf{X}, a^*, \mathrm{E}_p)}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\right]\right) \\
= \; & n_1^{1/4} O_p(n^{-1/2}) \\
= \; & o_p(1).
\end{aligned}
$$

Therefore, (D.13) leads to

$$
\begin{aligned}
& \sqrt{n_1}(\widetilde{\theta}-\theta) \\
= \; & n^{-1/2}\sum_{i=1}^{n}\left[\sqrt{\pi}\phi^*_{\text{eff}}(\mathbf{x}_i, r_i, r_i y_i) + \frac{r_i}{\sqrt{\pi}}\{b^*(\mathbf{x}_i, a^*, \mathrm{E}_p) - y_i\}\{\rho^*(y_i) - \rho(y_i)\}\right] + o_p(1).
\end{aligned}
$$

□

D.9 Proposed doubly flexible estimation for $\boldsymbol{\theta}$ such that $\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$

D.9.1 General approach

Let $\dim(\mathbf{U}) = \dim(\boldsymbol{\theta})$, $\mathrm{E}_q\{\partial \mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}$ be invertible, and $\mathbf{A} \equiv [\mathrm{E}_q\{\partial \mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}]^{-1}$. We first establish $\mathcal{F}$, the family of all influence functions for estimating $\boldsymbol{\theta}$. In Section D.11.1, we show that

$$
\begin{aligned}
\mathcal{F} \; \equiv \; & \left[\frac{r}{\pi}[\rho(y)\mathbf{A}\{\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta}) - \mathbf{b}(\mathbf{x})\} + \mathbf{c}] + \frac{1-r}{1-\pi}\{\mathbf{A}\mathbf{b}(\mathbf{x}) - \mathbf{c}\}\right. \\
& \left. : \mathrm{E}\{\mathbf{b}(\mathbf{X}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}, \forall \mathbf{c}\right].
\end{aligned}
$$

We now derive the efficient influence function $\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}, r, ry)$ which corresponds to the semiparametric efficiency bound and also provides guidance on constructing flexible estimators for $\boldsymbol{\theta}$. The derivation is provided in Section D.11.2.

Proposition D.9.1. *The efficient influence function $\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}, r, ry)$ for $\boldsymbol{\theta}$ is*

$$
\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}, r, ry) \; = \; \frac{r}{\pi}\rho(y)\mathbf{A}\left[\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta}) - \frac{\mathrm{E}_p\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho^2(Y) + \mathbf{a}(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) + \pi/(1-\pi)\rho(Y) \mid \mathbf{x}\}}\right]
$$

$$+\frac{1-r}{1-\pi}\frac{\mathbf{A}\mathrm{E}_p\{\mathbf{U}(\mathbf{x},Y,\boldsymbol{\theta})\rho^2(Y)+\mathbf{a}(Y)\rho(Y)\mid\mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y)+\pi/(1-\pi)\rho(Y)\mid\mathbf{x}\}},$$

where $\mathbf{a}(y)$ *satisfies*

$$E\left[\frac{\mathrm{E}_p\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\rho^2(Y)+\mathbf{a}(Y)\rho(Y)\mid\mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y)+\pi/(1-\pi)\rho(Y)\mid\mathbf{X}\}}\mid y\right]=E\{\mathbf{U}(\mathbf{X},y,\boldsymbol{\theta})\mid y\}.$$

The efficient estimator can be obtained by solving the estimating equation

$$\sum_{i=1}^{n}\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}_i,r_i,r_iy_i)=\mathbf{0},$$

where the efficient influence function $\boldsymbol{\phi}_{\text{eff}}(\mathbf{x},r,ry)$ is given in Proposition D.9.1. However, constructing $\boldsymbol{\phi}_{\text{eff}}(\mathbf{x},r,ry)$ requires two possibly unknown functions $p_{Y|\mathbf{X}}(y,\mathbf{x})$ and $\rho(y)$. We will consider replacing the two functions by working models $p^{\star}_{Y|\mathbf{X}}(y,\mathbf{x})$ and $\rho^*(y)$.

D.9.2 Proposed doubly flexible estimator

Under the adopted working models $p^{\star}_{Y|\mathbf{X}}(y,\mathbf{x})$ and $\rho^*(y)$, we obtain an estimator $\widehat{\boldsymbol{\theta}}_t$ by solving the estimating equation $\sum_{i=1}^{n}\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{x}_i,r_i,r_iy_i)=\mathbf{0}$. Given that $\mathbf{A}$ is invertible, this is equivalent to solving

$$\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\{\mathbf{U}(\mathbf{x}_i,y_i,\boldsymbol{\theta})-\mathbf{b}^{**}(\mathbf{x}_i,\boldsymbol{\theta})\}+\frac{1-r_i}{1-\pi}\mathbf{b}^{**}(\mathbf{x}_i,\boldsymbol{\theta})\right]=\mathbf{0},\tag{D.14}$$

where

$$\mathbf{b}^{**}(\mathbf{x},\boldsymbol{\theta})\equiv\frac{\mathrm{E}^{\star}_p\{\mathbf{U}(\mathbf{x},Y,\boldsymbol{\theta})\rho^{*2}(Y)+\mathbf{a}^{**}(Y,\boldsymbol{\theta})\rho^*(Y)\mid\mathbf{x}\}}{\mathrm{E}^{\star}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid\mathbf{x}\}},$$

and $\mathbf{a}^{**}(y,\boldsymbol{\theta})$ is a solution for

$$\mathrm{E}\left[\frac{\mathrm{E}^{\star}_p\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\rho^{*2}(Y)+\mathbf{a}^{**}(Y,\boldsymbol{\theta})\rho^*(Y)\mid\mathbf{X}\}}{\mathrm{E}^{\star}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid\mathbf{X}\}}\mid y\right]=\mathrm{E}\{\mathbf{U}(\mathbf{X},y,\boldsymbol{\theta})\mid y\}.\tag{D.15}$$

Our analysis shows that $\widehat{\boldsymbol{\theta}}_t$ is still consistent for $\boldsymbol{\theta}$ under suitable conditions. We state this result as a proposition below, and provide its proof in Section D.11.3.

Proposition D.9.2. *Assume* $E\{\partial\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{X},R,RY,\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\}$ *is invertible,* $\boldsymbol{\theta}\in\boldsymbol{\theta}$*, where* $\boldsymbol{\theta}$ *is compact. Assume* $E\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{X},R,RY,\boldsymbol{\theta})\|_2\}<\infty$*. Then* $\widehat{\boldsymbol{\theta}}_t$ *is consistent for* $\boldsymbol{\theta}$*.*

Practically, we may choose to adopt a parametric working model $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta})$ instead of a fixed function $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x})$. In such construction, a natural way will be to find an estimator for $\boldsymbol{\zeta}$, say $\widehat{\boldsymbol{\zeta}}$, using samples drawn from the population $\mathcal{P}$, and use the estimated model $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x}, \widehat{\boldsymbol{\zeta}})$ to form the estimating equation (D.14). We will analyze this strategy in the sequel.

In addition, constructing (D.14) requires $g(\mathbf{x}, y)$ since $\mathbf{a}^{*\star}(y, \boldsymbol{\theta})$ is obtained by solving (D.15). To overcome this issue, we incorporate nonparametric estimation. Note that directly estimating $g(\mathbf{x}, y)$ is subject to the curse of dimensionality since $\mathbf{x}$ is possibly high-dimensional. Hence, instead, we estimate the both sides of (D.15) using nonparametric regression. In summary, we approximate (D.15) using a kernel-based estimator of $\mathrm{E}(\cdot \mid y)$ as

$$\begin{aligned} & \int \mathbf{a}^{*\star}(t, \boldsymbol{\theta}) \rho^{*}(t) \sum_{i=1}^{n} \frac{p^{\star}_{Y|\mathbf{X}}(t, \mathbf{x}_i) r_i K_h(y - y_i)}{\mathrm{E}^{\star}_{p}\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^{*}(Y) \mid \mathbf{x}_i\}} dt \\ = & \sum_{i=1}^{n} \left[\mathbf{U}(\mathbf{x}_i, y, \boldsymbol{\theta}) - \frac{\mathrm{E}^{\star}_{p}\{\mathbf{U}(\mathbf{x}_i, Y, \boldsymbol{\theta}) \rho^{*2}(Y) \mid \mathbf{x}_i\}}{\mathrm{E}^{\star}_{p}\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^{*}(Y) \mid \mathbf{x}_i\}} \right] r_i K_h(y - y_i), \quad \text{(D.16)} \end{aligned}$$

and solve this integral equation with respect to $\mathbf{a}^{*\star}(\cdot, \boldsymbol{\theta})$ to obtain $\widehat{\mathbf{a}}^{*\star}(\cdot, \boldsymbol{\theta})$.

We summarize the estimation procedure in Algorithm 5.

Algorithm 5 Proposed Estimator $\widehat{\boldsymbol{\theta}}$: Doubly Flexible in $\rho^{*}(y)$ and $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x})$

Input: data from population $\mathcal{P}$: $(y_i, \mathbf{x}_i, r_i = 1)$, $i = 1, \ldots, n_1$, data from population $\mathcal{Q}$: $(\mathbf{x}_j, r_j = 0)$, $j = n_1 + 1, \ldots, n$, and value $\pi = n_1/n$.

do

(a) adopt a working model for $\rho(y)$, denoted as $\rho^{*}(y)$;

(b) adopt a working model for $p_{Y|\mathbf{X}}(y, \mathbf{x})$, denoted as $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x})$ or $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x}, \widehat{\boldsymbol{\zeta}})$;

(c) compute $w_i = [\mathrm{E}^{\star}_{p}\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^{*}(Y) \mid \mathbf{x}_i\}]^{-1}$ for $i = 1, \ldots, n$;

(d) obtain $\widehat{\mathbf{a}}^{*\star}(\cdot, \boldsymbol{\theta})$ by solving the integral equation (D.16);

(e) compute $\widehat{\mathbf{b}}^{*\star}(\mathbf{x}_i, \boldsymbol{\theta}) = w_i \mathrm{E}^{\star}_{p}\{\mathbf{U}(\mathbf{x}_i, Y, \boldsymbol{\theta}) \rho^{*2}(Y) + \widehat{\mathbf{a}}^{*\star}(Y, \boldsymbol{\theta}) \rho^{*}(Y) \mid \mathbf{x}_i\}$ for $i = 1, \ldots, n$;

(f) obtain $\widehat{\boldsymbol{\theta}}$ by solving the estimating equation

$$\sum_{i=1}^{n} \left[\frac{r_i}{\pi} \rho^{*}(y_i) \{\mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) - \widehat{\mathbf{b}}^{*\star}(\mathbf{x}_i, \boldsymbol{\theta})\} + \frac{1 - r_i}{1 - \pi} \widehat{\mathbf{b}}^{*\star}(\mathbf{x}_i, \boldsymbol{\theta}) \right] = \mathbf{0}.$$

Output: $\widehat{\boldsymbol{\theta}}$.

We now study the theoretical properties of $\widehat{\boldsymbol{\theta}}$. The main technical challenge arises from the gap between the solutions for (D.15) and (D.16). For quantifying the gap, we introduce some notations. Let

$$\begin{aligned}
\mathrm{E}_p^\star\{\mathbf{a}(\mathbf{x}, Y, \boldsymbol{\theta}) \mid \mathbf{x}, \widehat{\boldsymbol{\zeta}}\} &\equiv \int \mathbf{a}(\mathbf{x}, y, \boldsymbol{\theta}) p^\star_{Y|\mathbf{X}}(y, \mathbf{x}, \widehat{\boldsymbol{\zeta}}) dy, \\
u^{**}(t, y) &\equiv p_Y(y) \int \frac{\rho^*(t) p^\star_{Y|\mathbf{X}}(t, \mathbf{x}, \boldsymbol{\zeta})}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}\}} g(\mathbf{x}, y) d\mathbf{x}, \\
\mathcal{L}^{**}(\mathbf{a})(y, \boldsymbol{\theta}) &\equiv p_Y(y) \mathrm{E}\left[\frac{\mathrm{E}_p^\star\{\mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{X}\}} \mid y \right] \\
&= \int \mathbf{a}(t, \boldsymbol{\theta}) u^{**}(t, y) dt,
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{L}^{**}_{i,h}(\mathbf{a})(y, \boldsymbol{\theta}) &\equiv r_i K_h(y - y_i) \frac{\mathrm{E}_p^\star\{\mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}_i\}}, \\
\widehat{\mathcal{L}}^{**}(\mathbf{a})(y, \boldsymbol{\theta}) &\equiv n_1^{-1} \sum_{i=1}^n r_i K_h(y - y_i) \frac{\mathrm{E}_p^\star\{\mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i, \widehat{\boldsymbol{\zeta}}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) \mid \mathbf{x}_i, \widehat{\boldsymbol{\zeta}}\} + \pi/(1-\pi)\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}_i, \widehat{\boldsymbol{\zeta}}\}}, \\
\mathbf{v}^{**}(y, \boldsymbol{\theta}) &\equiv p_Y(y) \mathrm{E}\left[\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) - \frac{\mathrm{E}_p^\star\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\rho^{*2}(Y) \mid \mathbf{X}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{X}\}} \mid y \right], \\
\mathbf{v}^{**}_{i,h}(y, \boldsymbol{\theta}) &\equiv r_i K_h(y - y_i) \left[\mathbf{U}(\mathbf{x}_i, y, \boldsymbol{\theta}) - \frac{\mathrm{E}_p^\star\{\mathbf{U}(\mathbf{x}_i, Y, \boldsymbol{\theta})\rho^{*2}(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \right], \\
\widehat{\mathbf{v}}^{**}(y, \boldsymbol{\theta}) &\equiv n_1^{-1} \sum_{i=1}^n r_i K_h(y - y_i) \left[\mathbf{U}(\mathbf{x}_i, y, \boldsymbol{\theta}) - \frac{\mathrm{E}_p^\star\{\mathbf{U}(\mathbf{x}_i, Y, \boldsymbol{\theta})\rho^{*2}(Y) \mid \mathbf{x}_i, \widehat{\boldsymbol{\zeta}}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i, \widehat{\boldsymbol{\zeta}}\}} \right].
\end{aligned}$$

By the definitions of the solutions for (D.15) and (D.16), $\mathbf{a}^{**}(y, \boldsymbol{\theta})$ and $\widehat{\mathbf{a}}^{**}(y, \boldsymbol{\theta})$ satisfy $\mathcal{L}^{**}(\mathbf{a}^{**})(y, \boldsymbol{\theta}) = \mathbf{v}^{**}(y, \boldsymbol{\theta})$ and $\widehat{\mathcal{L}}^{**}(\widehat{\mathbf{a}}^{**})(y, \boldsymbol{\theta}) = \widehat{\mathbf{v}}^{**}(y, \boldsymbol{\theta})$.

We list a set of regularity conditions.

(B1) $\mathbf{A}^{**} \equiv [\mathrm{E}_q^{**}\{\partial \mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}]^{-1}$ and $\mathrm{E}\{\partial \boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}$ are invertible, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $\boldsymbol{\Theta}$ is compact, and $\mathrm{E}\{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\|_2\} < \infty$. $\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta})$ is twice differentiable with respect to y and its derivative is bounded.

(B2) $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$ is complete.

(B3) $\rho^*(y) > \delta$ for all y on the support of $p_Y(y)$, where δ is a constant such that $\delta > 0$. $\rho^*(y)$ is twice differentiable and its derivative is bounded.

(B4) The function $u^{**}(t,y)$ is bounded and has bounded derivatives with respect to t and y on its support. $\mathbf{a}^{**}(y,\boldsymbol{\theta})$ in (D.15) is bounded.

(B5) The support sets of $g(\mathbf{x},y), p_Y(y), \rho^*(y)$ are compact.

(B6) The kernel function $K(\cdot) \geqslant 0$ is symmetric, bounded, and twice differentiable with bounded first derivative. It has support on $(-1,1)$ and satisfies $\int_{-1}^{1} K(t)dt = 1$.

(B7) The bandwidth h satisfies $n_1(\log n_1)^{-4}h^2 \to \infty$ and $n^2 n_1^{-1} h^4 \to 0$.

Lemma D.9.1. *For* $\mathbf{a}(y,\boldsymbol{\theta}) = \{a_1(y,\boldsymbol{\theta}),\ldots,a_d(y,\boldsymbol{\theta})\}^{\mathrm{T}}$, *let* $\|\mathbf{a}\|_\infty \equiv \max_{k=1\ldots d}\|a_k\|_\infty$. *Under the regularity conditions (B1)-(B5), the linear operator* $\mathcal{L}^{**}: L^\infty(R^d) \to L^\infty(R^d)$ *is invertible. In addition, there exist constants* c_1, c_2 *such that* $0 < c_1, c_2 < \infty$ *and for all* $\mathbf{a}(y) \in L^\infty(R^d)$,

(i) $c_1\|\mathbf{a}\|_\infty \leqslant \|\mathcal{L}^{**}(\mathbf{a})\|_\infty \leqslant c_2\|\mathbf{a}\|_\infty$,

(ii) $\|\mathcal{L}^{**-1}(\mathbf{a})\|_\infty \leqslant c_1^{-1}\|\mathbf{a}\|_\infty$.

The proof of Lemma D.9.1 is in Section D.11.4. Below, we present the asymptotic normality of $\widehat{\boldsymbol{\theta}}$, and provide its proof in Section D.11.5.

Theorem D.9.1. *Assume* $\widehat{\boldsymbol{\zeta}}$ *satisfies* $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$ *and* $\mathrm{E}_p^\star\{\|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y,\mathbf{x},\boldsymbol{\zeta})\|_2 \mid \mathbf{x}\}$ *is bounded, where* $\mathbf{S}_{\boldsymbol{\zeta}}^\star(y,\mathbf{x},\boldsymbol{\zeta}) \equiv \partial \log p_{Y|\mathbf{X}}^\star(y,\mathbf{x},\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}$. *For any choice of* $p_{Y|\mathbf{X}}^\star(y,\mathbf{x},\boldsymbol{\zeta})$ *and* $\rho^*(y)$, *under Conditions (B1)-(B7),*

$$\sqrt{n_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N\left\{\mathbf{0}, \mathbf{A}\mathbf{A}^{**-1}\boldsymbol{\Sigma}(\mathbf{A}\mathbf{A}^{**-1})^{\mathrm{T}}\right\}$$

in distribution as $n_1 \to \infty$, *where*

$$\boldsymbol{\Sigma} \equiv var\left[\sqrt{\pi}\boldsymbol{\phi}_{\mathrm{eff}}^{**}(\mathbf{X},R,RY,\boldsymbol{\theta}) + \frac{R}{\sqrt{\pi}}\mathbf{A}^{**}\left\{\mathbf{b}^{**}(\mathbf{X},\boldsymbol{\theta}) - \mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\right\}\left\{\rho^*(Y) - \rho(\mathbf{X})\right\}\right] \quad (D.17)$$

Remark D.9.1. *In Theorem D.9.1,* $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$ *is the only assumption imposed on* $\widehat{\boldsymbol{\zeta}}$. *That is, the result in Theorem D.9.1 holds as long as* $\widehat{\boldsymbol{\zeta}}$ *is* $\sqrt{n_1}$*-consistent for* $\boldsymbol{\zeta}$, *regardless of the asymptotic variance of* $\sqrt{n_1}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})$. *This is easily achievable by constructing a standard MLE or moment based estimator for* $\boldsymbol{\zeta}$ *in* $p_{Y|\mathbf{X}}^\star(y,\mathbf{x},\boldsymbol{\zeta})$, *based on the* n_1 *observations from population* $\mathcal{P}$.

In addition to the above remark, we can see from Theorem D.9.1 that $\widehat{\boldsymbol{\theta}}$ is indeed the efficient estimator for $\boldsymbol{\theta}$ when the working models $p_{Y|\mathbf{X}}^\star(y,\mathbf{x},\boldsymbol{\zeta})$ and $\rho^*(y)$ are correctly

specified, and $\widehat{\boldsymbol{\zeta}}$ satisfies $\|\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$. We formally state this result as Corollary D.9.1.

Corollary D.9.1. *Assume* $\widehat{\boldsymbol{\zeta}}$ *satisfies* $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$ *and* $\mathrm{E}_p^{\star}\{\|\mathbf{S}_{\boldsymbol{\zeta}}^{\star}(Y, \mathbf{x}, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}\}$ *is bounded. If* $p_{Y|\mathbf{X}}^{\star}(y, \mathbf{x}, \boldsymbol{\zeta}) = p_{Y|\mathbf{X}}(y, \mathbf{x})$ *and* $\rho^*(y) = \rho(y)$*, under Conditions (B1)-(B7),*

$$\sqrt{n_1}(\widehat{\boldsymbol{\theta}}_{\text{eff}} - \boldsymbol{\theta}) \to N\left[0, var\left\{\sqrt{\pi}\boldsymbol{\phi}_{\text{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\right\}\right]$$

in distribution as $n_1 \to \infty$.

D.10 Alternative singly flexible estimator

We now consider replacing $\mathrm{E}_p(\cdot \mid \mathbf{x})$ in the efficient influence function in Proposition D.9.1 by an arbitrary estimator $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$ with convergence rate faster than $n_1^{-1/4}$. We first describe the estimation procedure in Algorithm 6.

Algorithm 6 Alternative Estimator $\widetilde{\boldsymbol{\theta}}$: Singly Flexible in $\rho^*(y)$

Input: data from population $\mathcal{P}$: $(y_i, \mathbf{x}_i, r_i = 1)$, $i = 1, \ldots, n_1$, data from population $\mathcal{Q}$: $(\mathbf{x}_j, r_j = 0)$, $j = n_1 + 1, \ldots, n$, and value $\pi = n_1/n$.

do

(a) adopt a working model for $\rho(y)$, denoted as $\rho^*(y)$;

(b) adopt a nonparametric or machine learning algorithm for estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$, denoted as $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$;

(c) compute $\widehat{w}_i = [\widehat{\mathrm{E}}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}]^{-1}$ for $i = 1, \ldots, n$;

(d) obtain $\widehat{\mathbf{a}}^*(\cdot, \boldsymbol{\theta})$ by solving the integral equation (D.16) with E_p replaced by $\widehat{\mathrm{E}}_p$;

(e) compute $\widehat{\mathbf{b}}^*(\mathbf{x}_i, \boldsymbol{\theta}) = \widehat{w}_i\widehat{\mathrm{E}}_p\{\mathbf{U}(\mathbf{x}_i, Y, \boldsymbol{\theta})\rho^{*2}(Y) + \widehat{\mathbf{a}}^*(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}$ for $i = 1, \ldots, n$;

(f) obtain $\widehat{\boldsymbol{\theta}}$ by solving the estimating equation

$$\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\{\mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) - \widehat{\mathbf{b}}^*(\mathbf{x}_i, \boldsymbol{\theta})\} + \frac{1-r_i}{1-\pi}\widehat{\mathbf{b}}^*(\mathbf{x}_i, \boldsymbol{\theta})\right] = \mathbf{0}.$$

Output: $\widetilde{\boldsymbol{\theta}}$.

For establishing the theoretical properties of $\widetilde{\boldsymbol{\theta}}$, we define

$$u^*(t, y) \equiv p_Y(y)\int \frac{\rho^*(t)p_{Y|\mathbf{X}}(t, \mathbf{x})}{\mathrm{E}_p\{\rho^{*2}(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{x}\}}g(\mathbf{x}, y)d\mathbf{x},$$

$$
\begin{aligned}
\mathcal{L}^*(\mathbf{a})(y,\boldsymbol{\theta}) &\equiv p_Y(y)\mathrm{E}\left[\frac{\mathrm{E}_p\{\mathbf{a}(Y,\boldsymbol{\theta})\rho^*(Y)\mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^{*2}(Y)\mid \mathbf{X}\}+\pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y)\mid \mathbf{X}\}}\mid y\right]\\
&= \int \mathbf{a}(t,\boldsymbol{\theta})u^*(t,y)dt,\\
\mathcal{L}^*_{i,h}(\mathbf{a})(y,\boldsymbol{\theta}) &\equiv r_iK_h(y-y_i)\frac{\mathrm{E}_p\{\mathbf{a}(Y,\boldsymbol{\theta})\rho^*(Y)\mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y)\mid \mathbf{x}_i\}+\pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y)\mid \mathbf{x}_i\}},\\
\widehat{\mathcal{L}}^*(\mathbf{a})(y,\boldsymbol{\theta}) &\equiv n_1^{-1}\sum_{i=1}^n r_iK_h(y-y_i)\frac{\widehat{\mathrm{E}}_p\{\mathbf{a}(Y,\boldsymbol{\theta})\rho^*(Y)\mid \mathbf{x}_i\}}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y)\mid \mathbf{x}_i\}+\pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y)\mid \mathbf{x}_i\}},\\
\mathbf{v}^*(y,\boldsymbol{\theta}) &\equiv p_Y(y)\mathrm{E}\left[\mathbf{U}(\mathbf{X},y,\boldsymbol{\theta})-\frac{\mathrm{E}_p\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\rho^{*2}(Y)\mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{X}\}}\mid y\right],\\
\mathbf{v}^*_{i,h}(y,\boldsymbol{\theta}) &\equiv r_iK_h(y-y_i)\left[\mathbf{U}(\mathbf{x}_i,y,\boldsymbol{\theta})-\frac{\mathrm{E}_p\{\mathbf{U}(\mathbf{x}_i,Y,\boldsymbol{\theta})\rho^{*2}(Y)\mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{x}_i\}}\right],\\
\widehat{\mathbf{v}}^*(y,\boldsymbol{\theta}) &\equiv n_1^{-1}\sum_{i=1}^n r_iK_h(y-y_i)\left[\mathbf{U}(\mathbf{x}_i,y,\boldsymbol{\theta})-\frac{\widehat{\mathrm{E}}_p\{\mathbf{U}(\mathbf{x}_i,Y,\boldsymbol{\theta})\rho^{*2}(Y)\mid \mathbf{x}_i\}}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{x}_i\}}\right].
\end{aligned}
$$

Then the solutions for (D.15) and (D.16), $\mathbf{a}^*(y,\boldsymbol{\theta})$ and $\widehat{\mathbf{a}}^*(y,\boldsymbol{\theta})$ satisfy $\mathcal{L}^*(\mathbf{a}^*)(y,\boldsymbol{\theta}) = \mathbf{v}^*(y,\boldsymbol{\theta})$ and $\widehat{\mathcal{L}}^*(\widehat{\mathbf{a}}^*)(y,\boldsymbol{\theta}) = \widehat{\mathbf{v}}^*(y,\boldsymbol{\theta})$.

Theorem D.10.1. *Assume $\widehat{E}_p$ satisfies $\|\widehat{E}_p\{\mathbf{a}(Y)\mid \mathbf{x}\}-\mathrm{E}_p\{\mathbf{a}(Y)\mid \mathbf{x}\}\|_\infty = o_p(n_1^{-1/4})$ for any bounded function $\mathbf{a}(y)$. For any choice of $\rho^*(y)$, under Conditions (B1)-(B7),*

$$\sqrt{n_1}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}) \to N\left\{\mathbf{0}, \mathbf{A}\mathbf{A}^{*-1}\boldsymbol{\Sigma}(\mathbf{A}\mathbf{A}^{*-1})^{\mathrm{T}}\right\}$$

in distribution as $n_1 \to \infty$, where

$$\boldsymbol{\Sigma} \equiv var\left[\sqrt{\pi}\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X},R,RY,\boldsymbol{\theta})+\frac{R}{\sqrt{\pi}}\mathbf{A}^*\{\mathbf{b}^*(\mathbf{X},\boldsymbol{\theta})-\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\}\{\rho^*(Y)-\rho(Y)\}\right].$$

The proof of Theorem D.10.1 is in Section D.11.6. Theorem D.10.1 further confirms that when the shifting model $\rho^*(y)$ is correctly posited, the estimator $\widetilde{\boldsymbol{\theta}}$ achieves the efficiency bound. We point out this result as Corollary D.10.1.

Corollary D.10.1. *Assume $\widehat{E}_p$ satisfies $\|\widehat{E}_p\{\mathbf{a}(Y)\mid \mathbf{x}\}-\mathrm{E}_p\{\mathbf{a}(Y)\mid \mathbf{x}\}\|_\infty = o_p(n_1^{-1/4})$ for any bounded function $\mathbf{a}(y)$. If $\rho^*(y)=\rho(y)$, under Conditions (B1)-(B7),*

$$\sqrt{n_1}(\widetilde{\boldsymbol{\theta}}_{\mathrm{eff}}-\boldsymbol{\theta}) \to N\left[0, var\{\sqrt{\pi}\boldsymbol{\phi}_{\mathrm{eff}}(\mathbf{X},R,RY)\}\right]$$

in distribution as $n_1 \to \infty$.

D.11 Proofs of Section D.9

D.11.1 Derivation of influence functions

We first state and prove a result regarding the tangent space.

Proposition D.11.1. *The tangent space of (5.1) is* $\mathcal{T} \equiv \mathcal{T}_{\boldsymbol{\alpha}} \oplus (\mathcal{T}_{\boldsymbol{\beta}} + \mathcal{T}_{\boldsymbol{\gamma}})$*, where*

$$\begin{aligned}
\mathcal{T}_{\boldsymbol{\alpha}} &= [r\mathbf{a}_1(y) : \mathrm{E}_p\{\mathbf{a}_1(Y)\} = \mathbf{0}], \\
\mathcal{T}_{\boldsymbol{\beta}} &= [r\mathbf{a}_2(\mathbf{x}, y) + (1-r)\mathrm{E}_q\{\mathbf{a}_2(\mathbf{x}, Y) \mid \mathbf{x}\} : E\{\mathbf{a}_2(\mathbf{X}, y) \mid y\} = \mathbf{0}], \\
\mathcal{T}_{\boldsymbol{\gamma}} &= [(1-r)\mathrm{E}_q\{\mathbf{a}_3(Y) \mid \mathbf{x}\} : \mathrm{E}_q\{\mathbf{a}_3(Y)\} = \mathbf{0}].
\end{aligned}$$

Proof. Consider a parametric submodel of (5.1),

$$f_{\mathbf{X},R,RY}(\mathbf{x}, r, ry, \boldsymbol{\delta}) = \pi^r(1-\pi)^{1-r}\{g(\mathbf{x}, y, \boldsymbol{\beta})p_Y(y, \boldsymbol{\alpha})\}^r \left\{\int g(\mathbf{x}, y, \boldsymbol{\beta})q_Y(y, \boldsymbol{\gamma})dy\right\}^{1-r} \quad \text{(D.18)}$$

where $\boldsymbol{\delta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. We can derive that the score function associated with an arbitrary $\boldsymbol{\delta}$ is $\mathbf{S}_{\boldsymbol{\delta}} \equiv (\mathbf{S}_{\boldsymbol{\alpha}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, where

$$\begin{aligned}
\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x}, r, ry) &\equiv r\mathbf{a}_{\boldsymbol{\alpha}}(y), \\
\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x}, r, ry) &\equiv r\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{x}, y) + (1-r)\mathrm{E}_q\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{x}, Y) \mid \mathbf{x}\}, \\
\mathbf{S}_{\boldsymbol{\gamma}}(\mathbf{x}, r, ry) &\equiv (1-r)\mathrm{E}_q\{\mathbf{a}_{\boldsymbol{\gamma}}(Y) \mid \mathbf{x}\},
\end{aligned}$$

$\mathrm{E}_p\{\mathbf{a}_{\boldsymbol{\alpha}}(Y)\} = \mathbf{0}$, $\mathrm{E}_p\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, y) \mid y\} = \mathrm{E}_q\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, y) \mid y\} = \mathrm{E}\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, y) \mid y\} = \mathbf{0}$, and $\mathrm{E}_q\{\mathbf{a}_{\boldsymbol{\gamma}}(Y)\} = \mathbf{0}$. The above derivation directly leads to Proposition D.11.1. □

We now establish $\mathcal{F}$ in Proposition D.11.2.

Proposition D.11.2. *The set of the influence functions for* $\boldsymbol{\theta}$ *is*

$$\begin{aligned}
\mathcal{F} \equiv \Bigg[& \frac{r}{\pi}[\rho(y)\mathbf{A}\{\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta}) - \mathbf{b}(\mathbf{x})\} + \mathbf{c}] + \frac{1-r}{1-\pi}\{\mathbf{A}\mathbf{b}(\mathbf{x}) - \mathbf{c}\} \\
& : E\{\mathbf{b}(\mathbf{X}) \mid y\} = E\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}, \forall \mathbf{c}\Bigg].
\end{aligned}$$

Proof. Recall the submodel (D.18), then it can be verified that

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} = \mathbf{0},$$

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}^{\mathrm{T}}} = \mathbf{A}\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\mathbf{a}_{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{X}, Y)\} = \mathbf{A}\mathrm{E}_p\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\rho(Y)\mathbf{a}_{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{X}, Y)\},$$
$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} = \mathbf{A}\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\mathbf{a}_{\boldsymbol{\gamma}}^{\mathrm{T}}(Y)\},$$

where $\mathrm{E}\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, y) \mid y\} = \mathbf{0}$ and $\mathrm{E}_q\{\mathbf{a}_{\boldsymbol{\gamma}}(Y)\} = \mathbf{0}$. Now let $\boldsymbol{\phi}(\mathbf{x}, r, ry)$ be

$$\boldsymbol{\phi}(\mathbf{x}, r, ry) \equiv \frac{r}{\pi}\boldsymbol{\phi}_1(\mathbf{x}, y) + \frac{1-r}{1-\pi}\boldsymbol{\phi}_2(\mathbf{x}).$$

For $\boldsymbol{\phi}(\mathbf{x}, r, ry)$ to be an influence function, it must satisfy

$$\mathrm{E}(\boldsymbol{\phi}) = \mathrm{E}_p\{\boldsymbol{\phi}_1(\mathbf{X}, Y)\} + \mathrm{E}_q\{\boldsymbol{\phi}_2(\mathbf{X})\} = \mathbf{0} \tag{D.19}$$

and $\mathrm{E}(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\delta}}^{\mathrm{T}}) = \partial\boldsymbol{\theta}/\partial\boldsymbol{\delta}^{\mathrm{T}}$, where $\boldsymbol{\delta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{S}_{\boldsymbol{\delta}} = (\mathbf{S}_{\boldsymbol{\alpha}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$ is the score function of the submodel (D.18). First, $\mathrm{E}(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\alpha}}^{\mathrm{T}}) = \partial\boldsymbol{\theta}/\partial\boldsymbol{\alpha}^{\mathrm{T}}$ is equivalent to

$$\mathrm{E}\{\boldsymbol{\phi}_1(\mathbf{X}, y) \mid y\} = \mathbf{c} \tag{D.20}$$

for some constant vector $\mathbf{c}$. In addition, since $\mathrm{E}(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}) = \mathrm{E}_p[\{\boldsymbol{\phi}_1(\mathbf{X}, Y) + \boldsymbol{\phi}_2(\mathbf{X})\rho(Y)\}\mathbf{a}_{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{X}, Y)]$ and $\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{x}, y)$ satisfies $\mathrm{E}\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, y) \mid y\} = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}) = \partial\boldsymbol{\theta}/\partial\boldsymbol{\beta}^{\mathrm{T}}$ implies $\boldsymbol{\phi}_1(\mathbf{x}, y) + \boldsymbol{\phi}_2(\mathbf{x})\rho(y) = \mathbf{A}\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta})\rho(y) + \mathbf{a}(y)$ for some $\mathbf{a}(y)$. Then (D.20) yields

$$\boldsymbol{\phi}_1(\mathbf{x}, y) = \rho(y)[\mathbf{A}\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta}) - \boldsymbol{\phi}_2(\mathbf{x}) - \mathrm{E}\{\mathbf{A}\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) - \boldsymbol{\phi}_2(\mathbf{X}) \mid y\}] + \mathbf{c}.$$

Also, noting that $\mathrm{E}(\boldsymbol{\phi}\mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}}) = \partial\boldsymbol{\theta}/\partial\boldsymbol{\gamma}^{\mathrm{T}}$ is equivalent to $\mathrm{E}\{\boldsymbol{\phi}_2(\mathbf{X}) \mid y\} = \mathrm{E}\{\mathbf{A}\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\} + \mathbf{c}^*$ for some constant vector $\mathbf{c}^*$, we have from (D.19) and (D.20) that $\mathbf{c}^* = -\mathbf{c}$. Therefore, defining $\mathbf{b}(\mathbf{x}) \equiv \mathbf{A}^{-1}\{\boldsymbol{\phi}_2(\mathbf{x}) + \mathbf{c}\}$, the summary description of the influence function is

$$\boldsymbol{\phi}(\mathbf{x}, r, ry) = \frac{r}{\pi}[\rho(y)\mathbf{A}\{\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta}) - \mathbf{b}(\mathbf{x})\} + \mathbf{c}] + \frac{1-r}{1-\pi}\{\mathbf{A}\mathbf{b}(\mathbf{x}) - \mathbf{c}\},$$

where $\mathbf{b}(\mathbf{x})$ satisfies $\mathrm{E}\{\mathbf{b}(\mathbf{X}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$ and $\mathbf{c}$ is a constant vector. □

D.11.2 Proof of Proposition D.9.1

Note that

$$\frac{\mathrm{E}_p\{a(\mathbf{x}, Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho(Y) \mid \mathbf{x}\}} = \frac{\int a(\mathbf{x}, y)\rho(y)g(\mathbf{x}, y)p_Y(y)dy}{\int \rho(y)g(\mathbf{x}, y)p_Y(y)dy} = \frac{\int a(\mathbf{x}, y)g(\mathbf{x}, y)q_Y(y)dy}{\int g(\mathbf{x}, y)q_Y(y)dy}$$

$$= \mathrm{E}_q\{a(\mathbf{x}, Y) \mid \mathbf{x}\},$$

then $\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}, r, ry)$ can be alternatively written as

$$\begin{aligned}\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}, r, ry) &= \frac{r}{\pi}\rho(y)\mathbf{A}\left[\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta}) - \frac{\mathrm{E}_q\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho(Y) + \mathbf{a}(Y) \mid \mathbf{x}\}}{\mathrm{E}_q\{\rho(Y) + \pi/(1-\pi) \mid \mathbf{x}\}}\right] \\ &\quad + \frac{1-r}{1-\pi}\frac{\mathbf{A}\mathrm{E}_q\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho(Y) + \mathbf{a}(Y) \mid \mathbf{x}\}}{\mathrm{E}_q\{\rho(Y) + \pi/(1-\pi) \mid \mathbf{x}\}},\end{aligned}$$

where $\mathbf{a}(y)$ satisfies

$$\mathrm{E}\left[\frac{\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\rho(Y) + \mathbf{a}(Y) \mid \mathbf{X}\}}{\mathrm{E}_q\{\rho(Y) + \pi/(1-\pi) \mid \mathbf{X}\}} \mid y\right] = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}.$$

First, it is immediate that $\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}, r, ry)$ is an influence function for $\boldsymbol{\theta}$, i.e., belongs to $\mathcal{F}$ given in Proposition D.11.2 from letting

$$\begin{aligned}\mathbf{b}(\mathbf{x}) &\equiv \frac{\mathrm{E}_q\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho(Y) + \mathbf{a}(Y) \mid \mathbf{x}\}}{\mathrm{E}_q\{\rho(Y) + \pi/(1-\pi) \mid \mathbf{x}\}}, \\ \mathbf{c} &\equiv \mathbf{0}.\end{aligned}$$

Next, we show that $\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}, r, ry)$ is in the tangent space $\mathcal{T}$ given in Proposition D.11.1. We decompose $\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}, r, ry)$ into

$$\boldsymbol{\phi}_{\text{eff}}(\mathbf{x}, r, ry) = r\{\mathbf{a}_1(y) + \mathbf{a}_2(\mathbf{x}, y)\} + (1-r)[\mathrm{E}_q\{\mathbf{a}_2(\mathbf{x}, Y) \mid \mathbf{x}\} + \mathrm{E}_q\{\mathbf{a}_3(Y) \mid \mathbf{x}\}],$$

where

$$\begin{aligned}\mathbf{a}_1(y) &\equiv \mathbf{0}, \\ \mathbf{a}_2(\mathbf{x}, y) &\equiv \frac{1}{\pi}\rho(y)\mathbf{A}\{\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta}) - \mathbf{b}(\mathbf{x})\}, \\ \mathbf{a}_3(y) &\equiv \frac{1}{\pi}\mathbf{A}\mathbf{a}(y).\end{aligned}$$

Then it is easy to see that $r\mathbf{a}_1(y) \in \mathcal{T}_{\boldsymbol{\alpha}}$, and $r\mathbf{a}_2(\mathbf{x}, y) + (1-r)\mathrm{E}_q\{\mathbf{a}_2(\mathbf{x}, Y) \mid \mathbf{x}\} \in \mathcal{T}_{\boldsymbol{\beta}}$ because $\mathrm{E}\{\mathbf{b}(\mathbf{X}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$. Further, $(1-r)\mathrm{E}_q\{\mathbf{a}_3(Y) \mid \mathbf{x}\} \in \mathcal{T}_{\boldsymbol{\gamma}}$ since

$$\begin{aligned}\mathrm{E}_q\{\mathbf{a}_3(Y)\} &= \frac{1}{\pi}\mathbf{A}\mathrm{E}_q[\mathrm{E}_q\{\mathbf{a}(Y) \mid \mathbf{X}\}] \\ &= \frac{1}{\pi}\mathbf{A}\mathrm{E}_q\{\mathbf{b}(\mathbf{X})\mathrm{E}_q\{\rho(Y) \mid \mathbf{X}\}] + \frac{1}{1-\pi}\mathbf{A}\mathrm{E}_q\{\mathbf{b}(\mathbf{X})\}\end{aligned}$$

$$
\begin{aligned}
& -\frac{1}{\pi}\mathbf{A}\mathrm{E}_q[\mathrm{E}_q\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\rho(Y)\mid\mathbf{X}\}] \\
= \;& \frac{1}{\pi}\mathbf{A}\mathrm{E}_q[\mathrm{E}_q\{\mathbf{b}(\mathbf{X})-\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\mid Y\}\rho(Y)] \\
= \;& \mathbf{0},
\end{aligned}
$$

where the third equality holds since $\mathrm{E}_q\{\mathbf{b}(\mathbf{X})\}=\mathrm{E}_q\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\}=\mathbf{0}$ by the definition of $\boldsymbol{\theta}$. Hence $\boldsymbol{\phi}_{\text{eff}}(\mathbf{x},r,ry)$ belongs to $\mathcal{T}$. □

D.11.3 Proof of Proposition D.9.2

The invertibility of $\mathrm{E}\{\partial\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{X},R,RY,\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\}$ implies that $\mathrm{E}\{\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{X},R,RY,\boldsymbol{\theta})\}=\mathbf{0}$ has a unique solution in the neighborhood of $\boldsymbol{\theta}$, and the existence of $\partial\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{x},r,ry,\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}$ automatically implies that $\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{x},r,ry,\boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta}$. Then following Theorem 2.6 of Newey & McFadden (1994), it suffices to show that $\mathrm{E}\{\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{X},R,RY,\boldsymbol{\theta})\}=\mathbf{0}$. It is immediate that $\mathrm{E}\{\mathbf{b}^{**}(\mathbf{X},\boldsymbol{\theta})\mid y\}=\mathrm{E}\{\mathbf{U}(\mathbf{X},y,\boldsymbol{\theta})\mid y\}$ from the definition of $\mathbf{b}^{**}(\mathbf{x},\boldsymbol{\theta})$. Hence,

$$
\begin{aligned}
& \mathrm{E}\{\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{X},R,RY,\boldsymbol{\theta})\} \\
= \;& \mathbf{A}^{**}\mathrm{E}\left[\frac{R}{\pi}\rho^*(Y)\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})-\mathbf{b}^{**}(\mathbf{X},\boldsymbol{\theta})\}+\frac{1-R}{1-\pi}\mathbf{b}^{**}(\mathbf{X},\boldsymbol{\theta})\right] \\
= \;& \mathbf{A}^{**}\mathrm{E}_p\left[\rho^*(Y)\mathrm{E}\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})-\mathbf{b}^{**}(\mathbf{X},\boldsymbol{\theta})\mid Y\}\right]+\mathbf{A}^{**}\mathrm{E}_q\left[\mathrm{E}\left\{\mathbf{b}^{**}(\mathbf{X},\boldsymbol{\theta})\mid Y\right\}\right] \\
= \;& \mathbf{A}^{**}\mathrm{E}_q\left[\mathrm{E}\left\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\mid Y\right\}\right] \\
= \;& \mathbf{0}
\end{aligned}
$$

by the definition of $\boldsymbol{\theta}$. □

D.11.4 Proof of Lemma D.9.1

From the definition of $\mathcal{L}^{**}(\mathbf{a})$ and Conditions (B4)-(B5), it is immediate that there exists a constant $0<c_2<\infty$ such that $\|\mathcal{L}^{**}(\mathbf{a})\|_\infty\leqslant c_2\|\mathbf{a}\|_\infty$. Now we show there is a constant $0<c_1<\infty$ such that $c_1\|\mathbf{a}\|_\infty\leqslant\|\mathcal{L}^{**}(\mathbf{a})\|_\infty$. Note that if $\mathcal{L}^{**}$ is invertible, by the bounded inverse theorem we have $\|\mathcal{L}^{**-1}(\mathbf{v})\|_\infty\leqslant c_1^{-1}\|\mathbf{v}\|_\infty$ for some constant $0<c_1<\infty$, i.e., $c_1\|\mathbf{a}\|_\infty\leqslant\|\mathcal{L}^{**}(\mathbf{a})\|_\infty$. Hence it suffices to show that $\mathcal{L}^{**}$ is invertible. We prove this by contradiction. Suppose there are $\mathbf{a}_1(y,\boldsymbol{\theta})$ and $\mathbf{a}_2(y,\boldsymbol{\theta})$ such that $\mathcal{L}^{**}(\mathbf{a}_1)(y,\boldsymbol{\theta})=\mathcal{L}^{**}(\mathbf{a}_2)(y,\boldsymbol{\theta})=\mathbf{v}^{**}(y,\boldsymbol{\theta})$ and $\mathbf{a}_1\neq\mathbf{a}_2$. Then by Conditions (B1)-(B3), we have

$$
\mathbf{A}^{**}\mathrm{E}^{\star}_p\{\mathbf{a}_1(Y,\boldsymbol{\theta})\rho^*(Y)\mid\mathbf{x}\}\neq\mathbf{A}^{**}\mathrm{E}^{\star}_p\{\mathbf{a}_2(Y,\boldsymbol{\theta})\rho^*(Y)\mid\mathbf{x}\}.
$$

Now, the efficient score calculated under the posited models is

$$\begin{aligned}\boldsymbol{\phi}^{**}_{\text{eff}}(\mathbf{x}, r, ry, \boldsymbol{\theta}) &= \frac{r}{\pi}\rho^*(y)\mathbf{A}^{**}\left[\mathbf{U}(\mathbf{x}, y, \boldsymbol{\theta}) - \frac{\mathrm{E}^{\star}_p\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho^{*2}(Y) + \mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}^{\star}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}\right] \\ &\quad + \frac{1-r}{1-\pi}\frac{\mathbf{A}^{**}\mathrm{E}^{\star}_p\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho^{*2}(Y) + \mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}^{\star}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}},\end{aligned}$$

where $\mathbf{a}(y, \boldsymbol{\theta})$ satisfies $\mathcal{L}^{**}(\mathbf{a})(y, \boldsymbol{\theta}) = \mathbf{v}^{**}(y, \boldsymbol{\theta})$. Then letting $\mathbf{a} = \mathbf{a}_1$ and $\mathbf{a} = \mathbf{a}_2$ gives two distinct efficient scores, which contradicts the uniqueness of the efficient score. Therefore, there is a unique solution $\mathbf{a}^{**}(y, \boldsymbol{\theta})$ for $\mathcal{L}^{**}(\mathbf{a}^{**})(y, \boldsymbol{\theta}) = \mathbf{v}^{**}(y, \boldsymbol{\theta})$, hence $\mathcal{L}^{**}$ is invertible. □

D.11.5 Proof of Theorem D.9.1

We define

$$\mathbf{b}^{**}(\mathbf{x}, \mathbf{a}, \boldsymbol{\zeta}, \boldsymbol{\theta}) \equiv \frac{\mathrm{E}^{\star}_p\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho^{*2}(Y) + \mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}, \boldsymbol{\zeta}\}}{\mathrm{E}^{\star}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}, \boldsymbol{\zeta}\}},$$

then under Conditions (B1), (B4)-(B7) and $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$,

$$\begin{aligned}&\widehat{\mathcal{L}}^{**}(\mathbf{a}^{**})(y, \boldsymbol{\theta}) - \widehat{\mathbf{v}}^{**}(y, \boldsymbol{\theta}) \\ =\ & n_1^{-1}\sum_{i=1}^{n} r_i K_h(y - y_i)\{\mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \widehat{\boldsymbol{\zeta}}, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{x}_i, y, \boldsymbol{\theta})\} \\ =\ & n_1^{-1}\sum_{i=1}^{n} r_i K_h(y - y_i)\left\{\mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) + \frac{\partial \mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2})\right. \\ & \left. - \mathbf{U}(\mathbf{x}_i, y, \boldsymbol{\theta})\right\} \\ =\ & n_1^{-1}\sum_{i=1}^{n}\left\{\mathcal{L}^{**}_{i,h}(\mathbf{a}^{**})(y, \boldsymbol{\theta}) - \mathbf{v}^{**}_{i,h}(y, \boldsymbol{\theta})\right\} \\ & + \left[p_Y(y)\frac{\partial \mathrm{E}\{\mathbf{b}^{**}(\mathbf{X}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) \mid y\}}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} + o_p(1)\right](\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2}) \\ =\ & n_1^{-1}\sum_{i=1}^{n}\left\{\mathcal{L}^{**}_{i,h}(\mathbf{a}^{**})(y, \boldsymbol{\theta}) - \mathbf{v}^{**}_{i,h}(y, \boldsymbol{\theta})\right\} + o_p(n_1^{-1/2})\end{aligned}$$

uniformly in $(y, \boldsymbol{\theta})$ since $\mathrm{E}\{\mathbf{b}^{**}(\mathbf{X}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$. In addition, for any bounded function $\mathbf{a}(y, \boldsymbol{\theta})$, $\|(\widehat{\mathcal{L}}^{**} - \mathcal{L}^{**})(\mathbf{a})\|_\infty = O_p\left\{(n_1 h)^{-1/2}\log n_1 + h^2\right\} = o_p(n_1^{-1/4})$ under Conditions (B1), (B4)-(B7) and the assumption $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$. Similarly, $\|\widehat{\mathbf{v}}^{**} - \mathbf{v}^{**}\|_\infty = o_p(n_1^{-1/4})$. Then using that $\mathcal{L}^{**-1}$ is a bounded linear operator by Lemma

D.9.1, $\widehat{\mathbf{a}}^{**}(y,\boldsymbol{\theta})$ can be expressed as

$$\begin{aligned}
&\widehat{\mathbf{a}}^{**}(y,\boldsymbol{\theta}) \\
=\ & \{\mathcal{L}^{**} + (\widehat{\mathcal{L}}^{**} - \mathcal{L}^{**})\}^{-1}(\widehat{\mathbf{v}}^{**})(y,\boldsymbol{\theta}) \\
=\ & \{\mathcal{L}^{**-1} - \mathcal{L}^{**-1}(\widehat{\mathcal{L}}^{**} - \mathcal{L}^{**})\mathcal{L}^{**-1}\}\{\mathbf{v}^{**} + (\widehat{\mathbf{v}}^{**} - \mathbf{v}^{**})\}(y,\boldsymbol{\theta}) + o_p(n_1^{-1/2}) \\
=\ & \mathbf{a}^{**}(y,\boldsymbol{\theta}) + \mathcal{L}^{**-1}(\widehat{\mathbf{v}}^{**} - \mathbf{v}^{**})(y,\boldsymbol{\theta}) - \{\mathcal{L}^{**-1}(\widehat{\mathcal{L}}^{**} - \mathcal{L}^{**})\}(\mathbf{a}^{**})(y,\boldsymbol{\theta}) + o_p(n_1^{-1/2}) \\
=\ & \mathbf{a}^{**}(y,\boldsymbol{\theta}) + \mathcal{L}^{**-1}\{\widehat{\mathbf{v}}^{**} - \widehat{\mathcal{L}}^{**}(\mathbf{a}^{**})\}(y,\boldsymbol{\theta}) + o_p(n_1^{-1/2}) \\
=\ & \mathbf{a}^{**}(y,\boldsymbol{\theta}) + n_1^{-1}\sum_{i=1}^{n}\mathbf{g}_{i,h}(y,\boldsymbol{\theta}) + o_p(n_1^{-1/2}) \qquad \text{(D.21)}
\end{aligned}$$

uniformly in y, where $\mathbf{g}_{i,h}(y,\boldsymbol{\theta}) \equiv \mathcal{L}^{**-1}\{\mathbf{v}_{i,h}^{**} - \mathcal{L}_{i,h}^{**}(\mathbf{a}^{**})\}(y,\boldsymbol{\theta})$. Note that $\|E(\mathbf{V}_{i,h}^{**}) - \mathbf{v}^{**}\|_\infty = O(h^2)$ by Conditions (B5)-(B6), then since $\mathcal{L}^{**-1}$ is a bounded linear operator by Lemma D.9.1,

$$\|\mathrm{E}\{\mathcal{L}^{**-1}(\mathbf{V}_{i,h}^{**})\} - \mathcal{L}^{**-1}(\mathbf{v}^{**})\|_\infty = \|\mathcal{L}^{**-1}\{\mathrm{E}(\mathbf{V}_{i,h}^{**}) - \mathbf{v}^{**}\}\|_\infty = O(h^2).$$

Similarly, $\|\mathrm{E}\{(\mathcal{L}^{**-1}\mathcal{L}_{i,h}^{**})(\mathbf{a}^{**})\} - \mathbf{a}^{**}\|_\infty = O(h^2)$. Using $\mathcal{L}^{**-1}(\mathbf{v}^{**})(y,\boldsymbol{\theta}) = \mathbf{a}^{**}(y,\boldsymbol{\theta})$, we further get

$$\begin{aligned}
\|\mathrm{E}(\mathbf{G}_{i,h})\|_\infty &= \|\mathrm{E}\{\mathcal{L}^{**-1}(\mathbf{V}_{i,h}^{**})\} - \mathrm{E}\{(\mathcal{L}^{**-1}\mathcal{L}_{i,h}^{**})(\mathbf{a}^{**})\} - \{\mathcal{L}^{**-1}(\mathbf{v}^{**}) - \mathbf{a}^{**}\}\|_\infty \\
&\leqslant \|\mathrm{E}\{\mathcal{L}^{**-1}(\mathbf{V}_{i,h}^{**})\} - \mathcal{L}^{**-1}(\mathbf{v})\|_\infty + \|\mathrm{E}\{(\mathcal{L}^{**-1}\mathcal{L}_{i,h})(\mathbf{a}^{**})\} - \mathbf{a}^{**}\|_\infty \\
&= O(h^2), \qquad \text{(D.22)}
\end{aligned}$$

and for $i \neq j$,

$$\begin{aligned}
& \left\| \mathrm{E}\left[\frac{\mathrm{E}_p^\star\{\mathbf{G}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \mid \mathbf{x}_i \right] \right\|_\infty \\
=\ & \left\| \frac{\mathrm{E}_p^\star\{\mathrm{E}(\mathbf{G}_{j,h})(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \right\|_\infty \\
\leqslant\ & \|\mathrm{E}(\mathbf{G}_{j,h})\|_\infty \left| \frac{\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \right| \\
=\ & O(h^2). \qquad \text{(D.23)}
\end{aligned}$$

Also, we have

$$\left\| \frac{\mathrm{E}_p^\star\{\mathbf{a}(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \right\|_\infty$$

$$
\begin{aligned}
&\leqslant \|\mathbf{a}\|_\infty \left| \frac{\mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \right| \\
&= O(\|\mathbf{a}\|_\infty),
\end{aligned}
$$

and from (D.21)

$$
\begin{aligned}
&\left\| \frac{\partial \mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) - \frac{\partial \mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \right\|_\infty \\
= &\left\| \frac{\partial}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} \left[\frac{\mathrm{E}_p^\star\{(\widehat{\mathbf{a}}^{**} - \mathbf{a}^{**})(Y, \boldsymbol{\theta}) \rho^*(Y) \mid \mathbf{x}_i, \boldsymbol{\zeta}\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i, \boldsymbol{\zeta}\}} \right] (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \right\|_\infty \\
= &\Bigg\| \frac{\mathrm{E}_p^\star\{(\widehat{\mathbf{a}}^{**} - \mathbf{a}^{**})(Y, \boldsymbol{\theta}) \rho^*(Y) \mathbf{S}_{\boldsymbol{\zeta}}^{\star\mathrm{T}}(Y, \mathbf{x}_i, \boldsymbol{\zeta}) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \\
&- \mathrm{E}_p^\star\{(\widehat{\mathbf{a}}^{**} - \mathbf{a}^{**})(Y, \boldsymbol{\theta}) \rho^*(Y) \mid \mathbf{x}_i\} \\
&\times \frac{\mathrm{E}_p[\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y)\} \mathbf{S}_{\boldsymbol{\zeta}}^{\star\mathrm{T}}(Y, \mathbf{x}_i, \boldsymbol{\zeta}) \mid \mathbf{x}_i]}{[\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}]^2} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \Bigg\|_\infty \\
\leqslant &\|\widehat{\mathbf{a}}^{**} - \mathbf{a}^{**}\|_\infty \frac{\mathrm{E}_p^\star\{\rho^*(Y) \|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y, \mathbf{x}_i, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}_i\} \|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \\
&+ \|\widehat{\mathbf{a}}^{**} - \mathbf{a}^{**}\|_\infty \mathrm{E}_p^\star\{\rho^*(Y) \mid \mathbf{x}_i\} \\
&\times \frac{\mathrm{E}_p^\star[\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y)\} \|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y, \mathbf{x}_i, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}_i] \|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2}{[\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}]^2} \\
= &\, o_p(n_1^{-1/4}) O_p(n_1^{-1/2}) = o_p(n_1^{-1/2}),
\end{aligned}
$$

because $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$, $\mathrm{E}_p^\star\{\|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y, \mathbf{x}, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}\}$ is bounded, and $\|\widehat{\mathbf{a}}^{**} - \mathbf{a}^{**}\|_\infty = O_p\left\{(n_1 h)^{-1/2} \log n_1 + h^2\right\} = o_p(n^{-1/4})$ by (D.21) and Condition (B7). Hence, using (D.21) we get

$$
\begin{aligned}
&\mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \widehat{\boldsymbol{\zeta}}, \boldsymbol{\theta}) \qquad\qquad (\text{D.24}) \\
= &\, \mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) + \frac{\mathrm{E}_p^\star\{(\widehat{\mathbf{a}}^{**} - \mathbf{a}^{**})(Y, \boldsymbol{\theta}) \rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \\
&+ \frac{\partial \mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2}) \\
= &\, \mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) + n_1^{-1} \sum_{j=1}^{n} \frac{\mathrm{E}_p^\star\{\mathbf{g}_{j,h}(Y, \boldsymbol{\theta}) \rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p^\star\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \\
&+ \frac{\partial \mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} (\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2})
\end{aligned}
$$

uniformly in $\mathbf{x}_i$ by Condition (B5).

Now we show the consistency of $\widehat{\boldsymbol{\theta}}$ by following Theorem 2.1 of Newey & McFadden (1994). We can view the problem of finding the solution for $\mathrm{E}\{\boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\} = \mathbf{0}$ as maximizing the objective function $Q_0(\boldsymbol{\theta}) \equiv -\|\mathrm{E}\{\boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\}\|_2^2$, then Theorem 2.1 of Newey & McFadden (1994) is directly applicable. It is immediate that $\mathrm{E}\{\mathbf{b}^{**}(\mathbf{X}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$ from the definition of $\mathbf{b}^{**}(\mathbf{x}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta})$. Hence,

$$
\begin{aligned}
& \mathrm{E}\{\boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\} \\
= \; & \mathbf{A}^{**}\mathrm{E}\left[\frac{R}{\pi}\rho^*(Y)\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) - \mathbf{b}^{**}(\mathbf{X}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta})\} + \frac{1-R}{1-\pi}\mathbf{b}^{**}(\mathbf{X}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta})\right] \\
= \; & \mathbf{A}^{**}\mathrm{E}_p\left[\rho^*(Y)\mathrm{E}\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) - \mathbf{b}^{**}(\mathbf{X}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) \mid Y\}\right] \\
& + \mathbf{A}^{**}\mathrm{E}_q\left[\mathrm{E}\left\{\mathbf{b}^{**}(\mathbf{X}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) \mid Y\right\}\right] \\
= \; & \mathbf{A}^{**}\mathrm{E}_q\left[\mathrm{E}\left\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) \mid Y\right\}\right] \\
= \; & \mathbf{0},
\end{aligned}
$$

where the last step is by the definition of $\boldsymbol{\theta}$. Also, Condition (B1) implies that $\boldsymbol{\theta}$ is the unique solution for $\mathrm{E}\{\boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\} = \mathbf{0}$ in the neighborhood of $\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ which is compact, and $\boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{x}, r, ry, \boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta}$. Therefore, it suffices to show that the estimating equation converges in probability to $\mathrm{E}\{\boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\}$ uniformly in $\boldsymbol{\theta}$. Using (D.24), the estimating equation can be expressed as

$$
\begin{aligned}
& \mathbf{A}^{**}n^{-1}\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\left\{\mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) - \mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \widehat{\boldsymbol{\zeta}}, \boldsymbol{\theta})\right\} + \frac{1-r_i}{1-\pi}\mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \widehat{\boldsymbol{\zeta}}, \boldsymbol{\theta})\right] \\
= \; & n^{-1}\sum_{i=1}^{n}\boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{x}_i, r_i, r_i y_i, \boldsymbol{\theta}) \\
& + \mathbf{A}^{**}n^{-1}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i) - \frac{1-r_i}{1-\pi}\right\}\left\{\mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) - \mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \widehat{\boldsymbol{\zeta}}, \boldsymbol{\theta})\right\} \\
= \; & n^{-1}\sum_{i=1}^{n}\boldsymbol{\phi}^{**}_{\mathrm{eff}}(\mathbf{x}_i, r_i, r_i y_i, \boldsymbol{\theta}) - \mathbf{A}^{**}\{T_1(\boldsymbol{\theta}) + T_2(\boldsymbol{\theta})\} + o_p(n_1^{-1/2}),
\end{aligned}
\tag{D.25}
$$

where

$$
T_1(\boldsymbol{\theta}) \equiv n^{-1}n_1^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i) - \frac{1-r_i}{1-\pi}\right\}\frac{\mathrm{E}^{\star}_p\{\mathbf{g}_{j,h}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}^{\star}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}},
$$

$$T_2(\boldsymbol{\theta}) \equiv n^{-1}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i)-\frac{1-r_i}{1-\pi}\right\}\frac{\partial \mathbf{b}^{**\star}(\mathbf{x}_i,\mathbf{a}^{**\star},\boldsymbol{\zeta},\boldsymbol{\theta})}{\partial\boldsymbol{\zeta}^{\mathrm{T}}}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}).$$

Using the property of the U-statistic, Condition (B7), and (D.23), we can rewrite $T_1(\boldsymbol{\theta})$ as

$$\begin{aligned}
& T_1(\boldsymbol{\theta}) \\
= \; & n_1^{-1}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i)-\frac{1-r_i}{1-\pi}\right\}\mathrm{E}\left[\frac{\mathrm{E}_p^{\star}\{\mathbf{G}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid\mathbf{x}_i\}}{\mathrm{E}_p^{\star}\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid\mathbf{x}_i\}}\mid\mathbf{x}_i,r_i,r_iy_i\right] \\
& +n_1^{-1}\sum_{j=1}^{n}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\rho^*(Y_i)-\frac{1-R_i}{1-\pi}\right\}\frac{\mathrm{E}_p^{\star}\{\mathbf{g}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid\mathbf{X}_i\}}{\mathrm{E}_p^{\star}\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid\mathbf{X}_i\}}\mid\mathbf{x}_j,r_j,r_jy_j\right] \\
& -n^{1/2}n_1^{-1}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\rho^*(Y_i)-\frac{1-R_i}{1-\pi}\right\}\frac{\mathrm{E}_p^{\star}\{\mathbf{G}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid\mathbf{X}_i\}}{\mathrm{E}_p^{\star}\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid\mathbf{X}_i\}}\right]+O_p(n_1^{-1}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\rho^*(Y_i)-\frac{1-R_i}{1-\pi}\right\}\frac{\mathrm{E}_p^{\star}\{\mathbf{g}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid\mathbf{X}_i\}}{\mathrm{E}_p^{\star}\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid\mathbf{X}_i\}}\mid\mathbf{x}_j,r_j,r_jy_j\right] \\
& +O_p(n_1^{-1}nh^2+n_1^{-1}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}\int\{\rho^*(y)p_Y(y)-q_Y(y)\}\mathrm{E}\left[\frac{\mathrm{E}_p^{\star}\{\mathbf{g}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid\mathbf{X}\}}{\mathrm{E}_p^{\star}\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid\mathbf{X}\}}\mid y\right]dy+o_p(n_1^{-1/2}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}\int\mathcal{L}^{**\star}(\mathbf{g}_{j,h})(y,\boldsymbol{\theta})\left\{\rho^*(y)-\rho(y)\right\}dy+o_p(n_1^{-1/2}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}\int\{\mathbf{v}_{j,h}^{**\star}(y,\boldsymbol{\theta})-\mathcal{L}_{j,h}^{**\star}(\mathbf{a}^{**\star})(y,\boldsymbol{\theta})\}\left\{\rho^*(y)-\rho(y)\right\}dy+o_p(n_1^{-1/2}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}r_j\int K_h(y-y_j)\{\mathbf{U}(\mathbf{x}_j,y,\boldsymbol{\theta})-\mathbf{b}^{**\star}(\mathbf{x}_j,\mathbf{a}^{**\star},\boldsymbol{\zeta},\boldsymbol{\theta})\}\left\{\rho^*(y)-\rho(y)\right\}dy+o_p(n_1^{-1/2}),
\end{aligned}$$

where

$$\begin{aligned}
& \int K_h(y-y_j)\{\mathbf{U}(\mathbf{x}_j,y,\boldsymbol{\theta})-\mathbf{b}^{**\star}(\mathbf{x}_j,\mathbf{a}^{**\star},\boldsymbol{\zeta},\boldsymbol{\theta})\}\left\{\rho^*(y)-\rho(y)\right\}dy \\
= \; & \{\mathbf{U}(\mathbf{x}_j,y_j,\boldsymbol{\theta})-\mathbf{b}^{**\star}(\mathbf{x}_j,\mathbf{a}^{**\star},\boldsymbol{\zeta},\boldsymbol{\theta})\}\{\rho^*(y_j)-\rho(y_j)\} \\
& +\big[\{\mathbf{U}(\mathbf{x}_j,y_j,\boldsymbol{\theta})-\mathbf{b}^{**\star}(\mathbf{x}_j,\mathbf{a}^{**\star},\boldsymbol{\zeta},\boldsymbol{\theta})\}\{\rho^{*\prime\prime}(y_j)-\rho''(y_j)\}+2\mathbf{U}'_y(\mathbf{x}_j,y_j,\boldsymbol{\theta})\{\rho^{*\prime}(y_j)-\rho'(y_j)\} \\
& +\mathbf{U}''_{yy}(\mathbf{x}_j,y_j,\boldsymbol{\theta})\{\rho^*(y_j)-\rho(y_j)\}\big]\frac{h^2}{2}\int t^2K(t)dt+O(h^4) \\
= \; & \{\mathbf{U}(\mathbf{x}_j,y_j,\boldsymbol{\theta})-\mathbf{b}^{**\star}(\mathbf{x}_j,\mathbf{a}^{**\star},\boldsymbol{\zeta},\boldsymbol{\theta})\}\{\rho^*(y_j)-\rho(y_j)\}+o(n_1^{-1/2})
\end{aligned}$$

under Conditions (B1), (B3), (B6), and (B7). On the other hand, using $\mathrm{E}\{\mathbf{b}^{**\star}(\mathbf{X},\mathbf{a}^{**\star},\boldsymbol{\zeta},\boldsymbol{\theta})\mid$

$y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$, $T_2(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} T_2(\boldsymbol{\theta}) &= \mathrm{E}\left[\left\{\frac{R}{\pi}\rho^*(Y) - \frac{1-R}{1-\pi}\right\}\frac{\partial \mathbf{b}^{**}(\mathbf{X}, a^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}\right](\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2}) \\ &= \frac{\partial}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}\mathrm{E}\left[\left\{\frac{R}{\pi}\rho^*(Y) - \frac{1-R}{1-\pi}\right\}\mathrm{E}\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) \mid Y\}\right](\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(n_1^{-1/2}) \\ &= o_p(n_1^{-1/2}). \end{aligned}$$

Hence, (D.25) leads to

$$\begin{aligned} & \mathbf{A}^{**} n^{-1}\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\left\{\mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) - \mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \widehat{\boldsymbol{\zeta}}, \boldsymbol{\theta})\right\} + \frac{1-r_i}{1-\pi}\mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \widehat{\boldsymbol{\zeta}}, \boldsymbol{\theta})\right] \\ = \; & n^{-1}\sum_{i=1}^{n}\boldsymbol{\phi}_{\mathrm{eff}}^{**}(\mathbf{x}_i, r_i, r_i y_i, \boldsymbol{\theta}) \\ & + \mathbf{A}^{**} n_1^{-1}\sum_{i=1}^{n} r_i\{\mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta})\}\{\rho^*(y_i) - \rho(y_i)\} + o_p(n_1^{-1/2}) \\ = \; & \mathrm{E}\left\{\boldsymbol{\phi}_{\mathrm{eff}}^{**}(\mathbf{X}, R, RY, \boldsymbol{\theta})\right\} \\ & + \mathbf{A}^{**}\mathrm{E}_p\left[\{\mathbf{b}^{**}(\mathbf{X}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\}\{\rho^*(Y) - \rho(Y)\}\right] + O_p(n_1^{-1/2}) \\ = \; & \mathrm{E}\left\{\boldsymbol{\phi}_{\mathrm{eff}}^{**}(\mathbf{X}, R, RY, \boldsymbol{\theta})\right\} + O_p(n_1^{-1/2}), \end{aligned} \tag{D.26}$$

since $\mathrm{E}\{\mathbf{b}^{**}(\mathbf{X}, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$. This implies that the estimating equation converges in probability to $\mathrm{E}\{\boldsymbol{\phi}_{\mathrm{eff}}^{**}(\mathbf{X}, R, RY, \boldsymbol{\theta})\}$ uniformly in $\boldsymbol{\theta}$ by Condition (B1). Hence, $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$.

Finally, we derive the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$. By the definition of $\widehat{\boldsymbol{\theta}}$ and (D.26),

$$\begin{aligned} \mathbf{0} &= \sqrt{n_1}\mathbf{A}^{**} n^{-1}\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\left\{\mathbf{U}(\mathbf{x}_i, y_i, \widehat{\boldsymbol{\theta}}) - \mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\theta}})\right\}\right. \\ &\quad \left. + \frac{1-r_i}{1-\pi}\mathbf{b}^{**}(\mathbf{x}_i, \widehat{\mathbf{a}}^{**}, \widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\theta}})\right] \\ &= n^{-1/2}\sum_{i=1}^{n}\left[\sqrt{\pi}\boldsymbol{\phi}_{\mathrm{eff}}^{**}(\mathbf{x}_i, r_i, r_i y_i, \boldsymbol{\theta})\right. \\ &\quad \left. + \frac{r_i}{\sqrt{\pi}}\mathbf{A}^{**}\{\mathbf{b}^{**}(\mathbf{x}_i, \mathbf{a}^{**}, \boldsymbol{\zeta}, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta})\}\{\rho^*(y_i) - \rho(y_i)\}\right] \\ &\quad + \widehat{\mathbf{B}}\sqrt{n_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1), \end{aligned}$$

where

$$\begin{aligned}
\widehat{\mathbf{B}} &= \mathrm{E}\left\{\frac{\partial \boldsymbol{\phi}_{\mathrm{eff}}^{**\star}(\mathbf{X}, R, RY, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\} \\
&\quad + \mathbf{A}^{**\star}\mathrm{E}_p\left[\frac{\partial\{\mathbf{b}^{**\star}(\mathbf{X}, \mathbf{a}^{**\star}, \boldsymbol{\zeta}, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\{\rho^*(Y) - \rho(Y)\}\right] + o_p(1) \\
&= \mathrm{E}\left\{\frac{\partial \boldsymbol{\phi}_{\mathrm{eff}}^{**\star}(\mathbf{X}, R, RY, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\} \\
&\quad + \mathbf{A}^{**\star}\mathrm{E}_p\left[\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}E\{\mathbf{b}^{**\star}(\mathbf{X}, \mathbf{a}^{**\star}, \boldsymbol{\zeta}, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) \mid Y\}\{\rho^*(Y) - \rho(Y)\}\right] + o_p(1) \\
&= \mathrm{E}\left\{\frac{\partial \boldsymbol{\phi}_{\mathrm{eff}}^{**\star}(\mathbf{X}, R, RY, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\} + o_p(1)
\end{aligned}$$

because $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$ and $\mathrm{E}\{\mathbf{b}^{**\star}(\mathbf{X}, \mathbf{a}^{**\star}, \boldsymbol{\zeta}, \boldsymbol{\theta}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$. In addition,

$$\begin{aligned}
&\mathrm{E}\left\{\frac{\partial \boldsymbol{\phi}_{\mathrm{eff}}^{**\star}(\mathbf{X}, R, RY, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\} \\
&= \mathbf{A}^{**\star}\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\mathrm{E}\left[\frac{R}{\pi}\rho^*(Y)\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) - \mathbf{b}^{**\star}(\mathbf{X}, \boldsymbol{\theta})\} + \frac{1-R}{1-\pi}\mathbf{b}^{**\star}(\mathbf{X}, \boldsymbol{\theta})\right] \\
&= \mathbf{A}^{**\star}\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\left(\mathrm{E}_p\left[\rho^*(Y)\mathrm{E}\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) - \mathbf{b}^{**\star}(\mathbf{X}, \boldsymbol{\theta}) \mid Y\}\right] + \mathrm{E}_q\left[\mathrm{E}\left\{\mathbf{b}^{**\star}(\mathbf{X}, \boldsymbol{\theta}) \mid Y\right\}\right]\right) \\
&= \mathbf{A}^{**\star}\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\mathrm{E}_q\left[\mathrm{E}\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) \mid Y\}\right] \\
&= \mathbf{A}^{**\star}\mathbf{A}^{-1}.
\end{aligned}$$

Therefore, $\sqrt{n_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to $N\left\{\mathbf{0}, \mathbf{A}\mathbf{A}^{**\star -1}\boldsymbol{\Sigma}(\mathbf{A}\mathbf{A}^{**\star -1})^{\mathrm{T}}\right\}$ as $n_1 \to \infty$, where $\boldsymbol{\Sigma}$ is given in (D.17). □

D.11.6 Proof of Theorem D.10.1

We define

$$\mathbf{b}^*(\mathbf{x}, \mathbf{a}, \mathrm{E}_p, \boldsymbol{\theta}) \equiv \frac{\mathrm{E}_p\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho^{*2}(Y) + \mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}},$$

and for any function $g(\cdot, \mu)$, define its kth Gateaux derivative with respect to μ at μ_1 in the direction μ_2 as

$$\frac{\partial^k g(\cdot, \mu_1)}{\partial \mu^k}(\mu_2) \equiv \frac{\partial^k g(\cdot, \mu)}{\partial \mu^k}(\mu_2)\Big|_{\mu=\mu_1} \equiv \frac{\partial^k g(\cdot, \mu_1 + h\mu_2)}{\partial h^k}\Big|_{h=0}.$$

Then we have

$$
\begin{aligned}
& \frac{\partial \mathbf{b}^*(\mathbf{x}, \mathbf{a}, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= \; & \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho^{*2}(Y) + \mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}} \\
& -\mathrm{E}_p\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho^{*2}(Y) + \mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}\}\frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}{[\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}]^2} \\
= \; & o_p(n_1^{-1/4}),
\end{aligned}
$$

$$
\begin{aligned}
& \frac{\partial^2 \mathbf{b}^*(\mathbf{x}, \mathbf{a}, \mu, \boldsymbol{\theta})}{\partial {\mathrm{E}_p}^2}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= \; & \frac{-2(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}}{[\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}]^3} \\
& \times \Big[(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho^{*2}(Y) + \mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}\}\mu\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\} \\
& -\mu\{\mathbf{U}(\mathbf{x}, Y, \boldsymbol{\theta})\rho^{*2}(Y) + \mathbf{a}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}\}(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}\}\Big] \\
= \; & o_p(n_1^{-1/2})
\end{aligned}
$$

for any bounded $\mathbf{a}(y)$ and $\mu(\cdot \mid \mathbf{x})$ since $\|(\widehat{\mathrm{E}}_p - \mathrm{E}_p)(\cdot \mid \mathbf{x})\|_\infty = o_p(n_1^{-1/4})$ by the assumption, and these hold uniformly in $\mathbf{x}$ by Condition (B5). Then by the Taylor expansion and mean value theorem, for any bounded $\mathbf{a}(y)$ and some $\alpha \in (0, 1)$,

$$
\begin{aligned}
& \mathbf{b}^*(\mathbf{x}, \mathbf{a}, \widehat{\mathrm{E}}_p, \boldsymbol{\theta}) \\
= \; & \mathbf{b}^*(\mathbf{x}, \mathbf{a}, \mathrm{E}_p, \boldsymbol{\theta}) + \frac{\partial \mathbf{b}^*(\mathbf{x}, \mathbf{a}, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + \frac{1}{2}\frac{\partial^2 \mathbf{b}^*\{\mathbf{x}, \mathbf{a}, \mathrm{E}_p + \alpha(\widehat{\mathrm{E}}_p - \mathrm{E}_p), \boldsymbol{\theta}\}}{\partial {\mathrm{E}_p}^2}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= \; & \mathbf{b}^*(\mathbf{x}, \mathbf{a}, \mathrm{E}_p, \boldsymbol{\theta}) + \frac{\partial \mathbf{b}^*(\mathbf{x}, \mathbf{a}, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n_1^{-1/2}). \qquad \text{(D.27)}
\end{aligned}
$$

Noting that $\mathbf{a}^*(y, \boldsymbol{\theta})$ is bounded under Condition (B4), we further get

$$
\begin{aligned}
& \widehat{\mathcal{L}}^*(\mathbf{a}^*)(y, \boldsymbol{\theta}) - \widehat{\mathbf{v}}^*(y, \boldsymbol{\theta}) \\
= \; & n_1^{-1}\sum_{i=1}^{n} r_i K_h(y - y_i)\{\mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \widehat{\mathrm{E}}_p, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{x}_i, y, \boldsymbol{\theta})\} \\
= \; & n_1^{-1}\sum_{i=1}^{n} r_i K_h(y - y_i)\left\{\mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \frac{\partial \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n_1^{-1/2}) - \mathbf{U}(\mathbf{x}_i, y, \boldsymbol{\theta})\Big\} \\
&= n_1^{-1} \sum_{i=1}^{n} \left\{ \mathcal{L}^*_{i,h}(\mathbf{a}^*)(y, \boldsymbol{\theta}) - \mathbf{v}^*_{i,h}(y, \boldsymbol{\theta}) \right\} + o_p(n_1^{-1/2})
\end{aligned}
$$

uniformly in y by Condition (B5). The last equality above is because $\mathrm{E}\{\mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$ from the definition of $\mathbf{a}^*$, hence

$$
\begin{aligned}
& n_1^{-1} \sum_{i=1}^{n} r_i K_h(y - y_i) \frac{\partial \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= \; & n_1^{-1} \sum_{i=1}^{n} r_i K_h(y - y_i) \frac{\partial \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
& - \frac{\partial [p_Y(y) \mathrm{E}\{\mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) \mid y\}]}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
= \; & n_1^{-1/4} \Bigg[n_1^{-1} \sum_{i=1}^{n} r_i K_h(y - y_i) n_1^{1/4} \frac{\partial \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
& - p_Y(y) \mathrm{E}\left\{ n_1^{1/4} \frac{\partial \mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \mid y \right\} \Bigg] \\
= \; & n_1^{-1/4} O_p\left\{ (n_1 h)^{-1/2} \log n_1 + h^2 \right\} \\
= \; & o_p(n_1^{-1/2})
\end{aligned}
$$

under Conditions (B4)-(B7). In addition, for any bounded function $\mathbf{a}(y, \boldsymbol{\theta})$,

$$
\|(\widehat{\mathcal{L}}^* - \mathcal{L}^*)(\mathbf{a})\|_\infty = O_p\left\{ (n_1 h)^{-1/2} \log n_1 + h^2 \right\} = o_p(n_1^{-1/4})
$$

under Conditions (B1), (B4)-(B7) and the assumption $\|(\widehat{\mathrm{E}}_p - \mathrm{E}_p)(\cdot \mid \mathbf{x})\|_\infty = o_p(n_1^{-1/4})$. Similarly, $\|\widehat{\mathbf{v}}^* - \mathbf{v}^*\|_\infty = o_p(n_1^{-1/4})$. Then using that $\mathcal{L}^{*-1}$ is a bounded linear operator by Lemma D.9.1, $\widehat{\mathbf{a}}^*(y, \boldsymbol{\theta})$ can be expressed as

$$
\begin{aligned}
\widehat{\mathbf{a}}^*(y, \boldsymbol{\theta}) &= \{\mathcal{L}^* + (\widehat{\mathcal{L}}^* - \mathcal{L}^*)\}^{-1}(\widehat{\mathbf{v}}^*)(y, \boldsymbol{\theta}) \\
&= \{\mathcal{L}^{*-1} - \mathcal{L}^{*-1}(\widehat{\mathcal{L}}^* - \mathcal{L}^*)\mathcal{L}^{*-1}\}\{\mathbf{v}^* + (\widehat{\mathbf{v}}^* - \mathbf{v}^*)\}(y, \boldsymbol{\theta}) + o_p(n_1^{-1/2}) \\
&= \mathbf{a}^*(y, \boldsymbol{\theta}) + \mathcal{L}^{*-1}(\widehat{\mathbf{v}}^* - \mathbf{v}^*)(y, \boldsymbol{\theta}) - \{\mathcal{L}^{*-1}(\widehat{\mathcal{L}}^* - \mathcal{L}^*)\}(\mathbf{a}^*)(y, \boldsymbol{\theta}) + o_p(n_1^{-1/2}) \\
&= \mathbf{a}^*(y, \boldsymbol{\theta}) + \mathcal{L}^{*-1}\{\widehat{\mathbf{v}}^* - \widehat{\mathcal{L}}^*(\mathbf{a}^*)\}(y, \boldsymbol{\theta}) + o_p(n_1^{-1/2}) \\
&= \mathbf{a}^*(y, \boldsymbol{\theta}) + n_1^{-1} \sum_{i=1}^{n} \mathbf{g}_{i,h}(y, \boldsymbol{\theta}) + o_p(n_1^{-1/2}) \qquad \text{(D.28)}
\end{aligned}
$$

uniformly in y, where $\mathbf{g}_{i,h}(y,\boldsymbol{\theta}) \equiv \mathcal{L}^{*-1}\{\mathbf{v}^*_{i,h} - \mathcal{L}^*_{i,h}(\mathbf{a}^*)\}(y,\boldsymbol{\theta})$. We also have for any bounded $\mathbf{a}(y,\boldsymbol{\theta})$,

$$\begin{aligned} & \left\| \frac{\mathrm{E}_p\{\mathbf{a}(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \right\|_\infty \\ \leqslant & \ \|\mathbf{a}\|_\infty \left| \frac{\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \right| \\ = & \ O(\|\mathbf{a}\|_\infty), \end{aligned}$$

and

$$\begin{aligned} & \left\| \frac{\partial \mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) - \frac{\partial \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \right\|_\infty \\ = & \left\| \frac{\partial}{\partial \mathrm{E}_p} \left[\frac{\mathrm{E}_p\{(\widehat{\mathbf{a}}^* - \mathbf{a}^*)(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \right] (\widehat{\mathrm{E}}_p - \mathrm{E}_p) \right\|_\infty \\ = & \left\| \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{(\widehat{\mathbf{a}}^* - \mathbf{a}^*)(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \right. \\ & \left. -\mathrm{E}_p\{(\widehat{\mathbf{a}}^* - \mathbf{a}^*)(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\} \frac{(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}}{[\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}]^2} \right\|_\infty \\ \leqslant & \ n_1^{-1/4} \frac{\|(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{n_1^{1/4}(\widehat{\mathbf{a}}^* - \mathbf{a}^*)(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}\|_\infty}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \\ & + \|\widehat{\mathbf{a}}^* - \mathbf{a}^*\|_\infty \mathrm{E}_p\{\rho^*(Y) \mid \mathbf{x}_i\} \frac{|(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}|}{[\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}]^2} \\ = & \ n_1^{-1/4} o_p(n_1^{-1/4}) + o_p(n_1^{-1/4}) o_p(n_1^{-1/4}) \\ = & \ o_p(n_1^{-1/2}) \end{aligned}$$

because $\|(\widehat{\mathrm{E}}_p - \mathrm{E}_p)(\cdot \mid \mathbf{x})\|_\infty = o_p(n_1^{-1/4})$ and $\|\widehat{\mathbf{a}}^{*\star} - \mathbf{a}^{*\star}\|_\infty = O_p\left\{(n_1 h)^{-1/2} \log n_1 + h^2\right\} = o_p(n^{-1/4})$ by (D.22), (D.28), and Condition (B7). Hence using (D.27) and (D.28),

$$\begin{aligned} & \mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \widehat{\mathrm{E}}_p, \boldsymbol{\theta}) \qquad \text{(D.29)} \\ = & \ \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) + \frac{\mathrm{E}_p\{(\widehat{\mathbf{a}}^* - \mathbf{a}^*)(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \\ & + \frac{\partial \mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n_1^{-1/2}) \\ = & \ \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) + n_1^{-1} \sum_{j=1}^{n} \frac{\mathrm{E}_p\{\mathbf{g}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}} \end{aligned}$$

$$+\frac{\partial \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) + o_p(n_1^{-1/2})$$

uniformly in $\mathbf{x}_i$ by Condition (B5).

Now we show the consistency of $\widetilde{\boldsymbol{\theta}}$ by following Theorem 2.1 of Newey & McFadden (1994). We can view the problem of finding the solution for $\mathrm{E}\{\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\} = \mathbf{0}$ as maximizing the objective function $Q_0(\boldsymbol{\theta}) \equiv -\|\mathrm{E}\{\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\}\|_2^2$, then Theorem 2.1 of Newey & McFadden (1994) is directly applicable. It is immediate that $\mathrm{E}\{\mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$ from the definition of $\mathbf{b}^*(\mathbf{x}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})$. Hence,

$$\begin{aligned}
& \mathrm{E}\{\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\} \\
= \;& \mathbf{A}^*\mathrm{E}\left[\frac{R}{\pi}\rho^*(Y)\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) - \mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})\} + \frac{1-R}{1-\pi}\mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})\right] \\
= \;& \mathbf{A}^*\mathrm{E}_p\left[\rho^*(Y)\mathrm{E}\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) - \mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) \mid Y\}\right] + \mathbf{A}^*\mathrm{E}_q\left[\mathrm{E}\left\{\mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) \mid Y\right\}\right] \\
= \;& \mathbf{A}^*\mathrm{E}_q\left[\mathrm{E}\left\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) \mid Y\right\}\right] \\
= \;& \mathbf{0},
\end{aligned}$$

where the last step is by the definition of $\boldsymbol{\theta}$. Also, Condition (B1) implies that $\boldsymbol{\theta}$ is the unique solution for $\mathrm{E}\{\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\} = \mathbf{0}$ in the neighborhood of $\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ which is compact, and $\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{x}, r, ry, \boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta}$. Therefore, it suffices to show that the estimating equation converges in probability to $\mathrm{E}\{\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\}$ uniformly in $\boldsymbol{\theta}$. Using (D.29), the estimating equation can be expressed as

$$\begin{aligned}
& \mathbf{A}^* n^{-1}\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\left\{\mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) - \mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \widehat{\mathrm{E}}_p, \boldsymbol{\theta})\right\} + \frac{1-r_i}{1-\pi}\mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \widehat{\mathrm{E}}_p, \boldsymbol{\theta})\right] \\
= \;& n^{-1}\sum_{i=1}^{n}\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{x}_i, r_i, r_i y_i, \boldsymbol{\theta}) \\
& + \mathbf{A}^* n^{-1}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i) - \frac{1-r_i}{1-\pi}\right\}\left\{\mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) - \mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \widehat{\mathrm{E}}_p, \boldsymbol{\theta})\right\} \\
= \;& n^{-1}\sum_{i=1}^{n}\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{x}_i, r_i, r_i y_i, \boldsymbol{\theta}) - \mathbf{A}^*\{T_1(\boldsymbol{\theta}) + T_2(\boldsymbol{\theta})\} + o_p(n_1^{-1/2}),
\end{aligned} \tag{D.30}$$

where

$$T_1(\boldsymbol{\theta}) \;\equiv\; n^{-1}n_1^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i) - \frac{1-r_i}{1-\pi}\right\}\frac{\mathrm{E}_p\{\mathbf{g}_{j,h}(Y, \boldsymbol{\theta})\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y) + \pi/(1-\pi)\rho^*(Y) \mid \mathbf{x}_i\}},$$

$$T_2(\boldsymbol{\theta}) \equiv n^{-1}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i)-\frac{1-r_i}{1-\pi}\right\}\frac{\partial \mathbf{b}^*(\mathbf{x}_i,\mathbf{a}^*,\mathrm{E}_p,\boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p-\mathrm{E}_p).$$

Using the property of the U-statistic, Condition (B7), and (D.23), we can rewrite $T_1(\boldsymbol{\theta})$ as

$$\begin{aligned}
& T_1(\boldsymbol{\theta}) \\
= \; & n_1^{-1}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i)-\frac{1-r_i}{1-\pi}\right\}\mathrm{E}\left[\frac{\mathrm{E}_p\{\mathbf{G}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid \mathbf{x}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{x}_i\}}\mid \mathbf{x}_i, r_i, r_i y_i\right] \\
& +n_1^{-1}\sum_{j=1}^{n}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\rho^*(Y_i)-\frac{1-R_i}{1-\pi}\right\}\frac{\mathrm{E}_p\{\mathbf{g}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid \mathbf{X}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{X}_i\}}\mid \mathbf{x}_j, r_j, r_j y_j\right] \\
& -n^{1/2}n_1^{-1}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\rho^*(Y_i)-\frac{1-R_i}{1-\pi}\right\}\frac{\mathrm{E}_p\{\mathbf{G}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid \mathbf{X}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{X}_i\}}\right]+O_p(n_1^{-1}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}\mathrm{E}\left[\left\{\frac{R_i}{\pi}\rho^*(Y_i)-\frac{1-R_i}{1-\pi}\right\}\frac{\mathrm{E}_p\{\mathbf{g}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid \mathbf{X}_i\}}{\mathrm{E}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{X}_i\}}\mid \mathbf{x}_j, r_j, r_j y_j\right] \\
& +O_p(n_1^{-1}nh^2+n_1^{-1}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}\int\{\rho^*(y)p_Y(y)-q_Y(y)\}\mathrm{E}\left[\frac{\mathrm{E}_p\{\mathbf{g}_{j,h}(Y,\boldsymbol{\theta})\rho^*(Y)\mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^{*2}(Y)+\pi/(1-\pi)\rho^*(Y)\mid \mathbf{X}\}}\mid y\right]dy+o_p(n_1^{-1/2}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}\int\mathcal{L}^*(\mathbf{g}_{j,h})(y,\boldsymbol{\theta})\left\{\rho^*(y)-\rho(y)\right\}dy+o_p(n_1^{-1/2}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}\int\{\mathbf{v}^*_{j,h}(y,\boldsymbol{\theta})-\mathcal{L}^*_{j,h}(\mathbf{a}^*)(y,\boldsymbol{\theta})\}\left\{\rho^*(y)-\rho(y)\right\}dy+o_p(n_1^{-1/2}) \\
= \; & n_1^{-1}\sum_{j=1}^{n}r_j\int K_h(y-y_j)\{\mathbf{U}(\mathbf{x}_j,y,\boldsymbol{\theta})-\mathbf{b}^*(\mathbf{x}_j,\mathbf{a}^*,\mathrm{E}_p,\boldsymbol{\theta})\}\left\{\rho^*(y)-\rho(y)\right\}dy+o_p(n_1^{-1/2}),
\end{aligned}$$

where

$$\begin{aligned}
& \int K_h(y-y_j)\{\mathbf{U}(\mathbf{x}_j,y,\boldsymbol{\theta})-\mathbf{b}^*(\mathbf{x}_j,\mathbf{a}^*,\mathrm{E}_p,\boldsymbol{\theta})\}\left\{\rho^*(y)-\rho(y)\right\}dy \\
= \; & \{\mathbf{U}(\mathbf{x}_j,y_j,\boldsymbol{\theta})-\mathbf{b}^*(\mathbf{x}_j,\mathbf{a}^*,\mathrm{E}_p,\boldsymbol{\theta})\}\{\rho^*(y_j)-\rho(y_j)\} \\
& +\big[\{\mathbf{U}(\mathbf{x}_j,y_j,\boldsymbol{\theta})-\mathbf{b}^*(\mathbf{x}_j,\mathbf{a}^*,\mathrm{E}_p,\boldsymbol{\theta})\}\{\rho^{*\prime\prime}(y_j)-\rho''(y_j)\}+2\mathbf{U}'_y(\mathbf{x}_j,y_j,\boldsymbol{\theta})\{\rho^{*\prime}(y_j)-\rho'(y_j)\} \\
& +\mathbf{U}''_{yy}(\mathbf{x}_j,y_j,\boldsymbol{\theta})\{\rho^*(y_j)-\rho(y_j)\}\big]\frac{h^2}{2}\int t^2K(t)dt+O(h^4) \\
= \; & \{\mathbf{U}(\mathbf{x}_j,y_j,\boldsymbol{\theta})-\mathbf{b}^*(\mathbf{x}_j,\mathbf{a}^*,\mathrm{E}_p,\boldsymbol{\theta})\}\{\rho^*(y_j)-\rho(y_j)\}+o(n_1^{-1/2})
\end{aligned}$$

under Conditions (B1), (B3), (B6), and (B7). On the other hand, using $\mathrm{E}\{\mathbf{b}^*(\mathbf{X},\mathbf{a}^*,\mathrm{E}_p,\boldsymbol{\theta})\mid$

$y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$, $T_2(\boldsymbol{\theta})$ can be written as

$$\begin{aligned}
T_2(\boldsymbol{\theta}) &= n^{-1}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i) - \frac{1-r_i}{1-\pi}\right\}\frac{\partial \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
&\quad -\frac{\partial}{\partial \mathrm{E}_p}\mathrm{E}\left[\left\{\frac{R}{\pi}\rho^*(Y) - \frac{1-R}{1-\pi}\right\}\mathrm{E}\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta}) \mid y\}\right](\widehat{\mathrm{E}}_p - \mathrm{E}_p) \\
&= n_1^{-1/4}\left(n^{-1}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi}\rho^*(y_i) - \frac{1-r_i}{1-\pi}\right\}n_1^{1/4}\frac{\partial \mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\right. \\
&\quad \left. -\mathrm{E}\left[\left\{\frac{R}{\pi}\rho^*(Y) - \frac{1-R}{1-\pi}\right\}n_1^{1/4}\frac{\partial \mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta})}{\partial \mathrm{E}_p}(\widehat{\mathrm{E}}_p - \mathrm{E}_p)\right]\right) \\
&= n_1^{-1/4}O_p(n^{-1/2}) \\
&= o_p(n_1^{-1/2}).
\end{aligned}$$

Hence, (D.30) leads to

$$\begin{aligned}
&\mathbf{A}^* n^{-1}\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\left\{\mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) - \mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \widehat{\mathrm{E}}_p, \boldsymbol{\theta})\right\} + \frac{1-r_i}{1-\pi}\mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \widehat{\mathrm{E}}_p, \boldsymbol{\theta})\right] \\
= \;& n^{-1}\sum_{i=1}^{n}\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{x}_i, r_i, r_i y_i, \boldsymbol{\theta}) \\
&+\mathbf{A}^* n_1^{-1}\sum_{i=1}^{n} r_i\{\mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta})\}\{\rho^*(y_i) - \rho(y_i)\} + o_p(n_1^{-1/2}) \\
= \;& \mathrm{E}\left\{\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\right\} \\
&+\mathbf{A}^*\mathrm{E}_p\left[\{\mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\}\{\rho^*(Y) - \rho(Y)\}\right] + O_p(n_1^{-1/2}) \\
= \;& \mathrm{E}\left\{\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\right\} + O_p(n_1^{-1/2}),
\end{aligned} \tag{D.31}$$

since $\mathrm{E}\{\mathbf{b}^*(\mathbf{X}, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) \mid y\} = \mathrm{E}\{\mathbf{U}(\mathbf{X}, y, \boldsymbol{\theta}) \mid y\}$. This implies that the estimating equation converges in probability to $\mathrm{E}\{\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X}, R, RY, \boldsymbol{\theta})\}$ uniformly in $\boldsymbol{\theta}$ by Condition (B1). Hence, $\widetilde{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$.

Finally, we derive the asymptotic distribution of $\widetilde{\boldsymbol{\theta}}$. By the definition of $\widetilde{\boldsymbol{\theta}}$ and (D.31),

$$\begin{aligned}
\mathbf{0} &= \sqrt{n_1}\mathbf{A}^* n^{-1}\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho^*(y_i)\left\{\mathbf{U}(\mathbf{x}_i, y_i, \widetilde{\boldsymbol{\theta}}) - \mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \widehat{\mathrm{E}}_p, \widetilde{\boldsymbol{\theta}})\right\} + \frac{1-r_i}{1-\pi}\mathbf{b}^*(\mathbf{x}_i, \widehat{\mathbf{a}}^*, \widehat{\mathrm{E}}_p, \widetilde{\boldsymbol{\theta}})\right] \\
&= n^{-1/2}\sum_{i=1}^{n}\left[\sqrt{\pi}\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{x}_i, r_i, r_i y_i, \boldsymbol{\theta})\right. \\
&\quad \left. +\frac{r_i}{\sqrt{\pi}}\mathbf{A}^*\{\mathbf{b}^*(\mathbf{x}_i, \mathbf{a}^*, \mathrm{E}_p, \boldsymbol{\theta}) - \mathbf{U}(\mathbf{x}_i, y_i, \boldsymbol{\theta})\}\{\rho^*(y_i) - \rho(y_i)\}\right]
\end{aligned}$$

$$+\widehat{\mathbf{B}}\sqrt{n_1}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta})+o_p(1),$$

where

$$\begin{aligned}
\widehat{\mathbf{B}} &= \mathrm{E}\left\{\frac{\partial\boldsymbol{\phi}^*_{\mathrm{eff}}(\mathbf{X},R,RY,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\right\} \\
&\quad +\mathbf{A}^*\mathrm{E}_p\left[\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\mathrm{E}\{\mathbf{b}^*(\mathbf{X},\mathbf{a}^*,\mathrm{E}_p,\boldsymbol{\theta})-\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\mid Y\}\{\rho^*(Y)-\rho(Y)\}\right]+o_p(1) \\
&= \mathbf{A}^*\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\mathrm{E}\left[\frac{R}{\pi}\rho^*(Y)\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})-\mathbf{b}^*(\mathbf{X},\boldsymbol{\theta})\}+\frac{1-R}{1-\pi}\mathbf{b}^*(\mathbf{X},\boldsymbol{\theta})\right]+o_p(1) \\
&= \mathbf{A}^*\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\mathrm{E}_q\left\{\mathbf{U}(\mathbf{X},Y,\boldsymbol{\theta})\right\}+o_p(1) \\
&= \mathbf{A}^*\mathbf{A}^{-1}+o_p(1),
\end{aligned}$$

because $\widetilde{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$ and $\mathrm{E}\{\mathbf{b}^*(\mathbf{X},\mathbf{a}^*,\mathrm{E}_p,\boldsymbol{\theta})\mid y\}=\mathrm{E}\{\mathbf{U}(\mathbf{X},y,\boldsymbol{\theta})\mid y\}$. Therefore, $\sqrt{n_1}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta})$ converges in distribution to $N\left\{\mathbf{0},\mathbf{A}\mathbf{A}^{*-1}\boldsymbol{\Sigma}(\mathbf{A}\mathbf{A}^{*-1})^{\mathrm{T}}\right\}$ as $n_1\to\infty$. □

Bibliography

Agarwal, G. G. & Studden, W. (1980), 'Asymptotic integrated mean square error using least squares and bias minimizing splines', *The Annals of Statistics* pp. 1307–1325.

Agresti, A. (2003), *Categorical data analysis*, John Wiley & Sons.

Alexandari, A., Kundaje, A. & Shrikumar, A. (2020), Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation, *in* 'International Conference on Machine Learning', PMLR, pp. 222–232.

Antonelli, J., Papadogeorgou, G. & Dominici, F. (2022), 'Causal inference in high dimensions: a marriage between bayesian modeling and good frequentist properties', *Biometrics* **78**(1), 100–114.

Ao, W., Calonico, S. & Lee, Y.-Y. (2021), 'Multivalued treatments and decomposition analysis: An application to the WIA program', *Journal of Business & Economic Statistics* **39**(1), 358–371.

Athey, S., Imbens, G. W. & Wager, S. (2018), 'Approximate residual balancing: debiased inference of average treatment effects in high dimensions', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 597–623.

Azizzadenesheli, K., Liu, A., Yang, F. & Anandkumar, A. (2019), 'Regularized learning for domain adaptation under label shifts', *arXiv preprint arXiv:1903.09734* .

Basu, A. & Rathouz, P. J. (2005), 'Estimating marginal and incremental effects on health outcomes using flexible link and variance function models', *Biostatistics* **6**, 93–109.

Belloni, A. & Chernozhukov, V. (2011), 'L1-penalized quantile regression in high-dimensional sparse models', *The Annals of Statistics* **39**(1), 82–130.

Belloni, A. & Chernozhukov, V. (2013), 'Least squares after model selection in high-dimensional sparse models', *Bernoulli* **19**(2), 521–547.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C. & Kato, K. (2018), 'High-dimensional econometrics and regularized GMM', *arXiv:1806.01888* .

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York.

Bickel, P. J., Klaassen, J., Ritov, Y. & Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press Baltimore.

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

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S. & Zuber, S. (2013), 'Data resource profile: the survey of health, ageing and retirement in europe (SHARE)', *International journal of epidemiology* **42**(4), 992–1001.

Bowman, A., Hall, P. & Prvan, T. (1998), 'Bandwidth selection for the smoothing of distribution functions', *Biometrika* **85**(4), 799–808.

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

Cantoni, E. & de Luna, X. (2020), 'Semiparametric inference with missing data: Robustness to outliers and model misspecification', *Econometrics and statistics* **16**, 108–120.

Cantoni, E. & Ronchetti, E. (2001), 'Robust inference for generalized linear models', *Journal of the American Statistical Association* **96**, 1022–1030.

Cattaneo, M. D. (2010), 'Efficient semiparametric estimation of multi-valued treatment effects under ignorability', *Journal of Econometrics* **155**(2), 138–154.

Cattaneo, M. D., Crump, R. K. & Jansson, M. (2010), 'Robust data-driven inference for density-weighted average derivatives', *Journal of the American Statistical Association* **105**(491), 1070–1083.

Cattaneo, M. D., Crump, R. K. & Jansson, M. (2013), 'Generalized jackknife estimators of weighted average derivatives', *Journal of the American Statistical Association* **108**(504), 1243–1256.

Chan, Y. S. & Ng, H. T. (2005), Word sense disambiguation with distribution estimation., *in* 'IJCAI', Vol. 5, Citeseer, pp. 1010–5.

Chaudhuri, P., Doksum, K. & Samarov, A. (1997), 'On average derivative quantile regression', *The Annals of Statistics* **25**(2), 715–744.

Chen, X. & Liao, Z. (2014), 'Sieve M inference on irregular parameters', *Journal of Econometrics* **182**, 70–86.

Chen, X., Liao, Z. & Sun, Y. (2014), 'Sieve inference on possibly misspecified semi-nonparametric time series models', *Journal of Econometrics* **178**, 639–658.

Chen, X. & White, H. (1999), 'Improved rates and asymptotic normality for nonparametric neural network estimators', *IEEE Transactions on Information Theory* **45**(2), 682–691.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning', *The Econometrics Journal* **21**(1).

De Boor, C. (1978), *A Practical Guide to Splines*, Vol. 27, Springer, New York.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

DeVore, R. A. & Lorentz, G. G. (1993), *Constructive Approximation*, Vol. 303, Springer, New York.

d'Haultfoeuille, X. (2011), 'On the completeness condition in nonparametric instrumental problems', *Econometric Theory* **27**(3), 460–471.

Drineas, P., Mahoney, M. W. & Muthukrishnan, S. (2006), Sampling algorithms for L2 regression and applications, *in* 'Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm', SODA '06, Society for Industrial and Applied Mathematics, USA, p. 1127–1136.

Du Plessis, M. C. & Sugiyama, M. (2014), 'Semi-supervised learning of class balance under class-prior change by distribution matching', *Neural Networks* **50**, 110–119.

Dukes, O., Avagyan, V. & Vansteelandt, S. (2020), 'Doubly robust tests of exposure effects under high-dimensional confounding', *Biometrics* **76**(4), 1190–1200.

Fan, J. & Gijbels, I. (1996), *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Vol. 66, CRC Press.

Fan, J., Imai, K., Lee, I., Liu, H., Ning, Y. & Yang, X. (2022), 'Optimal covariate balancing conditions in propensity score estimation', *Journal of Business & Economic Statistics* **41**(1), 97–110.

Farrell, M. H. (2015), 'Robust inference on average treatment effects with possibly more covariates than observations', *Journal of Econometrics* **189**(1), 1–23.

Fernihough, A. (2019), 'Marginal effects for generalized linear models: The mfx package for R'.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A. & Neumann, T. C. (2012), 'Estimating the effects of length of exposure to instruction in a training program: The case of job corps', *Review of Economics and Statistics* **94**(1), 153–171.

Fong, C., Hazlett, C. & Imai, K. (2018), 'Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements', *The Annals of Applied Statistics* **12**(1), 156–177.

Galvao, A. F. & Wang, L. (2015), 'Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment', *Journal of the American Statistical Association* **110**(512), 1528–1542.

Garg, S., Wu, Y., Balakrishnan, S. & Lipton, Z. (2020), 'A unified view of label shift estimation', *Advances in Neural Information Processing Systems* **33**, 3290–3300.

GBD 2015 Obesity Collaborators (2017), 'Health effects of overweight and obesity in 195 countries over 25 years', *New England journal of medicine* **377**(1), 13–27.

Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., Foote Jr, J., Moseley, E. T., Grant, D. W., Tyler, P. D. et al. (2018), 'Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives', *PloS one* **13**(2), e0192360.

Genbäck, M., Ng, N., Stanghellini, E. & de Luna, X. (2018), 'Predictors of decline in self-reported health: addressing non-ignorable dropout in longitudinal studies of aging', *European Journal of Ageing* **15**, 211–220.

Gerfin, M. (1996), 'Parametric and semi-parametric estimation of the binary response model of labour market participation', *Journal of Applied Econometrics* **11**, 321–339.

Greene, W. H. (2000), *Econometric analysis*, 4th edn, Prentice Hall, New Jersey.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K. & Schölkopf, B. (2009), 'Covariate shift by kernel mean matching', *Dataset Shift in Machine Learning* **3**(4), 5.

Guo, J., Gong, M., Liu, T., Zhang, K. & Tao, D. (2020), LTF: A label transformation framework for correcting label shift, *in* 'International Conference on Machine Learning', PMLR, pp. 3843–3853.

Hahn, J. (1998), 'On the role of the propensity score in efficient semiparametric estimation of average treatment effects', *Econometrica* pp. 315–331.

Hansen, P. C. (1992), 'Numerical tools for analysis and solution of fredholm integral equations of the first kind', *Inverse problems* **8**(6), 849.

Härdle, W. & Stoker, T. M. (1989), 'Investigating smooth multiple regression by the method of average derivatives', *Journal of the American Statistical Association* **84**(408), 986–995.

Hart, J. D. & Yi, S. (1998), 'One-sided cross-validation', *Journal of the American Statistical Association* **93**, 620–631.

Heinze-Deml, C., Peters, J. & Meinshausen, N. (2018), 'Invariant causal prediction for nonlinear models', *Journal of Causal Inference* **6**(2).

Hernán, M. A. & Taubman, S. L. (2008), 'Does obesity shorten life? the importance of well-defined interventions to answer causal questions', *International journal of obesity* **32**(3), S8–S14.

Hirano, K. & Imbens, G. W. (2004), 'The propensity score with continuous treatments', *Applied Bayesian modeling and causal inference from incomplete-data perspectives* **226164**, 73–84.

Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z. & Wang, K. (2020), 'Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost', *Journal of translational medicine* **18**(1), 1–14.

Hripcsak, G. & Albers, D. J. (2013), 'Next-generation phenotyping of electronic health records', *Journal of the American Medical Informatics Association* **20**(1), 117–121.

Hristache, M., Juditsky, A. & Spokoiny, V. (2001), 'Direct estimation of the index coefficient in a single-index model', *Annals of Statistics* pp. 595–623.

Hu, Y. & Shiu, J.-L. (2018), 'Nonparametric identification using instrumental variables: sufficient conditions for completeness', *Econometric Theory* **34**(3), 659–693.

Huang, A. & Rathouz, P. J. (2012), 'Proportional likelihood ratio models for mean regression', *Biometrika* **99**, 223–229.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B. & Smola, A. (2006), 'Correcting sample selection bias by unlabeled data', *Advances in neural information processing systems* **19**.

Huber, M., Hsu, Y.-C., Lee, Y.-Y. & Lettry, L. (2020), 'Direct and indirect effects of continuous treatments based on generalized propensity score weighting', *Journal of Applied Econometrics* **35**(7), 814–840.

Idler, E. L. & Benyamini, Y. (1997), 'Self-rated health and mortality: a review of twenty-seven community studies', *Journal of health and social behavior* pp. 21–37.

Imai, K. & Ratkovic, M. (2014), 'Covariate balancing propensity score', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 243–263.

Imbens, G. W. (2000), 'The role of the propensity score in estimating dose-response functions', *Biometrika* **87**(3), 706–710.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica: Journal of the Econometric Society* **62**(2), 467–475.

Iyer, A., Nath, S. & Sarawagi, S. (2014), Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection, *in* 'International Conference on Machine Learning', PMLR, pp. 530–538.

Jiang, F., Ma, Y. & Carroll, R. J. (2018), 'A spline-assisted semiparametric approach to non-parametric measurement error models', *arXiv preprint arXiv:1804.00793* .

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L. & Mark, R. G. (2016), 'MIMIC-III, a freely accessible critical care database', *Scientific data* **3**(1), 1–9.

Kang, J. D. & Schafer, J. L. (2007), 'Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data', *Statistical science* **22**, 523–539.

Kennedy, E. H., Ma, Z., McHugh, M. D. & Small, D. S. (2017), 'Non-parametric methods for doubly robust estimation of continuous treatment effects', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1229–1245.

Kleiber, C. & Zeileis, A. (2008), *Applied econometrics with R*, Springer, New York.

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press.

Kooperberg, C. & Stone, C. J. (1991), 'A study of logspline density estimation', *Computational Statistics & Data Analysis* **12**(3), 327–347.

Kouw, W. M. & Loog, M. (2019), 'A review of domain adaptation without target labels', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(3), 766–785.

Kpotufe, S. & Martinet, G. (2021), 'Marginal singularity and the benefits of labels in covariate-shift', *The Annals of Statistics* **49**(6), 3299–3323.

Künsch, H., Stefanski, L. & Carroll, R. J. (1989), 'Conditionally unbiased bounded-infuence estimation in general regression models with applications to generalized linear models', *Journal of the American Statistical Association* **84**, 460–466.

Laan, M. J. & Robins, J. M. (2003), *Unified methods for censored longitudinal data and causality*, Springer.

Landweber, L. (1951), 'An iteration formula for fredholm integral equations of the first kind', *American Journal of Mathematics* **73**(3), 615–624.

Lee, Y.-Y. (2018), 'Efficient propensity score regression estimators of multivalued treatment effects for the treated', *Journal of Econometrics* **204**(2), 207–222.

Lin, L., Liu, L., Cui, X. & Wang, K. (2021), 'A generalized semiparametric regression and its efficient estimation', *Scandinavian Journal of Statistics* **48**, 1–24.

Lipton, Z., Wang, Y.-X. & Smola, A. (2018), Detecting and correcting for label shift with black box predictors, *in* 'International Conference on Machine Learning', PMLR, pp. 3122–3130.

Luo, X. & Tsai, W. Y. (2012), 'A proportional likelihood ratio model', *Biometrika* **99**, 211–222.

Maity, S., Sun, Y. & Banerjee, M. (2022), 'Minimax optimal approaches to the label shift problem in non-parametric settings', *Journal of Machine Learning Research* **23**(346), 1–45.

Manning, W. G., Basu, A. & Mullahy, J. (2005), 'Generalized modeling approaches to risk adjustment of skewed outcomes data', *Journal of Health Economics* **24**, 465–488.

Martinez, J., Hossain, R., Romero, J. & Little, J. J. (2017), A simple yet effective baseline for 3D human pose estimation, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2640–2649.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman & Hall, London.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. & Herrera, F. (2012), 'A unifying view on dataset shift in classification', *Pattern Recognition* **45**(1), 521–530.

Newey, W. K. & McFadden, D. (1994), 'Large sample estimation and hypothesis testing', *Handbook of econometrics* **4**, 2111–2245.

Newey, W. K. & Powell, J. L. (2003), 'Instrumental variable estimation of nonparametric models', *Econometrica* **71**(5), 1565–1578.

Newey, W. K. & Stoker, T. M. (1993), 'Efficiency of weighted average derivative estimators and index models', *Econometrica: Journal of the Econometric Society* pp. 1199–1223.

Ng, M., Liu, P., Thomson, B. & Murray, C. J. (2016), 'A novel method for estimating distributions of body mass index', *Population health metrics* **14**(1), 1–7.

Nguyen, T. D., Christoffel, M. & Sugiyama, M. (2016), Continuous target shift adaptation in supervised learning, *in* 'Asian Conference on Machine Learning', PMLR, pp. 285–300.

Ning, Y., Zhao, T., Liu, H. et al. (2017), 'A likelihood ratio framework for high-dimensional semiparametric regression', *Annals of Statistics* **45**, 2299–2327.

Norton, E. C., Dowd, B. E. & Maciejewski, M. L. (2019), 'Marginal effects—quantifying the effect of changes in risk factors in logistic regression models', *Journal of the American Medical Association* **321**, 1304–1305.

Parzen, E. (2004), 'Quantile probability and statistical data modeling', *Statistical Science* **19**, 652–662.

Qin, J., Zhang, B. & Leung, D. H. (2017), 'Efficient augmented inverse probability weighted estimation in missing data problems', *Journal of Business & Economic Statistics* **35**(1), 86–97.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. (2008), *Dataset Shift in Machine Learning*, MIT Press.

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

Robins, J. M. & Rotnitzky, A. (1995), 'Semiparametric efficiency in multivariate regression models with missing data', *Journal of the American Statistical Association* **90**(429), 122–129.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American statistical Association* **89**(427), 846–866.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Rubin, D. B. & Thomas, N. (2000), 'Combining propensity score matching with additional adjustments for prognostic covariates', *Journal of the American Statistical Association* **95**(450), 573–585.

Saerens, M., Latinne, P. & Decaestecker, C. (2002), 'Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure', *Neural Computation* **14**(1), 21–41.

Sant'Anna, P. H., Song, X. & Xu, Q. (2022), 'Covariate distribution balance via propensity scores', *Journal of Applied Econometrics* **37**(6), 1093–1120.

Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999), 'Adjusting for nonignorable drop-out using semiparametric nonresponse models', *Journal of the American Statistical Association* **94**(448), 1096–1120.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K. & Mooij, J. (2012), On causal and anticausal learning, *in* '29th International Conference on Machine Learning (ICML 2012)', Omnipress, pp. 1255–1262.

Shen, X. & Wong, W. H. (1994), 'Convergence rate of sieve estimates', *Annals of Statistics* **22**, 580–615.

Shimodaira, H. (2000), 'Improving predictive inference under covariate shift by weighting the log-likelihood function', *Journal of Statistical Planning and Inference* **90**(2), 227–244.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M. et al. (2016), 'The third international consensus definitions for sepsis and septic shock (sepsis-3)', *JAMA* **315**(8), 801–810.

Stefanski, L., Carroll, R. J. & Ruppert, D. (1986), 'Optimally bounded score functions for generalized linear models with applications to logistic regression', *Biometrika* **73**, 413–424.

Stone, C. J. (1982), 'Optimal global rates of convergence for nonparametric regression', *The annals of statistics* pp. 1040–1053.

Stone, C. J. (1990), 'Large-sample inference for log-spline models', *The Annals of Statistics* **18**(2), 717–741.

Storkey, A. (2009), 'When training and test sets are different: characterizing learning transfer', *Dataset shift in machine learning* **30**, 3–28.

Sugiyama, M. & Kawanabe, M. (2012), *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*, MIT press.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P. & Kawanabe, M. (2008), 'Direct importance estimation for covariate shift adaptation', *Annals of the Institute of Statistical Mathematics* **60**(4), 699–746.

Tan, W. K. & Heagerty, P. J. (2020), 'Predictive case control designs for modification learning', *arXiv preprint arXiv:2011.14529* .

Tasche, D. (2017), 'Fisher consistency for prior probability shift', *The Journal of Machine Learning Research* **18**(1), 3338–3369.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, Springer, New York.

Waernbaum, I. (2010), 'Propensity score model specification for estimation of average treatment effects', *Journal of Statistical Planning and Inference* **140**(7), 1948–1956.

Wager, S. & Walther, G. (2015), 'Adaptive concentration of regression trees, with application to random forests', *arXiv preprint arXiv:1503.06388* .

Wang, H. & Ma, Y. (2021), 'Optimal subsampling for quantile regression in big data', *Biometrika* **108**(1), 99–112.

Wang, H., Zhu, R. & Ma, P. (2018), 'Optimal subsampling for large sample logistic regression', *Journal of the American Statistical Association* **113**(522), 829–844.

Wang, Y. & Zubizarreta, J. R. (2020), 'Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations', *Biometrika* **107**(1), 93–105.

Wong, R. K. & Chan, K. C. G. (2018), 'Kernel-based covariate functional balancing for observational studies', *Biometrika* **105**(1), 199–213.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E. & Kadziola, Z. (2016), 'Propensity score matching and subclassification in observational studies with multi-level treatments', *Biometrics* **72**(4), 1055–1065.

Yang, X., Sun, H., Sun, X., Yan, M., Guo, Z. & Fu, K. (2018), 'Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network', *IEEE Access* **6**, 50839–50849.

Yin, Z., Tong, J., Chen, Y., Hubbard, R. A. & Tang, C. Y. (2022), 'A cost-effective chart review sampling design to account for phenotyping error in electronic health records (EHR) data', *Journal of the American Medical Informatics Association* **29**(1), 52–61.

Zhang, K., Schölkopf, B., Muandet, K. & Wang, Z. (2013), Domain adaptation under target and conditional shift, *in* 'International Conference on Machine Learning', PMLR, pp. 819–827.

Zhang, L., Ding, X., Ma, Y., Muthu, N., Ajmal, I., Moore, J., Herman, D. & Chen, J. (2020), 'A maximum likelihood approach to electronic health record phenotyping using positive and unlabeled patients', *Journal of the American Medical Informatics Association* **27**, 119–126.

Zhou, S., Shen, X. & Wolfe, D. (1998), 'Local asymptotics for regression splines and confidence regions', *Annals of Statistics* **26**, 1760–1782.

Zhou, T., Tong, G., Li, F. & Thomas, L. E. (2020), 'PSweight: An R package for propensity score weighting analysis', *arXiv preprint arXiv:2010.08893* .

Zubizarreta, J. R. (2015), 'Stable weights that balance covariates for estimation with incomplete outcome data', *Journal of the American Statistical Association* **110**(511), 910–922.

Vita

Seong-ho Lee

Education

- Ph.D. in Statistics, The Pennsylvania State University, PA, USA, 2023.
  Dissertation: Contributions to semiparametric inference and its applications.
  Advisor: Dr. Yanyuan Ma
- M.A. in Applied Statistics, Yonsei University, Seoul, South Korea, 2018.
  Thesis: HDLSS outlier detection using distance vector and dual random rotation.
  Advisor: Dr. Yongho Jeon
- B.S. in Mathematics and B.A. in Applied Statistics, Yonsei University, Seoul, South Korea, 2016.

Publications and Manuscripts

- Lee, S., Richardson, B. D., Ma, Y., & Garcia, T. P. (2023), 'Doubly robust estimation under a randomly censored covariate', in preparation.
- Lee, S., Ma, Y., & Zhao, J. (2023), 'Doubly flexible estimation under label shift', under review.
- Lee, S., Ma, Y., & Ronchetti, E. (2023), 'Semiparametric approach to estimation of marginal mean effects and marginal quantile effects', *Journal of Econometrics*.
- Lee, S., Ma, Y., Wei, Y., & Chen, J. (2023), 'Optimal sampling for positive only electronic health record data', *Biometrics*.
- Lee, S., Ma, Y., & de Luna, X. (2022), 'Covariate balancing for causal inference on categorical and continuous treatments', *Econometrics and Statistics*.

Honors and Awards

- Boyd Harshbarger Travel Award, *Southern Regional Council on Statistics*, USA, Oct 2022.
- Eberly College of Science Innovative Teaching Award in Statistics, *The Pennsylvania State University*, USA, Fall 2019.
- Student Paper Presentation Award, *Korean Statistical Society*, South Korea, Nov 2017.
- Scholarship for Academic Excellence, *Yonsei University*, South Korea, Fall 2009, Fall 2013, Fall 2014, Spring 2015.
- National Science and Technology Scholarship, *Korea Student Aid Foundation*, South Korea, Spring 2009.