Ai, Chunrong; Linton, Oliver Bruce; Motegi, Kaiji; Zhang, Zheng

Working Paper

A unified framework for efficient estimation of general treatment models

cemmap working paper, No. CWP64/19

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Ai, Chunrong; Linton, Oliver Bruce; Motegi, Kaiji; Zhang, Zheng (2019) : A unified framework for efficient estimation of general treatment models, cemmap working paper, No. CWP64/19, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2019.6419

This Version is available at:
https://hdl.handle.net/10419/211157

cemmap
centre for microdata methods and practice

A Unified Framework for Efficient Estimation of General Treatment Models

Chunrong Ai
Oliver Linton
Kaiji Motegi
Zheng Zhang

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP64/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

A Unified Framework for Efficient Estimation of General Treatment Models

Chunrong Ai [*⋆], Oliver Linton [†⋆], Kaiji Motegi [‡†], and Zheng Zhang [§‡]

[⋆]*Department of Economics, University of Florida*
[⋆]*Faculty of Economics, University of Cambridge*
[†]*Graduate School of Economics, Kobe University*
[‡]*Institute of Statistics & Big Data, Renmin University of China*

November 23, 2019

Abstract

This paper presents a weighted optimization framework that unifies the binary, multi-valued, continuous, as well as mixture of discrete and continuous treatment, under unconfounded treatment assignment. With a general loss function, the framework includes the average, quantile and asymmetric least squares causal effect of treatment as special cases. For this general framework, we first derive the semiparametric efficiency bound for the causal effect of treatment, extending the existing bound results to a wider class of models. We then propose a generalized optimization estimator for the causal effect with weights estimated by solving an expanding set of equations. Under some sufficient conditions, we establish the consistency and asymptotic normality of the proposed estimator of the causal effect and show that the estimator attains the semiparametric efficiency bound, thereby extending the existing literature on efficient estimation of causal effect to a wider class of applications. Finally, we discuss estimation of some causal effect functionals such as the treatment effect curve and the average outcome. To evaluate the finite sample performance of the proposed procedure, we conduct a small-scale simulation study and find that the proposed estimation has practical value. To illustrate the applicability of the procedure, we revisit the literature on campaign advertising and campaign contributions. Unlike the existing procedures, which produce mixed results, we find no evidence of campaign advertising on campaign contribution.

[*]E-mail: tsinghua@ufl.edu
[†]E-mail: obl20@cam.ac.uk
[‡]E-mail: motegi@econ.kobe-u.ac.jp
[§]E-mail: zhengzhang@ruc.edu.cn

Keywords: Causal effect; Entropy maximization; Treatment effect; Semiparametric efficiency; Sieve method; Stabilized Weights.

1 Introduction

Modeling and estimating the causal effect of treatment has received considerable attention from both the econometrics and statistics communities (see, e.g., Hirano, Imbens, and Ridder, 2003, Imbens, 2004, Abadie, 2005, Heckman and Vytlacil, 2005, Angrist and Pischke, 2008, Imbens and Wooldridge, 2009, Fan and Park, 2010, Chernozhukov, Fernández-Val, and Melly, 2013, Rothe, 2017, Athey, Imbens, and Wager, 2018, Słoczyński and Wooldridge, 2018, Wager and Athey, 2018). Most existing studies focus on the binary treatment where an individual either receives the treatment or does not, ignoring the treatment intensity. In many applications, however, the treatment intensity is a part of the treatment, and its causal effect is also of great interest to decision makers. For example, in evaluating how financial incentives affect health care providers, the causal effect may depend on not only the introduction of incentive but also the level of incentive. Similarly, in studying how taxes affect addictive substance usages, the causal effect may depend on the imposition of tax as well as on the actual tax rate. In finance, there are many plausible examples of interest. For example, in evaluating the effect of corporate bond purchase schemes on market quality, the causal effect may depend not just on whether the bond is selected into the scheme but on how much of it is purchased (see Boneva, Elliott, Kaminska, Linton, McLaren, and Morley, 2018). In recognition of the importance of the treatment intensity, the binary treatment literature has been extended to the multi-valued treatment (e.g., Imbens, 2000, Cattaneo, 2010) and continuous treatment (e.g., Hirano and Imbens, 2004, Imai and van Dyk, 2004, Florens, Heckman, Meghir, and Vytlacil, 2008, Fong, Hazlett, and Imai, 2018, Yiu and Su, 2018). The parameter of primary interest in this literature is the average causal effect of treatment, defined as the difference in response to two levels of treatment by the same individual, averaged over a set of individuals. The identification and estimation difficulty is that each individual only receives one level of treatment. To overcome this difficulty, researchers impose the *unconfounded treatment assignment* condition, which allows them to find statistical matches for each observed individual from all other treatment levels.

The main objective of this paper is to present a weighted optimization estimation framework that unifies the binary, multi-valued, continuous, as well as the mixture of discrete and continuous treatments, and allows for a general loss function (causal effect parameter) under the unconfounded treatment assignment condition. The weights are called the

stabilized weights by Robins, Hernán, and Brumback (2000) and are defined as the ratio of the marginal probability distribution of the treatment status over the conditional probability distribution of the treatment status given covariates. We first compute the semiparametric efficiency bound, Bickel, Klaassen, Ritov, and Wellner (1993), of the causal effect of treatment, extending the results of Hahn (1998), Firpo (2007), and Cattaneo (2010) from the binary treatment to a variety of treatments and to a general loss function. Our bound reveals that the weighted optimization with known stabilized weights does not produce efficient estimation since it fails to account for the information restricting the stabilized weights. This observation was made by Hirano, Imbens, and Ridder (2003) in the binary treatment case; here we show that their observation holds true for a much wider class of treatment models. We exploit the information that the stabilized weights satisfy certain moment conditions (an expanding number thereof) by estimating the stabilized weights from those equations by a novel entropy maximization method; we then estimate the causal effect by the generalized optimization method with the true stabilized weights replaced by the estimated weights. Under some sufficient conditions, we show that our proposed estimator is consistent and asymptotically normally distributed and, more importantly, it attains our semiparametric efficiency bound. We also propose consistent standard errors based on the same sieve methodology. We propose a tuning parameter selection methodology to guide the practical implementation. We also discuss estimation of the effect curve and establish its pointwise asymptotic normality and uniform consistency. We next present some simulation evidence that our estimation and inference methodology works well in finite samples and is robust to misspecification, whereas the Fong, Hazlett, and Imai (2018) is fragile. We apply our methodology to the study of the effect of political advertisements on campaign contributions using data considered by Urban and Niebler (2014) and Fong, Hazlett, and Imai (2018). We find that the evidence obtained by the Fong, Hazlett, and Imai (2018) method depends on the specification, and for some choices yields significant parameter estimates, whereas our method unambiguously finds effects that are economically small and statistically insignificant.

Literature Review. In the binary treatment case with *unconfounded treatment assignment*, the average causal effect is estimated by the difference of the weighted average responses with the propensity scores as weights (see, e.g., Rosenbaum and Rubin, 1983, Hirano, Imbens, and Ridder, 2003, Busso, DiNardo, and McCrary, 2014). Other popular methods include regression adjustment (Rubin, 1977, Angrist and Pischke, 2008), matching (Imbens, 2004, Abadie and Imbens, 2006, 2011, 2012, 2016) , imputation (Heckman, Ichimura, and Todd, 1998, Cattaneo and Farrell, 2011), and hybrid method (Farrell, 2015, Słoczyński and Wooldridge, 2018, Chernozhukov, Escanciano, Ichimura, Newey, and

Robins, 2018). The efficiency bound of the average causal effect in this model is derived by Robins, Rotnitzky, and Zhao (1994) and Hahn (1998), and efficient estimation is proposed by Robins, Rotnitzky, and Zhao (1994), Hahn (1998), Hirano, Imbens, and Ridder (2003), Bang and Robins (2005), Qin and Zhang (2007), Cao, Tsiatis, and Davidian (2009), Tan (2010), Vansteelandt, Bekaert, and Claeskens (2010), Graham, Pinto, and Egel (2012), and Chan, Yam, and Zhang (2016). Of particular interest in this literature is the study by Hirano, Imbens, and Ridder (2003) which shows that the weighted average difference estimator attains the semiparametric efficiency bound if the weights are estimated by the empirical likelihood estimation. In the multi-valued treatment case, Imbens (2000) generalizes the propensity score, and Cattaneo (2010) derives the efficiency bound and proposes an estimator that attains the efficiency bound. In the continuous treatment case, Hirano and Imbens (2004) and Imai and van Dyk (2004) parameterize the generalized propensity score function and propose a consistent estimator of the average causal effect. Their estimators are not efficient and could be biased if the generalized propensity score function is misspecified. Florens, Heckman, Meghir, and Vytlacil (2008) use a control function approach to identify the average causal effect in the continuous treatment and propose a consistent estimation. It is unclear if their estimation is efficient. Galvao and Wang (2015) estimate the continuous treatment effects through stabilized weighting. They do not study how to construct the stabilized weights such that their estimation is efficient. Kennedy, Ma, McHugh, and Small (2017) propose a nonparametric kernel estimator for the treatment effects curve, again the efficient estimation is still unclear. Fong, Hazlett, and Imai (2018) propose an estimator of the average causal effect of continuous treatment but do not establish consistency of their estimation. In fact, their simulation results indicate their estimation could be seriously biased. Yiu and Su (2018) study the average causal effect of both discrete and continuous treatment by parameterizing the propensity score. Their estimator is generally biased if their parameterization is incorrect.

In addition to the average causal effect of treatment (ATE), it is also important to investigate the distributional impact of treatment. For instance, a decision maker may be interested in the causal effect of a treatment on the outcome dispersion or on the lower tail of the outcome distribution. Doksum (1974) and Lehmann (1975) introduce the quantile causal effect of treatment (QTE). Firpo (2007) computes the efficiency bound and proposes an efficient estimation of QTE for the binary treatment. For additional studies on QTE, we refer to Abadie, Angrist, and Imbens (1998), Chernozhukov and Hansen (2005), Angrist and Pischke (2008), Frölich and Melly (2013), and Donald and Hsu (2014).

To the best of our knowledge, we are unaware of any previous work that computes the efficiency bound and proposes efficient estimation of the causal effect in the continuous or

mixture of discrete and continuous treatment under a general loss function that permits ATE and QTE.

The paper is organized as follows. Section 2 sets up the basic framework, Section 3 computes the semiparametric efficiency bound of the causal effect of treatment, Section 4 presents the generalized optimization estimator, Section 5 establishes the large sample properties of the proposed estimator, while Section 6 presents a consistent covariance matrix. In Section 7 we propose two data-driven approaches for selecting tuning parameters. In Section 8 we discuss some extensions. Section 9 reports on a simulation study, while Section 10 presents an empirical application, followed by some concluding remarks in Section 11. All technical proofs and extra simulation results are relegated to the supplemental material Ai, Linton, Motegi, and Zhang (2019).

2 Basic framework and notation

Let T denote the observed treatment status variable with support $\mathcal{T} \subset \mathbb{R}$, where $\mathcal{T}$ is either a discrete set, a continuum or a mixture of discrete and continuum subsets, and T has a marginal probability distribution function $F_T(t)$. Let $Y^*(t)$ denote the potential response when treatment $T = t$ is assigned. Let $L(\cdot)$ denote a known convex loss function whose derivative, denoted by $L'(\cdot)$, exists almost everywhere. For the leading part of the paper, we shall maintain that there exists a parametric causal effect function $g(t;\beta)$ with the unknown value $\boldsymbol{\beta}^* \in \mathbb{R}^p$ (with $p \in \mathbb{N}$) uniquely solving the minimization problem below, i.e.,

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \int_{\mathcal{T}} \mathbb{E}\left[L\left(Y^*(t) - g(t;\boldsymbol{\beta})\right)\right] dF_T(t). \tag{2.1}$$

The parameterization of the causal effect is restrictive. Some extensions to the unspecified causal effect function shall be discussed later in the paper (see Section 8).

The generality of model (2.1) permits many important already considered models. For example, it includes: the average causal effect of binary treatment studied in Hahn (1998) and Hirano, Imbens, and Ridder (2003) (i.e., $\mathcal{T} = \{0,1\}$, $L(v) = v^2$ and $g(t;\ \boldsymbol{\beta}) = \beta_0 + \beta_1 t$), the quantile causal effect of binary treatment studied in Firpo (2007) (i.e., $\mathcal{T} = \{0,1\}$, $L(v) = v(\tau - I(v \leq 0))$ is an almost everywhere differentiable function with $\tau \in (0,1)$ and $g(t;\boldsymbol{\beta}) = t\beta_1 + (1-t)\beta_0$), the average causal effect of multi-valued treatment studied in Cattaneo (2010) (i.e., $\mathcal{T} = \{0,1,\ldots,J\}$ for some $J \in \mathbb{N}$, $L(v) = v^2$ and $g(t;\boldsymbol{\beta}) = \sum_{j=0}^{J} \beta_j I(t=j)$), and the average causal effect of continuous treatment studied in Hirano and Imbens (2004) (i.e., $L(v) = v^2$ and $\mathbb{E}[Y^*(t)] = g(t;\boldsymbol{\beta})$ is a parametric model indexed by $\boldsymbol{\beta}$ for the potential outcome means, which is also termed by *marginal*

structural model in Robins, Hernán, and Brumback (2000). Examples include the linear marginal structure model $\mathbb{E}[Y^*(t)] = \beta_0 + \beta_1 \cdot t$, and the nonlinear marginal structure model $\mathbb{E}[Y^*(t)] = \beta_0 \cdot t + 1/(t+\beta_1)^2$ studied in Hirano and Imbens (2004)). It also includes the quantile causal effect of multi-valued (i.e., $L(v) = v(\tau - I(v \leq 0))$ with $\tau \in (0,1)$ and $g(t;\boldsymbol{\beta}) = \sum_{j=0}^{J} \beta_j I(t=j)$) and continuous treatment (i.e., $L(v) = v(\tau - I(v \leq 0))$ and $\inf\{q : \mathbb{P}(Y^*(t) \geq q) \leq \tau\} = g(t;\boldsymbol{\beta})$ is a parametric model indexed by $\boldsymbol{\beta}$ for the potential outcome quantiles. Examples include the linear model $\inf\{q : \mathbb{P}(Y^*(t) \geq q) \leq \tau\} = \beta_0 + \beta_1 \cdot t$ and the Box-Cox transformation model $\inf\{q : \mathbb{P}(Y^*(t) \geq q) \leq \tau\} = h_\lambda(\beta_0 + \beta_1 \cdot t)$ studied in Buchinsky (1995), where $h_\lambda(z) = (\lambda z + 1)^{-1/\lambda}$). The latter has so far not been covered by the existing literature. Moreover, with $L(v) = v^2 |\tau - I(v \leq 0)|$, it covers asymmetric least squares estimation of the causal effect of (binary, multi-valued, continuous, mixture of discrete and continuous) treatment. The asymmetric least squares regression received attention from some noted econometricians (see Newey and Powell, 1987) but zero attention in the causal effect literature. Our framework can also accommodate non-scalar treatment by introducing a dummy variable. For example, when studying the treatment effect of gender on salary, we can consider a dummy variable $T \in \{0,1\}$ to describe gender, where $T=1$ denotes male while $T=0$ denotes female.

The problem with (2.1) is that the potential outcome $Y^*(t)$ is not observed for all t. Let $Y := Y^*(T)$ denote the observed response. One may attempt to solve the following:

$$\min_{\boldsymbol{\beta}} \mathbb{E}[L(Y - g(T;\boldsymbol{\beta}))].$$

However, if there exists a selection into treatment, the true value $\boldsymbol{\beta}_0$ does not solve the above minimization problem. Indeed, in this case, the observed response and treatment assignment data alone cannot identify $\boldsymbol{\beta}^*$. To address this identification issue, most studies in the literature impose a selection on observable condition (e.g., Hirano, Imbens, and Ridder, 2003, Imai and van Dyk, 2004, Fong, Hazlett, and Imai, 2018). Specifically, let $\boldsymbol{X}$ denote a vector of covariates. The following condition shall be maintained throughout the paper.

Assumption 1 (*Unconfounded Treatment Assignment*)**.** *For all $t \in \mathcal{T}$, given $\boldsymbol{X}$, T is independent of $Y^*(t)$, i.e., $Y^*(t) \perp T|\boldsymbol{X}$, for all $t \in \mathcal{T}$.*

Let $F_{T|X}$ denote the conditional probability distribution of T given the observed covariates $\boldsymbol{X}$ and let $dF_{T|X}$ denote the probability measure. In the literature, $dF_{T|X}$ is called the *generalized propensity score* (Hirano and Imbens, 2004, Imai and van Dyk, 2004). Suppose that $dF_{T|X}(T|\boldsymbol{X})$ is positive everywhere and let

$$\pi_0(T, \boldsymbol{X}) := \frac{dF_T(T)}{dF_{T|X}(T|\ \boldsymbol{X})}.$$

The function $\pi_0(T, \boldsymbol{X})$ is called the *stabilized weight* in Robins, Hernán, and Brumback (2000). Under Assumption 1, we obtain

$$\mathbb{E}[\pi_0(T, \boldsymbol{X})L(Y - g(T; \boldsymbol{\beta}))] = \int \mathbb{E}\left[L(Y^*(t) - g(t; \boldsymbol{\beta}))\right] dF_T(t) \tag{2.2}$$

(see Appendix A), and hence the true value $\boldsymbol{\beta}^*$ solves the weighted optimization problem:

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}[\pi_0(T, \boldsymbol{X})L(Y - g(T; \boldsymbol{\beta}))]. \tag{2.3}$$

This result is very insightful. It tells us that the selection bias in the *unconfounded treatment assignment* can be corrected through covariate-balancing. More importantly, it says that the true value $\boldsymbol{\beta}^*$ can be identified from the observed data. The weighted optimization (2.3) provides a unified framework for estimating the causal effect of a variety of treatments, including binary, multi-level, continuous, and mixture of discrete and continuous treatment, and under a general loss function. The goal of this paper is to compute the semiparametric efficiency bound and present an efficient estimation of $\boldsymbol{\beta}^*$ under this general framework.

Although the parametric specification of $g(t; \boldsymbol{\beta})$ is somewhat restrictive, it is useful from a practical point of view. First, if T is a discrete variable, model misspecification is not an issue since the coefficient $\boldsymbol{\beta}^*$ has a clear causal interpretation. Second, if T is a continuous variable, usually a parametric specification may suffer from the model misspecification problem. Since T is univariate, the true response model can be well approximated through several polynomials of t. Third, a parametric specification of $g(t; \boldsymbol{\beta})$ allows us to infer the parameters at $\sqrt{N}$-consistent rate and construct the most efficient estimator. Fourth, the proposed framework (2.1) is more general than the existing literature of continuous treatment (Hirano and Imbens, 2004, Fong, Hazlett, and Imai, 2018), where either a regression model $\mathbb{E}[Y|T, \boldsymbol{X}]$ or a response model $\mathbb{E}[T|\boldsymbol{X}]$ is often required. In Section 8, we also consider fully nonparametric estimation of $g(t)$ under several important cases. The fully nonparametric estimation of $g(t)$ within the general framework (2.1) is beyond the scope of this article, and it will be pursued in a future work.

3 Efficiency bound

We begin by applying the approach of Bickel, Klaassen, Ritov, and Wellner (1993) to compute the semiparametric efficiency bound of the parameter $\boldsymbol{\beta}^*$ defined by (2.1) under Assumption 1. This gives the least possible variance achievable by a regular estimator in the semiparametric model. The result is presented in the following theorem.

Theorem 1. *Suppose that $g(T;\boldsymbol{\beta})$ is twice differentiable with respect to $\boldsymbol{\beta}$ in the parameter space $\Theta \subset \mathbb{R}^p$, with $m(T;\boldsymbol{\beta}^*) := \nabla_\beta g(T;\boldsymbol{\beta}^*)$, and $\mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}))|Y,\boldsymbol{X}\right]$ is differentiable with respect to $\boldsymbol{\beta} \in \Theta$. Denote $\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*) := \mathbb{E}[L'(Y-g(T;\boldsymbol{\beta}^*))|T,\boldsymbol{X}]$, $H_0 := -\nabla_\beta \mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\ \boldsymbol{\beta}))m(T;\boldsymbol{\beta})\right]\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$, and*

$$\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*) := \pi_0(T,\ \boldsymbol{X})m(T;\boldsymbol{\beta}^*)L'(Y-g(T;\boldsymbol{\beta}^*)) - \pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)$$
$$+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}_0)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|T\right] + \mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}_0)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}\right].$$

Suppose that H_0 is nonsingular and $\mathbb{E}\left[\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)\psi(Y,T,\boldsymbol{X};\ \boldsymbol{\beta}^*)^\top\right]$ exists and is finite. Under Assumption 1, namely $Y^*(t) \perp T|\boldsymbol{X}$ for all $t \in \mathcal{T}$, and model (2.1), the efficient influence function of $\boldsymbol{\beta}^*$ is given by

$$S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*) = H_0^{-1}\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*).$$

Consequently, the efficient variance bound of $\boldsymbol{\beta}^*$ is

$$V_{eff} = \mathbb{E}\left[S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)^\top\right].$$

The proof of Theorem 1 is given in the supplemental material Ai, Linton, Motegi, and Zhang (2019, Section 2.1). We can rewrite the influence function $\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)$ defined in Theorem 1 in a more intuitive form. Letting $\varrho\left(Y,T,\boldsymbol{X};\ \boldsymbol{\beta}\right) := \pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta})L'(Y-g(T;\boldsymbol{\beta}))$, we have

$$\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*) = \varrho\left(T,\boldsymbol{X},Y;\boldsymbol{\beta}^*\right) - res_{add}\varrho\left(T,\boldsymbol{X},Y;\boldsymbol{\beta}^*\right),$$

where the operator $res_{add}(\cdot)$ is defined by

$$res_{add}f\left(Y,T,\boldsymbol{X}\right) := \mathbb{E}\left[f\left(T,\boldsymbol{X},Y\right)|T,\boldsymbol{X}\right] - \mathbb{E}_{add}\left[f(T,\boldsymbol{X},Y)|T,\boldsymbol{X}\right],$$
$$\mathbb{E}_{add}\left[f(T,\boldsymbol{X},Y)|T,\boldsymbol{X}\right] := \mathbb{E}\left[f(T,\boldsymbol{X},Y)|T\right] + \mathbb{E}\left[f(T,\boldsymbol{X},Y)|\boldsymbol{X}\right].$$

where the operator $\mathbb{E}_{add}[\cdot]$ projects a random variable on to the space of additive functions

$$\{g(T,\boldsymbol{X}) : g(T,\boldsymbol{X}) = h_T(T) + h_X(\boldsymbol{X})\}$$

inside the space generated by $T,\boldsymbol{X}$, except that the projection is with respect to product measure $dF_T(t) \times dF_X(\boldsymbol{x})$ (Nielsen and Linton, 1998).

In the continuous case, $\pi_0(T,\boldsymbol{X})$ can be written as

$$\pi_0(T,\boldsymbol{X}) = \frac{f_T(T)f_X(\boldsymbol{X})}{f_{T,X}(T,\boldsymbol{X})}$$

and we know that $-\mathbb{E}[\log \pi_0(T, \boldsymbol{X})]$ is the Kullback-Leibler divergence of the joint density from the product of the marginals. The property of $\pi_0(T, \boldsymbol{X})$ in Theorem 1 can also be stated as that for any function $g(T, \boldsymbol{X})$:

$$\mathbb{E}\left[\pi_0(T, \boldsymbol{X}) g(T, \boldsymbol{X})\right] = \int \int g(t, \boldsymbol{x}) f_T(t) dt f_X(\boldsymbol{x}) d\boldsymbol{x},$$

which is the expectation of $g(T, \boldsymbol{X})$ taken with respect to the product measure $f_T(t) f_X(\boldsymbol{x}) dt d\boldsymbol{x}$. In the case where $g(T, \boldsymbol{X})$ is separable the resulting moment factorizes, that is,

$$\mathbb{E}\left[\pi_0(T, \boldsymbol{X}) u(T) v(\boldsymbol{X})\right] = \mathbb{E}\left[u(T)\right] \mathbb{E}\left[v(\boldsymbol{X})\right].$$

Kernel estimators will not satisfy the sample version of this property but they will satisfy the smoothed empirical version, that is:

$$\int \widehat{\pi}_0(t, \boldsymbol{x}) u(t) v(\boldsymbol{x}) dF_N(t, \boldsymbol{x}) \neq \int u(t) dF_N(t) \int v(\boldsymbol{x}) dF_N(\boldsymbol{x}),$$

where $F_N(t, \boldsymbol{x})$ is the joint empirical measure and $F_N(t)$ and $F_N(\boldsymbol{x})$ are the marginals, but

$$\int \widehat{\pi}_0(t, \boldsymbol{x}) u(t) v(\boldsymbol{x}) dF_N^*(t, \boldsymbol{x}) = \int u(t) dF_N^*(t) \int v(\boldsymbol{x}) dF_N^*(\boldsymbol{x}),$$

where $F_N^*(t, \boldsymbol{x})$ is the smoothed empirical distribution function (i.e., $dF_N^*(t, \boldsymbol{x})$ is the kernel density estimator used in constructing $\widehat{\pi}_0(t, \boldsymbol{x})$).

It is worth noting that our bound V_{eff} is equal to: the bound of Hahn (1998) for the case of binary average treatment, the bound of Cattaneo (2010) for the case of multi-valued average treatment, and the bound of Firpo (2007) for the case of binary quantile treatment (see Ai, Linton, Motegi, and Zhang, 2019, Sections 2.2-2.4). Moreover, our bound applies to a much wider class of models, including quantile causal effect of multi-valued, continuous, and mixture of discrete and continuous treatment as well as the asymmetric least squares estimation of the causal effect of all kinds of treatments.

Based on the expression of the efficient influence function, many papers construct an efficient estimator by solving the estimated efficient score equation (Athey, Imbens, Pham, and Wager, 2017, Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018) . Such estimators typically have the double or multiple robustness property. However, in our case the efficient influence function $S_{eff}(T, \boldsymbol{X}, Y; \boldsymbol{\beta})$ involves five unknown functionals $f_T(T)$, $f_{T|X}(T|\boldsymbol{X})$, $\varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta})$, $\mathbb{E}[\pi_0(T, \boldsymbol{X}) \varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}) m(T, \boldsymbol{\beta})|T]$, and $\mathbb{E}[\pi_0(T, \boldsymbol{X}) \varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}) m(T, \boldsymbol{\beta})|\boldsymbol{X}]$. Estimation of these functionals is quite difficult in practice, and we expect that the finite sample performance of the estimated $\boldsymbol{\beta}^*$ would be poor. Instead of explicitly estimating the efficient influence function S_{eff}, we propose a

simple weighted optimization estimator based on (2.3) by estimating the stabilized weights $\pi_0(T, \boldsymbol{X})$. This procedure is remarkably stable numerically and performs well statistically in small samples as we demonstrate in the Monte Carlo section.

It is also worth noting that, if the stabilized weights are known and $g(t; \boldsymbol{\beta}^*)$ is correctly specified, one can estimate $\boldsymbol{\beta}^*$ by solving the sample analogue of the weighted optimization (2.3). The asymptotic variance of this estimator is

$$V_{ineff} = \mathbb{E}\left[S_{ineff}(Y, T, \boldsymbol{X}; \boldsymbol{\beta}^*) S_{ineff}(Y, T, \boldsymbol{X}; \boldsymbol{\beta}^*)^\top\right],$$

with

$$S_{ineff}(Y, T, \boldsymbol{X}; \boldsymbol{\beta}^*) = H_0^{-1} \cdot \pi_0(T, \boldsymbol{X}) m(T; \boldsymbol{\beta}^*) L' \left\{Y - g(T; \boldsymbol{\beta}^*)\right\}.$$

It is easy to show that $V_{ineff} > V_{eff}$ (see Proposition C.1 of Appendix C), implying that the weighted optimization estimator is not efficient. This follows because the weighted optimization does not account for the restriction on the stabilized weight $\pi_0(t, \boldsymbol{x})$ that

$$\mathbb{E}\left[\pi_0(T, \boldsymbol{X}) u(T) v(\boldsymbol{X})\right] = \mathbb{E}[u(T)] \cdot \mathbb{E}[v(\boldsymbol{X})] \tag{3.1}$$

holds for any suitable functions $u(t)$ and $v(\boldsymbol{x})$. Incorporating restriction (3.1) into the estimation of the causal effect can improve efficiency. A similar observation was made by Hirano, Imbens, and Ridder (2003) in the binary treatment. Exactly how to incorporate restriction (3.1) into the estimation is the subject of the next section.

4 Efficient estimation

One way to incorporate (3.1) into the estimation is to estimate the stabilized weights from (3.1) and then implement (2.3) with the estimated weights. But before doing so, we must verify that (3.1) uniquely identifies $\pi_0(T, \boldsymbol{X})$.

Theorem 2. *For any integrable functions* $u(T)$ *and* $v(\boldsymbol{X})$, $\mathbb{E}\left[\pi(T, \boldsymbol{X}) u(T) v(\boldsymbol{X})\right] = \mathbb{E}[u(T)] \cdot \mathbb{E}[v(\boldsymbol{X})]$ *holds if and only if* $\pi(T, \boldsymbol{X}) = \pi_0(T, \boldsymbol{X})$ *a.s.*.

The proof is presented in Appendix B. Therefore, condition (3.1) identifies the stabilized weights. The challenge now is that (3.1) implies an infinite number of moment conditions. With a finite sample of observations, it is impossible to solve an infinite number of equations. To overcome this difficulty, we approximate the (infinite dimensional) function space with the (finite dimensional) sieve space. Specifically, let $u_{K_1}(T) = (u_{K_1,1}(T), \ldots, u_{K_1,K_1}(T))^\top$ and $v_{K_2}(\boldsymbol{X}) = (v_{K_2,1}(\boldsymbol{X}), \ldots, v_{K_2,K_2}(\boldsymbol{X}))^\top$ denote the known basis functions with dimensions $K_1 \in \mathbb{N}$ and $K_2 \in \mathbb{N}$ respectively, and let $K := K_1 \cdot K_2$. The

functions $u_{K_1}(t)$ and $v_{K_2}(\boldsymbol{x})$ are called the *approximation sieves* that can approximate any suitable functions $u(t)$ and $v(\boldsymbol{x})$ arbitrarily well (see Newey, 1997, Chen, 2007, for more discussion on sieve approximation). Since the sieve approximating space is also a subspace of the function space, $\pi_0(T, \boldsymbol{X})$ satisfies

$$\mathbb{E}\left[\pi_0(T, \boldsymbol{X}) u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right] = \mathbb{E}[u_{K_1}(T)] \cdot \mathbb{E}[v_{K_2}(\boldsymbol{X})]^\top. \tag{4.1}$$

Unfortunately, it is not the only solution. Indeed, for any monotonic increasing and globally concave function $\rho(v)$, with

$$\Lambda^*_{K_1 \times K_2} = \arg \max_{\Lambda \in \mathbb{R}^{K_1 \times K_2}} \mathbb{E}\left[\rho(u_{K_1}(T)^\top \Lambda v_{K_2}(\boldsymbol{X}))\right] - \mathbb{E}[u_{K_1}(T)]^\top \Lambda \mathbb{E}[v_{K_2}(\boldsymbol{X})], \tag{4.2}$$

$\pi^*_K(T, \mathbf{X}) = \rho'\left(u_{K_1}(T)^\top \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{X})\right)$ also solves (4.1), where $\rho'(v)$ denotes the first derivative. Let $\pi_K(T, \mathbf{X}) = \rho'\left(u_{K_1}(T)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{X})\right)$ denote the best approximation of $\pi_0(T, \boldsymbol{X})$ under L_∞ norm and suppose that $\|\pi_K - \pi_0\|_\infty = O(K^{-\alpha})$ for some $\alpha > 0$. Then, $\|\pi^*_K - \pi_0\|_{L^2} = O\left(K^{-\alpha}\right)$ (see Ai, Linton, Motegi, and Zhang, 2019, Lemma 3.1).

Let $\{T_i, \boldsymbol{X}_i, Y_i\}_{i=1}^N$ denote an independently and identically distributed sample of observations drawn from the joint distribution of $(T, \mathbf{X}, Y)$. We propose to estimate the stabilized weights $\pi_i = \pi_0(T_i, \mathbf{X}_i)$ by solving the entropy maximization problem:

$$\begin{cases} \max\left\{-\sum_{i=1}^N \pi_i \log \pi_i\right\} \\ \text{subject to } \frac{1}{N}\sum_{i=1}^N \pi_i u_{K_1}(T_i) v_{K_2}(\boldsymbol{X}_i)^\top = \left(\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\right)\left(\frac{1}{N}\sum_{j=1}^N v_{K_2}(\boldsymbol{X}_j)^\top\right). \end{cases} \tag{4.3}$$

Noting $\sum_{i=1}^N N^{-1}\pi_i = 1$ (since both $u_{K_1}(T)$ and $v_{K_2}(\boldsymbol{X})$ contain the constant 1) and

$$\max\left\{-\sum_{i=1}^N \pi_i \log \pi_i\right\} = -\min\left\{\sum_{i=1}^N \{N^{-1}\pi_i\} \cdot \log \frac{N^{-1}\pi_i}{N^{-1}}\right\},$$

the formulation (4.3) can be interpreted as the minimization of the Kullback-Leibler divergence between the estimated weights $\{N^{-1}\pi_i\}_{i=1}^N$ and the empirical frequencies $\{N^{-1}\}$ subject to the empirical moment constraints (4.1). This idea is similar to the exponential tilting (ET) idea developed in Kitamura and Stutzer (1997) and Imbens, Spady, and Johnson (1998). The difference is that they consider a parametric problem and we consider a nonparametric problem.

The primal problem (4.3) is difficult to compute. We instead consider its dual problem, which can be solved by numerically efficient and stable algorithms. Specifically, let $\rho(v) := -e^{-v-1}$ for any $v \in \mathbb{R}$, by Tseng and Bertsekas (1991), we can show that the dual solution is given by

$$\hat{\pi}_K(T_i, \boldsymbol{X}_i) := \rho'\left(u_{K_1}(T_i)^\top \hat{\Lambda}_{K_1 \times K_2} v_{K_2}(\boldsymbol{X}_i)\right), \tag{4.4}$$

where $\hat{\Lambda}_{K_1\times K_2}$ is the maximizer of the strictly concave function $\hat{G}_{K_1\times K_2}$ defined by

$$\widehat{\Lambda}_{K_1\times K_2} = \arg\max_{\Lambda} \hat{G}_{K_1\times K_2}(\Lambda) := \frac{1}{N}\sum_{i=1}^{N} \rho\left(u_{K_1}(T_i)^\top \Lambda v_{K_2}(\boldsymbol{X}_i)\right) - \left(\frac{1}{N}\sum_{i=1}^{N} u_{K_1}(T_i)\right)^\top \Lambda \left(\frac{1}{N}\sum_{j=1}^{N} v_{K_2}(\boldsymbol{X}_j)\right). \tag{4.5}$$

By the first order condition, the constraints of (4.3) are automatically satisfied by $\{\hat{\pi}_K(T_i, \boldsymbol{X}_i)\}_{i=1}^N$. The duality between (4.3) and (4.5) is shown in Appendix D. By Ai, Linton, Motegi, and Zhang (2019, Corollary 3.3), we have

$$\int_{\mathcal{T}\times\mathcal{X}} \left|\hat{\pi}_K(t,\boldsymbol{x}) - \pi^*_K(t,\boldsymbol{x})\right|^2 dF_{T,\boldsymbol{X}}(t,\boldsymbol{x}) = O_p\left(\sqrt{\frac{K}{N}}\right).$$

Having estimated the weights, we now estimate $\boldsymbol{\beta}^*$ by solving the generalized optimization, that is,

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \hat{\pi}_K(T_i, \boldsymbol{X}_i) L\left(Y_i - g(T_i;\boldsymbol{\beta})\right). \tag{4.6}$$

Remarks:

1. Alternatively, one can estimate the stabilized weights by estimating the generalized propensity score function as well as the marginal distribution of the treatment variable nonparametrically (e.g., kernel estimation). But these alternatively estimated weights do not satisfy the empirical moment in (4.3) and may not result in efficient estimation of the causal effect.

2. The primal problem (4.3) is different from the empirical likelihood (Smith, 1997, Imbens, 2002). Notice that $\rho(v) = -e^{-v-1}$ satisfies the invariance property (i.e., $-\rho''(v) = \rho'(v)$). It turns out that this invariance property is critical for establishing consistency of the generalized optimization estimator. Any other choice of $\rho(\cdot)$ that does not have the invariance property may result in biased causal effect estimation.

3. The proposed estimation (4.6) is a semiparametric estimation problem that contains both finite dimensional and infinite unknown parameters. The general semiparametric estimation problems have been studied by Ai and Chen (2003) and Chen, Linton, and Van Keilegom (2003). Ai and Chen (2003) study the large sample properties under smooth objective functions, and Chen, Linton, and Van Keilegom (2003) extend those to nonsmooth criterion functions. Equation (4.6) is a special case of the general setting of Chen, Linton, and Van Keilegom (2003), and we will indeed apply their Theorem 2 (page 1594) to derive the asymptotic properties of $\hat{\boldsymbol{\beta}}$. There is

a major difference between the present paper and Chen, Linton, and Van Keilegom (2003), however. Our focus is on the efficiency bound derivation and efficient estimation, whereas their focus is on deriving the asymptotic properties of the sequential estimator under high level conditions (e.g., Condition 2.6, page 1594). These high level conditions are nontrivial to verify. Most of our derivations are indeed verifying those high level conditions, see Section 4.2 of the supplemental material Ai, Linton, Motegi, and Zhang (2019).

Related methods

In the binary treatment effect model with $T \in \{0,1\}$, the propensity score is defined by $\pi(\boldsymbol{X}) := P(T=1|\boldsymbol{X})$. Then the stabilized weight reduces to $\pi_0(T,\boldsymbol{X}) = T\pi^{-1}(\boldsymbol{X}) \cdot P(T=1) + (1-T)\{1-\pi(\boldsymbol{X})\}^{-1} \cdot P(T=0)$. By setting $u(T) = (T, 1-T)^\top$ in (3.1), we obtain the covariate balancing equation of propensity score:

$$\mathbb{E}\left[T \cdot \pi(\boldsymbol{X})^{-1} v(\boldsymbol{X})\right] = \mathbb{E}[v(\boldsymbol{X})] = \mathbb{E}\left[\{1-T\} \cdot \{1-\pi(\boldsymbol{X})\}^{-1} v(\boldsymbol{X})\right]. \tag{4.7}$$

Our proposed estimator of stabilized weights (4.4) becomes

$$\widehat{\pi}_K(T_i, \boldsymbol{X_i}) = T_i \rho'\left(\widehat{\lambda}_{1K}^\top v_K(\boldsymbol{X}_i)\right) + (1-T_i)\rho'\left(\widehat{\lambda}_{2K}^\top v_K(\boldsymbol{X}_i)\right),$$

where

$$\widehat{\lambda}_{1K} = \arg\max_{\lambda_1} \left\{ \frac{\sum_{i=1}^N T_i \rho\left(\lambda_1^\top v_K(\boldsymbol{X}_i)\right)}{\sum_{i=1}^N T_i} - \frac{1}{N}\sum_{i=1}^N \lambda_1^\top v_K(\boldsymbol{X}_i) \right\},$$
$$\widehat{\lambda}_{2K} = \arg\max_{\lambda_2} \left\{ \frac{\sum_{i=1}^N \{1-T_i\} \rho\left(\lambda_2^\top v_K(\boldsymbol{X}_i)\right)}{\sum_{i=1}^N \{1-T_i\}} - \frac{1}{N}\sum_{i=1}^N \lambda_2^\top v_K(\boldsymbol{X}_i) \right\}.$$

Based on the covariate balancing moment (4.7), various estimators of average treatment effects have been proposed in the existing literature. Hirano, Imbens, and Ridder (2003) propose a nonparametric sieve MLE for the propensity score, which is denoted by $\widehat{\pi}(\boldsymbol{X}) = \pi(\widehat{\lambda}^\top v_K(\boldsymbol{X}))$, where $\pi(z) = \exp(z)/\{1+\exp(z)\}$ and $\widehat{\lambda}_K$ maximizes the log-likelihood function $\sum_{i=1}^N \{T_i \log \pi(\widehat{\lambda}^\top v_K(\boldsymbol{X}_i)) + (1-T_i)\log(1-\pi(\widehat{\lambda}^\top v_K(\boldsymbol{X}_i)))\}$. Their estimator attains the efficiency bound of ATE developed by Hahn (1998). From the first order condition, the covariates between treated and control groups are balanced, i.e. $\sum_{i=1}^N T_i \cdot \widehat{\pi}^{-1}(\boldsymbol{X}) v_K(\boldsymbol{X}) = \sum_{i=1}^N \{1-T_i\} \cdot \{1-\widehat{\pi}(\boldsymbol{X})\}^{-1} v_K(\boldsymbol{X})$, but the covariate balance between treated and combined groups is not guaranteed. In contrast, our proposed estimator of stabilized weights does not require the estimation of propensity score, and it

satisfies the empirical moment of (4.7) that balances the covariate among the treated, control and combined groups simultaneously. Moreover, in the continuous treatment framework, the ratio function $\pi_0(T, \boldsymbol{X})$ does not produce the likelihood function, hence the application of nonparametric MLE method in the general treatment framework is not straightforward.

Graham, Pinto, and Egel (2012) parametrically model the propensity score $\pi(\boldsymbol{X}) = \pi(\gamma^\top v^*(\boldsymbol{X}))$ by a finite dimensional parameter γ and known $v^*(\boldsymbol{X})$. They estimate γ by solving the empirical moment of (4.7) with $v(\boldsymbol{X}) = v^*(\boldsymbol{X})$. Their estimator attains the efficiency bound if both the propensity score function is correctly specified and the conditional potential outcomes $\{\mathbb{E}[Y^*(t)|\boldsymbol{X}], t \in \{0, 1\}\}$ are linear function of $v^*(\boldsymbol{X})$. Imai and Ratkovic (2014) parametrically model the propensity score by $\pi(\boldsymbol{X}; \gamma)$ and consider the overidentified moment condition with $v(\boldsymbol{X}) = v_K(\boldsymbol{X})$ being a specified K-dimensional vector of covariates, where K is possibly larger than the dimension of γ. They propose to estimate γ through generalized method of moments (GMM) and empirical likelihood (EL). We note neither GMM nor EL leads to the empirical moment of (4.7) because both of them are defined to be the maximizer of certain criteria function rather than directly solving the empirical moment of (4.7). In addition, the estimation of Imai and Ratkovic (2014) is not guaranteed to attain the efficiency bound of ATE developed by Hahn (1998).

5 Large sample properties

To establish the large sample properties of the generalized optimization estimator, we first show that the estimated weight function $\hat{\pi}_K(t, \boldsymbol{x})$ is consistent and compute its convergence rates under both the L_∞ norm and the L_2 norm. The following conditions shall be imposed.

Assumption 2. *(i) The support $\mathcal{X}$ of $\boldsymbol{X}$ is a compact subset of $\mathbb{R}^r$. The support $\mathcal{T}$ of the treatment variable T is a compact subset of $\mathbb{R}$. (ii) There exist two positive constants η_1 and η_2 such that*

$$0 < \eta_1 \leq \pi_0(t, \boldsymbol{x}) \leq \eta_2 < \infty \ , \ \forall (t, \boldsymbol{x}) \in \mathcal{T} \times \mathcal{X} \ .$$

Assumption 3. *There exist $\Lambda_{K_1 \times K_2} \in \mathbb{R}^{K_1 \times K_2}$ and a positive constant $\alpha > 0$ such that*

$$\sup_{(t,\boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}} \left| (\rho'^{-1}(\pi_0(t, \boldsymbol{x})) - u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right| = O(K^{-\alpha}).$$

Assumption 4. *(i) For every K_1 and K_2, the smallest eigenvalues of $\mathbb{E}\left[u_{K_1}(T) u_{K_1}(T)^\top\right]$ and $\mathbb{E}\left[v_{K_2}(\boldsymbol{X}) v_{K_2}(\boldsymbol{X})^\top\right]$ are bounded away from zero uniformly in K_1 and K_2. (ii) There are two sequences of constants $\zeta_1(K_1)$ and $\zeta_2(K_2)$ satisfying $\sup_{t \in \mathcal{T}} \|u_{K_1}(t)\| \leq \zeta_1(K_1)$ and $\sup_{\boldsymbol{x} \in \mathcal{X}} \|v_{K_2}(\boldsymbol{x})\| \leq \zeta_2(K_2)$, $K = K_1(N) K_2(N)$ and $\zeta(K) := \zeta_1(K_1)\zeta_2(K_2)$, such that $\zeta(K) K^{-\alpha} \to 0$ and $\zeta(K)\sqrt{K/N} \to 0$ as $N \to \infty$.*

Assumption 2 (i) restricts both the covariates and treatment level to be bounded. This condition is restrictive but convenient for computing the convergence rate under L_∞ norm. It is commonly imposed in the nonparametric regression literature. This condition can be relaxed, however, if we restrict the tail behavior of the joint distribution of $(\boldsymbol{X}, T)$. Assumption 2 (ii) restricts the weight function to be bounded and bounded away from zero. Given Assumption 2 (i), this condition is equivalent to $dF_{T|X}(T|\boldsymbol{X})$ being bounded away from zero, meaning that each type of individual (denoted by $\boldsymbol{X}$) always have a sufficient portion participating in each level of treatment. This restriction is important for our analysis since each individual participates only in one level of treatment and this condition allows us to construct her statistical counterparts from all other treatments. Although Assumption 2 (ii) is useful in causal analysis and establishing the convergence rates, it is not essential and could be relaxed by allowing η_1 (resp. η_2) to depend on N and to go to zero (resp. infinity) slowly, as $N \to \infty$. Notice that $u_{K_1}(t)^\top \Lambda v_{K_2}(\boldsymbol{x})$ is a linear sieve approximation to any suitable function of $(\boldsymbol{X}, T)$.

Assumption 3 requires the sieve approximation error of $\rho'^{-1}(\pi_0(t, \boldsymbol{x}))$ to shrink at a polynomial rate. This condition is satisfied for a variety of sieve basis functions. For example, if both $\boldsymbol{X}$ and T are discrete, then the approximation error is zero for sufficient large K and in this case Assumption 3 is satisfied with $\alpha = +\infty$. If some components of $(\boldsymbol{X}, T)$ are continuous, the polynomial rate depends positively on the smoothness of $\rho'^{-1}(\pi_0(t, \boldsymbol{x}))$ in continuous components and negatively on the number of the continuous components; indeed, for power series and B-splines, $\alpha = -s/r$, where s is the smoothness of approximand and r is the dimension of $\boldsymbol{X}$. Hence, the proposed method still suffers from the curse of dimensionality that typically occurs in nonparametric estimation. We will show that the convergence rate of the estimated weight function (and consequently the rate of the generalized optimization estimator) is bounded by this polynomial rate.

Assumption 4 (i) essentially ensures the sieve approximation estimator is non-degenerate. Similar conditions are common in the sieve regression literature (Andrews, 1991, Newey, 1997). If the approximation error is nonzero, Assumption 4 (ii) requires it to shrink to zero at an appropriate rate as the sample size increases. Newey (1994, 1997) show that if $u_{K_1}(t)$ (resp. $u_{K_2}(\boldsymbol{x})$) is a power series then $\zeta_1(K_1) = O(K_1)$ (resp. $\zeta_2(K_2) = O(K_2)$), and if $u_{K_1}(t)$ (resp. $u_{K_2}(\boldsymbol{x})$) is a B-spline then $\zeta_1(K_1) = O(\sqrt{K_1})$ (resp. $\zeta_2(K_2) = O(\sqrt{K_2})$).

Under these conditions, we are able to establish the following theorem:

Theorem 3. *Suppose that Assumptions 2-4 hold. Then, we obtain the following:*

$$\int_{\mathcal{T}\times\mathcal{X}} |\hat{\pi}_K(t, \boldsymbol{x}) - \pi_0(t, \boldsymbol{x})|^2 dF_{T,X}(t, \boldsymbol{x}) = O_p\left(\max\left\{K^{-2\alpha}, \frac{K}{N}\right\}\right),$$

$$\frac{1}{N}\sum_{i=1}^{N}|\hat{\pi}_K(T_i,\boldsymbol{X}_i)-\pi_0(T_i,\boldsymbol{X}_i)|^2=O_p\left(\max\left\{K^{-2\alpha},\frac{K}{N}\right\}\right).$$

The proof of Theorem 3 immediately follows from the supplemental material Ai, Linton, Motegi, and Zhang (2019, Lemma 3.1 & Corollary 3.3).

The following additional condition is needed to establish the consistency of the proposed estimator $\hat{\boldsymbol{\beta}}$.

Assumption 5. *(i) The parameter space $\Theta \subset \mathbb{R}^p$ is a compact set and the true parameter $\boldsymbol{\beta}_0$ is in the interior of Θ , where $p \in \mathbb{N}$. (ii) $L\left(Y-g(T;\boldsymbol{\beta})\right)$ is continuous in $\boldsymbol{\beta}$, $\sup_{\boldsymbol{\beta}\in\Theta}\mathbb{E}\left[|L\left(Y-g(T;\boldsymbol{\beta})\right)|^2\right]<\infty$ and $\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in\Theta}\left|L\left(Y-g(T;\boldsymbol{\beta})\right)\right|\right]<\infty$.*

Assumption 5 (i) is commonly imposed in the nonlinear regression literature, but can be relaxed if $g(t;\boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$. Assumption 5 (ii) is an envelope condition that is sufficient for the applicability of the uniform law of large numbers. A similar condition is also imposed in Newey and McFadden (1994, Lemma 2.4).

Under these and other conditions, we establish the consistency of the generalized optimization estimator. The proof of Theorem 4 is given in the supplemental material Ai, Linton, Motegi, and Zhang (2019, Section 4.1)

Theorem 4. *Suppose that Assumptions 1-5 hold. Then, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|\xrightarrow{p}0$.*

To establish the asymptotic distribution of the proposed estimator, we need some smoothness condition on the regression function and some under-smoothing condition on the sieve approximation (i.e., larger K than needed for consistency). We also have to address the possibility of a nonsmooth loss function. These conditions are presented below.

Assumption 6.

(i) The loss function $L(v)$ is differentiable almost everywhere, $g(t;\boldsymbol{\beta})$ is twice continuously differentiable in $\boldsymbol{\beta}\in\Theta$ and we denote its first derivative by $m(t;\boldsymbol{\beta}):=\nabla_\beta g(t;\boldsymbol{\beta})$;

(ii) $\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\boldsymbol{\beta}))m(T;\boldsymbol{\beta})\right]$ is differentiable with respect to $\boldsymbol{\beta}$ and $H_0:=-\nabla_\beta\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\boldsymbol{\beta}))m(T;\ \boldsymbol{\beta})\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ is nonsingular;

(iii) $\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*):=\mathbb{E}[L'(Y-g(T;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}]$ is continuously differentiable in $(t,\boldsymbol{x})$;

(iv) Suppose that $N^{-1}\sum_{i=1}^{N}\hat{\pi}_K(T_i,\boldsymbol{X}_i)L'\left(Y_i-g(T_i;\hat{\boldsymbol{\beta}})\right)m(T_i;\hat{\boldsymbol{\beta}})=o_p(N^{-1/2})$ holds with probability approaching one.

Assumption 7. *(i)* $\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in\Theta}|L'(Y-g(T;\boldsymbol{\beta}))|^{2+\delta}\right]<\infty$ *for some* $\delta>0$*; (ii) The function class* $\{L'(y-g(t;\ \boldsymbol{\beta})):\boldsymbol{\beta}\in\Theta\}$ *satisfies:*

$$\mathbb{E}\left[\sup_{\boldsymbol{\beta}_1:\|\boldsymbol{\beta}_1-\boldsymbol{\beta}\|<\delta}|L'(Y-g(T;\ \boldsymbol{\beta}_1))-L'(Y-g(T;\boldsymbol{\beta}))|^2\right]^{1/2}\leq a\cdot\delta^b$$

for any $\boldsymbol{\beta}\in\Theta$ *and any small* $\delta>0$ *and for some finite positive constants* a *and* b*.*

Assumption 6 (i) imposes sufficient regularity conditions on both regression function and loss function. These conditions permit nonsmooth loss functions and are satisfied by the example loss functions mentioned in previous sections. Assumption 6 (ii) ensures that the efficient variance to be finite. Assumption 6 (iv) is essentially saying that the almost sure first order condition is approximately satisfied, see Pakes and Pollard (1989). Assumption 7 is a stochastic equicontinuity condition, which is needed for establishing weak convergence, see Andrews (1994). Again, it is satisfied by widely used loss functions such as $L(v)=v^2$, $L(v)=v\{\tau-I(v\leq 0)\}$, and $L(v)=v^2\cdot|\tau-I(v\leq 0)|$ discussed in Section 2.

Under the above sufficient conditions, we have the following theorem.

Theorem 5. *Suppose that Assumptions 1-7 hold, and strengthen Assumption 4 (ii) to*

$$\textit{Assumption 4 (ii)}'\quad \zeta(K)\sqrt{K^2/N}\to 0 \textit{ and } \sqrt{N}K^{-\alpha}\to 0.$$

Then, $\sqrt{N}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\xrightarrow{d}\mathcal{N}(0,V_{eff})$*, where* $V_{eff}=\mathbb{E}\left[S_{eff}(T,\boldsymbol{X},Y;\boldsymbol{\beta}_0)S_{eff}(T,\boldsymbol{X},Y;\boldsymbol{\beta}_0)^\top\right]$*. Therefore,* $\hat{\boldsymbol{\beta}}$ *attains the semi-parametric efficiency bound of Theorem 1.*

Assumption 4 (ii)′ imposes further restrictions on the smoothing parameter (K) so that the sieve approximation is under-smoothed. This condition is stronger than Assumption 4 (ii) but it is commonly imposed in the semiparametric regression literature. The proof of Theorem 5 is given in the supplemental material Ai, Linton, Motegi, and Zhang (2019, Section 4).

6 Variance estimation

In order to conduct statistical inference, a consistent covariance matrix estimator is needed. Theorem 1 suggests that such consistent covariance can be obtained by replacing H_0 and $\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)$ with some consistent estimates. Since the nonsmooth loss function may invalidate the exchangeability between the expectation and derivative operator, some care

in the estimation of H_0 is warranted. Using the tower property of conditional expectation, we rewrite H_0 as:

$$\begin{aligned} H_0 =& -\nabla_\beta \mathbb{E}\left[\pi_0(T,\boldsymbol{X})\mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}))|T,\boldsymbol{X}\right]m(T;\boldsymbol{\beta})\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\ =& -\mathbb{E}\left[\pi_0(T,\boldsymbol{X})\nabla_\beta \mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}))|T,\boldsymbol{X}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} m(T;\boldsymbol{\beta}^*)^\top\right] \\ & -\mathbb{E}\left[\pi_0(T,\boldsymbol{X})\mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}^*))|T,\boldsymbol{X}\right]\nabla_\beta m(T;\boldsymbol{\beta}^*)\right]. \end{aligned}$$

Applying integration by parts (see Appendix E), we obtain

$$\begin{aligned} &\nabla_\beta \mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}))|T=t,\boldsymbol{X}=\boldsymbol{x}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\ =&\mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}^*))\frac{\partial}{\partial y}\log f_{Y,T,\boldsymbol{X}}(Y,T,\boldsymbol{X})\middle|T=t,\boldsymbol{X}=\boldsymbol{x}\right]m(t;\boldsymbol{\beta}^*) \end{aligned} \tag{6.1}$$

and consequently

$$H_0 = -\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\boldsymbol{\beta}^*))\left\{\frac{\partial}{\partial y}\log f_{Y,T,\boldsymbol{X}}(Y,T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)m(T;\boldsymbol{\beta})^\top + \nabla_\beta m(T;\boldsymbol{\beta}^*)^\top\right\}\right].$$

The log density $\log f_{Y,T,\boldsymbol{X}}(y,t,\boldsymbol{x})$ can be estimated via the widely used sieve extremum estimator (Chen, 2007, Example 2.6, page 5565):

$$\hat{f}_{Y,T,\boldsymbol{X}}(y,t,\boldsymbol{x}) := \frac{\exp\left(\hat{a}_{K_0}^\top r_{K_0}(y,t,\boldsymbol{x})\right)}{\int_{\mathcal{Y}\times\mathcal{T}\times\mathcal{X}}\exp\left(\hat{a}_{K_0}^\top r_{K_0}(y,t,\boldsymbol{x})\right)dydtd\boldsymbol{x}},$$

where $\hat{a}_{K_0}\in\mathbb{R}^{K_0}$ ($K_0\in\mathbb{N}$) maximizes the following concave objective function:

$$\hat{a}_{K_0} := \arg\max_{a\in\mathbb{R}^{K_0}}\frac{1}{N}\sum_{i=1}^N\left[a^\top r_{K_0}(Y_i,T_i,\boldsymbol{X}_i) - \log\int_{\mathcal{Y}\times\mathcal{T}\times\mathcal{X}}\exp\left(a^\top r_{K_0}(y,t,\boldsymbol{x})\right)dydtd\boldsymbol{x}\right],$$

and $r_{K_0}(t,y,\boldsymbol{x})$ is a K_0-dimensional sieve basis. Then H_0 can be estimated by

$$\hat{H} := -\frac{1}{N}\sum_{i=1}^N\hat{\pi}_K(T_i,\boldsymbol{X}_i)L'(Y_i-g(T_i;\hat{\boldsymbol{\beta}}))\left\{\hat{a}_{K_0}^\top\frac{\partial}{\partial y}r_{K_0}(Y_i,T_i,\boldsymbol{X}_i)m(T_i;\hat{\boldsymbol{\beta}})m(T_i;\hat{\boldsymbol{\beta}})^\top + \nabla_\beta m(T_i;\hat{\boldsymbol{\beta}})\right\}.$$

Also, $\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)$ can be directly estimated by the plug-in sieve estimator

$$\begin{aligned} \hat{\psi}(Y,T,\boldsymbol{X};\hat{\boldsymbol{\beta}}) =& \hat{\pi}_K(T,\boldsymbol{X})L'(Y-g(T;\hat{\boldsymbol{\beta}}))m(T;\hat{\boldsymbol{\beta}}) - \hat{\pi}_K(t,\boldsymbol{x})\hat{\mathbb{E}}\left[L'(Y-g(T;\hat{\boldsymbol{\beta}}))|T,\boldsymbol{X}\right]m(T;\hat{\boldsymbol{\beta}}) \\ &+\widehat{\mathbb{E}}\left[\hat{\pi}_K(T,\boldsymbol{X})L'(Y-g(T;\hat{\boldsymbol{\beta}}))|T\right]m(T;\hat{\boldsymbol{\beta}}) + \widehat{\mathbb{E}}\left[\hat{\pi}_K(T,\boldsymbol{X})L'(Y-g(T;\hat{\boldsymbol{\beta}}))|\boldsymbol{X}\right]m(T;\hat{\boldsymbol{\beta}}), \end{aligned}$$

$\widehat{\mathbb{E}}[\hat{\pi}_K(T,\boldsymbol{X})L'(Y-g(T;\hat{\boldsymbol{\beta}}))|T,\boldsymbol{X}]$ is the least square regression of $\hat{\pi}_K(T,\boldsymbol{X})L'(Y-g(T;\hat{\boldsymbol{\beta}}))$ on a sieve basis $w_{K_0}(T,\boldsymbol{X})$, $\widehat{\mathbb{E}}[L'(Y-g(T;\hat{\boldsymbol{\beta}}))|T]$ and $\widehat{\mathbb{E}}[\hat{\pi}_K(T,\boldsymbol{X})L'(Y-$

$g(T;\hat{\boldsymbol{\beta}}))|\boldsymbol{X}]$ are similarly defined. Finally, the asymptotic covariance matrix of the estimator is estimated by

$$\widehat{V} := \hat{H}^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\psi}(Y_i,T_i,\boldsymbol{X}_i;\hat{\boldsymbol{\beta}})\hat{\psi}(Y_i,T_i,\boldsymbol{X}_i;\hat{\boldsymbol{\beta}})^\top\right\}(\hat{H}^\top)^{-1}. \tag{6.2}$$

The sieve extreme estimator is uniformly strong consistent (in the almost sure sense), see Chen (2007, Theorem 3.1). Also from Theorems 3 and 4, we have $\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}|\hat{\pi}_K(t,\boldsymbol{x})-\pi_0(t,\boldsymbol{x})| = o_p(1)$ and $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|\to 0$. With these results, we obtain the consistency of $\hat{V}$.

Theorem 6. *Suppose that Assumptions 1-5 hold. Then, $\widehat{V}$ converges to V_{eff} in probability.*

7 Selection of tuning parameters

The large sample properties of the proposed estimator permit a wide range of values of K_1 and K_2. This presents a dilemma for applied researchers who have only one finite sample and would like to have some guidance on the selection of smoothing parameters. Several data-driven methods of selecting tuning parameters in series estimation have been discussed in Li (1987) and Li and Racine (2007, Section 15.2). Based on that background, we present two data-driven approaches to select K_1 and K_2. The first one is simply minimizing a (penalized) loss function. Define $\bar{L}(K_1,K_2) := N^{-1}\sum_{i=1}^{N}\hat{\pi}_K(T_i,\boldsymbol{X}_i)L(Y_i-g(T_i;\hat{\boldsymbol{\beta}}))$. There are several ways to penalize using large K_1 or K_2:

No penalty. $\mathcal{L}(K_1,K_2)=\bar{L}(K_1,K_2)$.

Additive penalty. $\mathcal{L}(K_1,K_2)=(1+2(K_1+K_2)/N)\times\bar{L}(K_1,K_2)$.

Multiplicative penalty. $\mathcal{L}(K_1,K_2)=(1+2K_1K_2/N)\times\bar{L}(K_1,K_2)$.

Choose (K_1^*,K_2^*) that minimizes $\mathcal{L}(K_1,K_2)$ in some choice sets $(K_1,K_2)\in\mathbb{K}_1\times\mathbb{K}_2$.

The second approach is the J-fold cross-validation (CV), which proceeds as follows.

1. Divide N samples into J groups, (say $J=5$ or 10), and let $n=N/J$. The data in the j^{th} group is denoted by $S_j=\{\boldsymbol{X}_i^{(j)},T_i^{(j)},Y_i^{(j)}:i=1,...,n\}$ for $j\in\{1,..,J\}$.
2. For each $j\in\{1,...,J\}$, we denote the dataset $S_{(-j)}=\{\boldsymbol{X}_i,T_i,Y_i\}_{i=1}^N/S_j$. We compute the following quantities based on $S_{(-j)}$:

$$\hat{\Lambda}_{K_1\times K_2}^{(-j)} = \arg\max_{\Lambda}\hat{G}_K^{(-j)}(\Lambda)$$

$$=\frac{1}{N-n}\sum_{i\in S_{(-j)}}\rho\left(u_{K_1}^\top(T_i)\Lambda v_{K_2}(\boldsymbol{X}_i)\right)-\left[\frac{1}{N-n}\sum_{i\in S_{(-j)}}u_{K_1}^\top(T_i)\right]\Lambda\left[\frac{1}{N-n}\sum_{i\in S_{(-j)}}v_{K_2}(\boldsymbol{X}_i)\right],$$

$$\hat{\pi}_K^{(-j)}(T, \boldsymbol{X}) = \rho'\left(u_{K_1}^\top(T)\hat{\Lambda}_{K_1\times K_2}^{(-j)} v_{K_2}(\boldsymbol{X})\right),$$
$$\hat{\boldsymbol{\beta}}_K^{(-j)} = \arg\min \sum_{i\in S_{(-j)}} \hat{\pi}_K^{(-j)}(T_i, \boldsymbol{X}_i)\left\{Y_i - g(T_i;\boldsymbol{\beta})\right\}^2.$$

3. Choose optimal K_1 and K_2 so that the following cross-validation criterion is minimized:

$$CV(K_1, K_2) = \sum_{j=1}^{J}\left[\sum_{k\in S_j} \hat{\pi}_K^{(-j)}(T_k, \boldsymbol{X}_k)\left\{Y_k - g\left(T_k; \hat{\boldsymbol{\beta}}_K^{(-j)}\right)\right\}^2\right].$$

When $J = 1$, the second approach coincides with the leave-out cross-validation (Stone, 1974). Li (1987) shows that the above procedures to select K_1 and K_2 are asymptotically optimal in the sense of minimizing a weighted loss function for regression.

It should be noted that the K_1 and K_2 chosen by the above criteria are not guaranteed to satisfy the undersmoothing conditions Assumption 4 (ii'), which has been pointed out by Li and Racine (2007, Section 15.2). Linton (1995) and Donald and Newey (2001) develop second order theory to determine the optimal tuning parameters with respect to higher order MSE for a class of semiparametric estimation problems. In general, the optimal rates for K_1 and K_2 according to this criterion are larger reflecting the need for undersmoothing. This suggests that in practice one should take the K_1 and K_2 determined by CV or $\mathcal{L}$ as a lower bound.

8 Some extensions

The condition (2.1) that the causal effect is parameterized may be restrictive for some applications. To relax this condition, we can consider the nonparametric specification:

$$\min_{g(\cdot)} \int_{\mathcal{T}} \mathbb{E}\left[L\left(Y^*(t) - g(t)\right)\right] dF_T(t).$$

Under Assumption 1, the above optimization is equivalent to

$$\min_{g(\cdot)} \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) L(f(Y) - g(T))\right].$$

We can estimate $g(\cdot)$ through the weighted nonparametric sieve regression:

$$\min_{g(\cdot)\in\mathcal{H}_{K_1}} \sum_{i=1}^{N} \hat{\pi}_K(T_i, \boldsymbol{X}_i) L(f(Y_i) - g(T_i)),$$

where $\mathcal{H}_{K_1} := \left\{g(\cdot) : \mathcal{T} \to \mathbb{R}, g(t) = \lambda^\top u_{K_1}(t) : \lambda \in \mathbb{R}^{K_1}\right\}$ is a specified sieve space. The extension to the general loss function requires considerable derivation and shall be dealt with in a separate paper. In this section, we only consider three specific cases: first, the treatment effect curve $\theta_t := \mathbb{E}[Y^*(t)]$, which corresponds to $L(v) = v^2$; second, the average treatment effects (ATE), which is defined by $\theta_{t_1,t_0} := \mathbb{E}[Y^*(t_1) - Y^*(t_0)]$ for $t_1 \neq t_0$; third, the average treatment effects on the treated (ATT), which is defined by $\theta_{t_1,t_0|t_0} := \mathbb{E}[Y^*(t_1) - Y^*(t_0)|T = t_0]$ for $t_1 \neq t_0$.

8.1 Estimation of effect curve and average treatment effects

We begin with estimation of θ_t. Note that, for all $t \in \mathcal{T}$ and under Assumption 1, we can rewrite θ_t as

$$\theta_t := \mathbb{E}[Y^*(t)] = \mathbb{E}\left[\pi_0(T, \boldsymbol{X})Y|T = t\right].$$

With $\pi_0(T, \boldsymbol{X})$ replaced by $\widehat{\pi}_K(T, \boldsymbol{X})$, we estimate θ_t by regressing $\widehat{\pi}_K(T, \boldsymbol{X})Y$ on $u_{K_1}(t)$:

$$\widehat{\theta}_t := \left[\sum_{i=1}^N \widehat{\pi}_K(T_i, \boldsymbol{X}_i)Y_i u_{K_1}(T_i)^\top\right]\left[\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1} u_{K_1}(t).$$

To aid presentation of the asymptotic properties of $\widehat{\theta}_t$, we define: $\Phi_{K_1\times K_1} := \mathbb{E}[u_{K_1}(T)u_{K_1}^\top(T)]$, and

$$\begin{aligned} b_{K_1}(T_i, \boldsymbol{X}_i, Y_i) :&= \pi_0(T_i, \boldsymbol{X}_i)Y_i \cdot u_{K_1}(T_i) - \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i)Y_i \cdot u_{K_1}(T_i)|T_i, \boldsymbol{X}_i\right] \\ &\quad + \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i)Y_i \cdot u_{K_1}(T_i)|\boldsymbol{X}_i\right] - \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i)Y_i \cdot u_{K_1}(T_i)\right], \end{aligned}$$

and

$$\begin{aligned} V_t :=& \mathbb{E}\left[\left\{u_{K_1}^\top(t)\Phi_{K_1\times K_1}^{-1} b_{K_1}(T_i, \boldsymbol{X}_i, Y_i)\right\}^2\right] \\ =& u_{K_1}^\top(t) \cdot \Phi_{K_1\times K_1}^{-1} \cdot \mathbb{E}\left[b_{K_1}(T_i, \boldsymbol{X}_i, Y_i)b_{K_1}^\top(T_i, \boldsymbol{X}_i, Y_i)\right] \cdot \Phi_{K_1\times K_1}^{-1} \cdot u_{K_1}(t). \end{aligned}$$

Theorem 7. *Suppose* $\sup_{t\in\mathcal{T}} |\theta_t - (\gamma^*)^\top u_{K_1}(t)| = O(K_1^{-\tilde{\alpha}})$ *holds for some* $\tilde{\alpha} > 0$ *and* $\gamma^* \in \mathbb{R}^{K_1}$, $\lambda_{\min}\left\{\mathbb{E}\left[b_{K_1}(T, \boldsymbol{X}, Y)b_{K_1}^\top(T, \boldsymbol{X}, Y)\right]\right\} \geq \underline{c} > 0$, *and Assumptions 1-4 hold. Then:*

1. *(Consistency)*

$$\int_{\mathcal{T}} |\widehat{\theta}_t - \theta_t|^2 dF_T(t) = O_p\left(\zeta(K)^2\left\{\frac{K}{N} + K^{-2\alpha}\right\} + K_1^{-2\tilde{\alpha}}\right)$$

$$\sup_{t\in\mathcal{T}} |\hat{\theta}_t - \theta_t| = O_p\left(\zeta_1(K_1)\left\{\zeta(K)\left(\sqrt{\frac{K}{N}} + K^{-\alpha}\right) + K_1^{-\tilde{\alpha}}\right\}\right).$$

2. *(Asymptotic Normality) suppose Assumption 4' and* $\sqrt{N}K_1^{-\tilde{\alpha}} \to 0$ *hold. Then for any fixed* $t \in \mathcal{T}$,

$$\sqrt{N}V_t^{-1/2}\left[\hat{\theta}_t - \theta_t\right] \xrightarrow{d} N(0,1).$$

See Ai, Linton, Motegi, and Zhang (2019, Section 5.1) for a proof of Theorem 7.

The proposed estimation procedure can also be used to estimate the average treatment effects (ATE) which is defined by

$$\theta_{t_1,t_0} := \mathbb{E}[Y^*(t_1) - Y^*(t_0)] = \theta_{t_1} - \theta_{t_0} \text{ for } t_1 \neq t_0.$$

The estimator of θ_{t_1,t_0} is defined by $\widehat{\theta}_{t_1,t_0} := \widehat{\theta}_{t_1} - \widehat{\theta}_{t_0}$. Let

$$V_{t_1,t_0} := \mathbb{E}\left[\left\{u_{K_1}^\top(t_1)\Phi_{K_1\times K_1}^{-1}b_{K_1}(T_i,\boldsymbol{X}_i,Y_i) - u_{K_1}^\top(t_0)\Phi_{K_1\times K_1}^{-1}b_{K_1}(T_i,\boldsymbol{X}_i,Y_i)\right\}^2\right]$$
$$= \left\{u_{K_1}(t_1) - u_{K_1}(t_0)\right\}^\top \Phi_{K_1\times K_1}^{-1}\mathbb{E}\left[b_{K_1}(T_i,\boldsymbol{X}_i,Y_i)b_{K_1}^\top(T_i,\boldsymbol{X}_i,Y_i)\right]\Phi_{K_1\times K_1}^{-1}\left\{u_{K_1}(t_1) - u_{K_1}(t_0)\right\}.$$

Similar to prove Theorem 7, we have the following corollary:

Corollary 8. *Suppose* $\sup_{t\in\mathcal{T}}|\theta_t - (\gamma^*)^\top u_{K_1}(t)| = O(K_1^{-\tilde{\alpha}})$ *holds for some* $\tilde{\alpha} > 0$ *and* $\gamma^* \in \mathbb{R}^{K_1}$, $\lambda_{\min}\left\{\mathbb{E}\left[b_{K_1}(T,\boldsymbol{X},Y)b_{K_1}^\top(T,\boldsymbol{X},Y)\right]\right\} \geq \underline{c} > 0$, *Assumptions 1-4' hold, and* $\sqrt{N}K_1^{-\tilde{\alpha}} \to 0$. *Then*

$$\sqrt{N}V_{t_1,t_0}^{-1/2}\left[\widehat{\theta}_{t_1,t_0} - \theta_{t_1,t_0}\right] \xrightarrow{d} N(0,1).$$

Feasible versions of the above CLT's are implemented using plug-in sieve estimation of the unknown quantities. For example, V_t can be estimated by

$$\widehat{V}_t = \frac{1}{N}\sum_{i=1}^N\left\{u_{K_1}^\top(t)\widehat{\Phi}_{K_1\times K_1}^{-1}\widehat{b}_{K_1}(T_i,\boldsymbol{X}_i,Y_i)\right\}^2,$$

where $\widehat{\Phi}_{K_1\times K_1} := N^{-1}\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}^\top(T_i)$,

$$\begin{aligned}\widehat{b}_{K_1}(T_i,\boldsymbol{X}_i,Y_i) :=& \widehat{\pi}_K(T_i,\boldsymbol{X}_i)Y_i\cdot u_{K_1}(T_i) - \widehat{\mathbb{E}}\left[\widehat{\pi}_K(T_i,\boldsymbol{X}_i)Y_i\cdot u_{K_1}(T_i)|T_i,\boldsymbol{X}_i\right]\\ &+\widehat{\mathbb{E}}[\widehat{\pi}_K(T_i,\boldsymbol{X}_i)Y_i\cdot u_{K_1}(T_i)|\boldsymbol{X}_i] - \widehat{\mathbb{E}}[\widehat{\pi}_K(T_i,\boldsymbol{X}_i)Y_i\cdot u_{K_1}(T_i)]\end{aligned}$$

is the plug-in estimates of $b_{K_1}(T_i,\boldsymbol{X}_i,Y_i)$, and $\widehat{\mathbb{E}}[\widehat{\pi}_K(T,\boldsymbol{X})Yu_{K_1}(T)|T,\boldsymbol{X}]$ is the least square regression of $\widehat{\pi}_K(T,\boldsymbol{X})Yu_{K_1}(T)$ on a sieve basis $w_{K_0}(T,\boldsymbol{X})$, and $\widehat{\mathbb{E}}[\widehat{\pi}_K(T,\boldsymbol{X})Y$ $u_{K_1}(T)|\boldsymbol{X}]$ is the least square regression of $\widehat{\pi}_K(T,\boldsymbol{X})Yu_{K_1}(T)$ on a sieve basis $v_{K_0}(\boldsymbol{X})$.

8.2 Average treatment effects on the treated

Another important parameter for program evaluation is the average treatment effects on the treated (ATT), which is defined by

$$\theta_{t_1,t_0|t_0} := \mathbb{E}\left[Y^*(t_1) - Y^*(t_0)|T = t_0\right] \equiv \theta_{t_1|t_0} - \theta_{t_0|t_0} \text{ for } t_1 \neq t_0.$$

Note that $\theta_{t_0|t_0} = \mathbb{E}[Y^*(t_0)|T = t_0] = \mathbb{E}[Y|T = t_0]$, so it can be estimated by regressing Y on $u_{K_1}(t_0)$:

$$\widehat{\theta}_{t_0|t_0} := \left[\sum_{i=1}^{N} Y_i \cdot u_{K_1}^{\top}(T_i)\right] \left[\sum_{i=1}^{N} u_{K_1}(T_i)u_{K_1}^{\top}(T_i)\right]^{-1} u_{K_1}(t_0).$$

The difficulty is to estimate $\theta_{t_1|t_0} = \mathbb{E}[Y^*(t_1)|T = t_0]$ owing to that $Y^*(t_1)$ cannot be observed under the treatment level $T = t_0$. Under Assumption 1, $\theta_{t_1|t_0}$ can be identified as follows:

$$\begin{aligned}
\theta_{t_1|t_0} =& \mathbb{E}\left[Y^*(t_1)|T = t_0\right] = \mathbb{E}\left[\mathbb{E}[Y^*(t_1)|\boldsymbol{X}, T = t_0]|T = t_0\right] \\
=& \mathbb{E}\left[\mathbb{E}[Y^*(t_1)|\boldsymbol{X}, T = t_1]|T = t_0\right] \quad \text{(by Assumption 1)} \\
=& \int \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, T = t_1] \cdot \frac{f_{X|T}(\boldsymbol{x}|t_0)}{f_{X|T}(\boldsymbol{x}|t_1)} \cdot f_{X|T}(\boldsymbol{x}|t_1)d\boldsymbol{x} \\
=& \int \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, T = t_1] \cdot \frac{f_T(t_1)/f_{T|X}(t_1|\boldsymbol{x})}{f_T(t_0)/f_{T|X}(t_0|\boldsymbol{x})} \cdot f_{X|T}(\boldsymbol{x}|t_1)d\boldsymbol{x} \\
=& \mathbb{E}\left[\frac{\pi_0(T, \boldsymbol{X})}{\pi_0(t_0, \boldsymbol{X})} \cdot Y \middle| T = t_1\right] \\
=& \mathbb{E}\left[\frac{\pi_0(T, \boldsymbol{X})}{\pi_0(T - \delta, \boldsymbol{X})} \cdot Y \middle| T = t_1\right], \qquad (8.1)
\end{aligned}$$

where $\delta := t_1 - t_0$. Based on (8.1), we replace $\pi_0(\cdot)$ by the estimator $\widehat{\pi}_K(\cdot)$ then apply sieve regression on $u_{K_1}(t_1)$, so that $\theta_{t_1|t_0}$ can be estimated by

$$\widehat{\theta}_{t_1|t_0} := \left[\sum_{i=1}^{N} \frac{\widehat{\pi}_K(T_i, \boldsymbol{X}_i)}{\widehat{\pi}_K(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}^{\top}(T_i)\right] \left[\sum_{i=1}^{N} u_{K_1}(T_i)u_{K_1}^{\top}(T_i)\right]^{-1} u_{K_1}(t_1).$$

Therefore, $\theta_{t_1,t_0|t_0}$ can be estimated by

$$\widehat{\theta}_{t_1,t_0|t_0} := \widehat{\theta}_{t_1|t_0} - \widehat{\theta}_{t_0|t_0}.$$

To aid presentation of the asymptotic properties of $\widehat{\theta}_{t_1|t_0}$, we define:

$$b_{1,K_1}(T_i, \boldsymbol{X}_i, Y_i) := \frac{f_{T|X}(T_i + \delta|\boldsymbol{X})}{f_{T|X}(T_i|\boldsymbol{X}_i)} \frac{\pi_0(T_i, \boldsymbol{X}_i)^2}{\pi_0(T_i - \delta, \boldsymbol{X}_i)^2} \cdot \mathbb{E}[Y_i|T_i, \boldsymbol{X}_i] \cdot u_{K_1}(T_i)$$

$$
\begin{aligned}
&- \mathbb{E}\left[\frac{f_{T|X}(T_i+\delta|\boldsymbol{X})}{f_{T|X}(T_i|\boldsymbol{X}_i)}\frac{\pi_0(T_i,\boldsymbol{X}_i)^2}{\pi_0(T_i-\delta,\boldsymbol{X}_i)^2}\cdot Y_i\cdot u_{K_1}(T_i)\middle|\boldsymbol{X}_i\right]\\
&- \mathbb{E}\left[\frac{f_{T|X}(T_i+\delta|\boldsymbol{X})}{f_{T|X}(T_i|\boldsymbol{X}_i)}\frac{\pi_0(T_i,\boldsymbol{X}_i)^2}{\pi_0(T_i-\delta,\boldsymbol{X}_i)^2}\cdot Y_i\cdot u_{K_1}(T_i)\middle|T_i\right]\\
&+ \mathbb{E}\left[\frac{f_{T|X}(T_i+\delta|\boldsymbol{X})}{f_{T|X}(T_i|\boldsymbol{X}_i)}\frac{\pi_0(T_i,\boldsymbol{X}_i)^2}{\pi_0(T_i-\delta,\boldsymbol{X}_i)^2}\cdot Y_i\cdot u_{K_1}(T_i)\right],
\end{aligned}
$$

and

$$
\begin{aligned}
b_{2,K_1}(T_i,\boldsymbol{X}_i,Y_i) :=& \frac{\pi_0(T_i,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}\cdot Y_i\cdot u_{K_1}(T_i) - \mathbb{E}\left[\frac{\pi_0(T_i,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}\cdot Y_i\cdot u_{K_1}(T_i)\middle|T_i,\boldsymbol{X}_i\right]\\
&+ \mathbb{E}\left[\frac{\pi_0(T_i,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}\cdot Y_i\cdot u_{K_1}(T_i)\middle|\boldsymbol{X}_i\right] - \mathbb{E}\left[\frac{\pi_0(T_i,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}\cdot Y_i\cdot u_{K_1}(T_i)\right],
\end{aligned}
$$

and

$$
b_{3,K_1}(T_i,Y_i) := u_{K_1}(T_i)\cdot\{Y_i-\mathbb{E}[Y_i|T_i]\}.
$$

Note that the expectation of b_{1,K_1}, b_{2,K_1} and b_{3,K_1} are zeros. Let

$$
V_{t_1,t_0|t_0} := \mathbb{E}\left[\left\{u_{K_1}^\top(t_1)\Phi_{K_1\times K_1}(b_{1,K_1}+b_{2,K_1}) - u_{K_1}^\top(t_0)\Phi_{K_1\times K_1}b_{3,K_1}\right\}^2\right] = \boldsymbol{w}^\top\Sigma_{2K_1\times 2K_1}\boldsymbol{w},
$$

where $\boldsymbol{w} := \left(u_{K_1}^\top(t_1)\cdot\Phi_{K_1\times K_1}, u_{K_1}^\top(t_0)\cdot\Phi_{K_1\times K_1}\right)^\top \in \mathbb{R}^{2K_1}$ and

$$
\Sigma_{2K_1\times 2K_1} := \mathbb{E}\begin{bmatrix}\{b_{1,K_1}+b_{2,K_1}\}\{b_{1,K_1}+b_{2,K_1}\}^\top, & -\{b_{1,K_1}+b_{2,K_1}\}b_{3,K_1}^\top\\ -b_{3,K_1}\{b_{1,K_1}+b_{2,K_1}\}^\top, & b_{3,K_1}b_{3,K_1}^\top\end{bmatrix}.
$$

Theorem 9. *Suppose* $\sup_{t\in\mathcal{T}}|\mathbb{E}[\pi_0(T,\boldsymbol{X})Y/\pi_0(T-\delta,\boldsymbol{X})|T=t]-(\gamma^*)^\top u_{K_1}(t)| = O(K_1^{-\tilde{\alpha}})$ *holds for some* $\tilde{\alpha}>0$ *and* $\gamma^*\in\mathbb{R}^{K_1}$, $\lambda_{\min}\left(\Sigma_{2K_1\times 2K_1}\right)\geq\underline{c}>0$, *Assumptions 1-4' hold, and* $\sqrt{N}K_1^{-\tilde{\alpha}}\to 0$. *Then*

$$
\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\left[\widehat{\theta}_{t_1,t_0|t_0}-\theta_{t_1,t_0|t_0}\right]\xrightarrow{d}N(0,1).
$$

See Ai, Linton, Motegi, and Zhang (2019, Section 5.2) for a proof of Theorem 9. Feasible versions of the above CLT's are implemented using plug-in sieve estimation of the unknown quantities.

9 Monte Carlo simulations

The large sample properties established in previous sections do not indicate how the generalized optimization estimator behaves in finite samples. To evaluate its finite sample performance, we conduct a simulation study on a continuous treatment. We present a simulation design in Section 9.1 and results in Section 9.2.

9.1 Simulation design

Let $\boldsymbol{X}_i = (X_{1i}, X_{2i})^\top$ be covariates, and assume that $\boldsymbol{X}_i \overset{i.i.d.}{\sim} N(0, I_2)$. Error terms are drawn mutually independently as $\xi_i \overset{i.i.d.}{\sim} N(0,1)$ and $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$. We consider four data generating processes (DGPs):

DGP-L1 $T = 1 + 0.2X_1 + \xi$ and $Y = 1 + X_1 + T + \epsilon$. (X_2 does not play any role, and X_1 affects T and Y linearly.)

DGP-NL1 $T = 0.1X_1^2 + \xi$ and $Y = X_1^2 + T + \epsilon$. (X_2 does not play any role, and X_1 affects T and Y non-linearly.)

DGP-L2 $T = 1 + 0.2\sum_{j=1}^2 X_j + \xi$ and $Y = 1 + (1/2)\sum_{j=1}^2 X_j + T + \epsilon$. ($X_1$ and X_2 affect T and Y linearly.)

DGP-NL2 $T = 0.1(\sum_{j=1}^2 X_j)^2 + \xi$ and $Y = 1/2 + [(1/2)\sum_{j=1}^2 X_j]^2 + T + \epsilon$. ($X_1$ and X_2 affect T and Y non-linearly.)

For each DGP, the true link function is $\mathbb{E}[Y(t)] = 1 + t$, a simple linear function with $\beta_1^* = \beta_2^* = 1$. Below we use a linear link function $g(T_i; \beta) = \beta_1 + \beta_2 T_i$, compute the generalized optimization estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^\top$, and examine its performance.

To compute the generalized optimization estimator, two approximating basis functions $u_{K_1}(T)$ and $v_{K_2}(\boldsymbol{X})$ need to be specified. For $u_{K_1}(T)$, $K_1 \in \{2,3,4\} \equiv \mathbb{K}_1$ is considered:

$$u_2(T) = (1, T)^\top,\ u_3(T) = (1, T, T^2)^\top,\ u_4(T) = (1, T, T^2, T^3)^\top.$$

For $v_{K_2}(\boldsymbol{X})$, the choice set $\mathbb{K}_2$ depends on the number of covariates. For DGP-L1 and DGP-NL1, $K_2 \in \{2,3,4\} \equiv \mathbb{K}_2^1$ is considered:

$$v_2(X_1) = (1, X_1)^\top,\ v_3(X_1) = (1, X_1, X_1^2)^\top,\ v_4(X_1) = (1, X_1, X_1^2, X_1^3)^\top. \tag{9.1}$$

For DGP-L2 and DGP-NL2, $K_2 \in \{3, 6, 10\} \equiv \mathbb{K}_2^2$ is considered:

$$\begin{aligned} v_3(\boldsymbol{X}) &= (1, X_1, X_2)^\top, \\ v_6(\boldsymbol{X}) &= (1, X_1, X_2, X_1^2, X_2^2, X_1X_2)^\top, \\ v_{10}(\boldsymbol{X}) &= (1, X_1, X_2, X_1^2, X_2^2, X_1X_2, X_1^3, X_2^3, X_1^2X_2, X_1X_2^2)^\top. \end{aligned} \tag{9.2}$$

In addition to fixed pairs of $(K_1, K_2) \in \mathbb{K}_1 \times \mathbb{K}_2$, the data-driven selections described in Section 7 are employed. First, the (penalized) loss function approaches are implemented

with $L(Y_i - g(T_i; \hat{\boldsymbol{\beta}})) = (Y_i - \hat{\beta}_1 - \hat{\beta}_2 T_i)^2$. Second, the J-folder cross validation is implemented with $J \in \{5, 10\}$.

We also compute Fong, Hazlett, and Imai's (2018) covariate balancing generalized propensity score estimator with a linear model specification and the quadratic loss function. The linear specification is correct under DGP-L1 and DGP-L2, while it is incorrect under DGP-NL1 and DGP-NL2. By comparing our estimator and the parametric estimator of Fong, Hazlett, and Imai (2018), we can highlight the robustness of the former to non-linear DGPs. Fong, Hazlett, and Imai (2018) also propose a nonparametric estimator in their Section 3.3. In their simulation study, the parametric and nonparametric estimators exhibit similar performance for each DGP considered (Fong, Hazlett, and Imai, 2018, Figure 2). Hence, the present paper focuses on the parametric version of their estimator to save space.

Our proposed estimator and the parametric version of Fong, Hazlett, and Imai's (2018) estimator are computed in a simulated sample with size $N \in \{100, 500, 1000\}$, after which another sample is generated and both estimators are computed again. This exercise is repeated $M = 1000$ times.

To evaluate the performance of point estimation, the bias, standard deviation, and root mean squared error (RMSE) of $\hat{\beta}_1$ and $\hat{\beta}_2$ are calculated from (a subset of) $M = 1000$ simulations. In a small portion of the $M = 1000$ samples, $\bar{\pi}_N \equiv (1/N) \sum_{i=1}^N \hat{\pi}_K(T_i, \boldsymbol{X}_i)$, which should be equal to 1 in theory, takes a value far from 1 due to numerical instability in the computation of $\Lambda^*_{K_1 \times K_2}$. The numerical maximization with respect to Λ should lead to a global maximizer $\Lambda^*_{K_1 \times K_2}$ in theory, but optimizing the $K_1 \times K_2$ elements of Λ all at once is often hard in practice. Hence, we calculate the bias, standard deviation, and RMSE from Monte Carlo samples such that $\bar{\pi}_N \in [0.5, 2]$. Other few samples having $\bar{\pi}_N \notin [0.5, 2]$ are simply discarded. (We admit that this computational problem becomes worse as the dimension of $\boldsymbol{X}$ becomes larger.)

The performance of the variance estimation is evaluated as follows. The true covariance matrix of $\hat{\boldsymbol{\beta}}$ is written as

$$V_{eff} = \begin{bmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{bmatrix}.$$

Different DGPs have different true values of (V_{11}, V_{12}, V_{22}), and they are computed in Section 6.2 of the supplemental material Ai, Linton, Motegi, and Zhang (2019). For each DGP, we compute $\hat{\boldsymbol{\beta}}$ based on (K_1, K_2) that leads to sharp point estimation. Then we use other sieve bases of dimension $(K_1', K_2') \in \mathbb{K}_1 \times \mathbb{K}_2$ to re-estimate the propensity score $\pi_{K'}(T, \boldsymbol{X})$. We allow (K_1, K_2) and (K_1', K_2') to be different from each other since $\hat{\pi}_K(T, \boldsymbol{X})$ that leads to sharp point estimation might be different from the $\hat{\pi}_{K'}(T, \boldsymbol{X})$ that

leads to sharp variance estimation.

Using $\hat{\pi}_{K'}(T, \boldsymbol{X})$ and variance-specific sieve bases $v_{M_0}(\boldsymbol{X})$ and $w_{K_0}(T, \boldsymbol{X})$, the variance estimator $\hat{V}_{eff}$ is computed. For $v_{M_0}(\boldsymbol{X})$, $M_0 \in \{2, 3\}$ is used for DGP-L1 and DGP-NL1 and $M_0 \in \{3, 6\}$ is used for DGP-L2 and DGP-NL2 (see (9.1) and (9.2)). For $w_{K_0}(T, \boldsymbol{X})$, $K_0 \in \{3, 5\}$ is used for DGP-L1 and DGP-NL1:

$$w_3(T, X_1) = (1, T, X_1)^\top, \quad w_5(T, X_1) = (1, T, X_1, T^2, X_1^2)^\top.$$

$K_0 \in \{4, 8\}$ is used for DGP-L2 and DGP-NL2:

$$w_4(T, \boldsymbol{X}) = (1, T, X_1, X_2)^\top, \quad w_8(T, \boldsymbol{X}) = (1, T, X_1, X_2, T^2, X_1^2, X_2^2, TX_1)^\top.$$

(Data-driven selection of (K_1', K_2', M_0, K_0) is beyond the scope of the present paper.)

9.2 Simulation results

We discuss point estimation first, and then discuss variance estimation. Since the slope parameter β_2 is economically more important than the intercept β_1, we only report the point estimation results of β_2 in order to conserve space. (Results of β_1 are available upon request.) See Tables 1-4 for the results under each of the four DGPs considered.

In Figure 1, we draw bar charts that depict the share of (K_1, K_2) selected by each data-driven method. To conserve space, we focus on the large sample case $N = 1000$. Results with $N \in \{100, 500\}$ are nearly identical to the results with $N = 1000$ for each method and DGP. Besides, we focus on the MSE-minimization with the additive and multiplicative penalties as well as the 5-folder cross validation in order to save space. (The MSE-minimization without penalty logically prefers the larger values of (K_1, K_2) than that with the penalties. Results with the 10-folder cross validation are almost identical to the results with the 5-folder cross validation. These omitted results are available upon request.)

Under DGP-L1, the generalized optimization estimator (labeled as GOE) has small enough RMSE for any fixed (K_1, K_2) (Table 1). It is not a surprising result since DGP-L1 has a simple linear structure. The data-driven methods often choose $(K_1^*, K_2^*) = (2, 2)$, the simplest possible approximation basis (Figure 1). The RMSE of the parametric version of the covariate balancing generalized propensity score estimator (labeled as CBGPS) is even smaller than the RMSE of GOE. It is not surprising since CBGPS has a correct parametric specification under DGP-L1.

Under DGP-NL1, GOE dominates CBGPS. GOE leads to small enough RMSE as long as $K_2 \geq 3$. The relatively large RMSE under $K_2 = 2$ suggests that X_1^2 needs to be included in $v_{K_2}(X_1)$ (see (9.1)). That is a reasonable result since DGP-NL1 has a quadratic structure.

As desired, any data-driven method considered often selects pairs with $K_2 \geq 3$ (Figure 1). CBGPS, in contrast, fails with the bias being around 0.2. The bias arises from the fact that the linear specification of CBGPS is incorrect under DGP-NL1. This result highlights that GOE performs well for both linear and nonlinear scenarios while CBGPS performs well for linear scenarios only.

The two-covariate scenarios yield similar implications to the single-covariate scenarios. Under DGP-L2, GOE with any fixed (K_1, K_2) has small RMSE (Table 3). The data-driven methods often choose $(K_1^*, K_2^*) = (2, 3)$, the simplest possible approximation basis (Figure 1). The RMSE of CBGPS is even smaller than the RMSE of GOE due to the linear structures of DGP-L2.

Under DGP-NL2, GOE with $K_2 \geq 6$ leads to small RMSE, and any data-driven method considered often selects pairs with $K_2 \geq 6$ as desired (Table 4 and Figure 1). CBGPS, in contrast, fails with substantial bias of around 0.17. This result again highlights the remarkable advantage of GOE relative to CBGPS.

To summarize the point estimation, the generalized optimization estimator performs well in finite samples, and its performance is still good even when the true DGP is nonlinear; in contrast, the existing parametric estimator of Fong, Hazlett, and Imai (2018) is sensitive to model misspecification.

We now discuss the variance estimation results. The values of the true covariance matrix, V_{eff}, are also provided in Tables 5-8. See Ai, Linton, Motegi, and Zhang (2019, Section 6) for how to compute the true values. For each DGP, we compute $\hat{\boldsymbol{\beta}}$ via $(K_1, K_2) = (2, 2)$ for DGP-L1, $(2, 3)$ for DGP-NL1, $(2, 3)$ for DGP-L2, and $(2, 6)$ for DGP-NL2. Recall from Tables 1-4 that those values are optimal values that lead to one of the smallest MSEs in point estimation. Then we present in Tables 5-8 the bias, standard deviation, and RMSE of $\hat{V}_{eff}$ with respect to V_{eff}, where $(K_1', K_2', M_0, K_0) = (3, 3, 3, 5)$ for DGP-L1, $(3, 3, 3, 5)$ for DGP-NL1, $(3, 3, 6, 8)$ for DGP-L2, and $(2, 10, 3, 4)$ for DGP-NL2. Under those values, we observe desired results that $\hat{V}_{eff}$ converges to V_{eff} as sample size N increases. When $N = 1000$, the bias and standard deviation are small enough. Under DGP-NL1 and DGP-NL2, CBGPS suffers from large bias in variance estimation (Tables 6 and 8). That is reasonable since the point estimation is already biased (Tables 2 and 4).

10 Empirical application

We revisit the U.S. presidential campaign data analyzed by Urban and Niebler (2014) and Fong, Hazlett, and Imai (2018). The motivation of the original study, Urban and Niebler (2014), is well summarized in Fong, Hazlett, and Imai (2018, Section 2):

Urban and Niebler (2014) explored the potential causal link between advertising and campaign contributions. Presidential campaigns ordinarily focus their advertising efforts on competitive states, but if political advertising drives more donations, then it may be worthwhile for candidates to also advertise in non-competitive states. The authors exploit the fact that media markets sometimes cross state boundaries. This means that candidates may inadvertently advertise in noncompetitive states when they purchase advertisements for media markets that mainly serve competitive states. By restricting their analysis to noncompetitive states, the authors attempt to isolate the effect of advertising from that of other campaigning, which do not incur these media market spillovers.

The treatment of interest, the number of political advertisements aired in each zip code, can be regarded as a continuous variable since it takes a range of values from 0 to 22379 across $N = 16265$ zip codes. Urban and Niebler (2014) restricted themselves to a binary treatment framework, and they dichotomized the treatment variable by examining whether a zip code received more than 1000 advertisements or not. Their empirical results suggest that advertising in non-competitive states had a significant impact on the level of campaign contributions.

Dichotomizing a continuous treatment variable requires an ad-hoc choice of a cut-off value, and it makes an empirical result hard to interpret. Fong, Hazlett, and Imai (2018) analyzed the continuous version of the treatment variable, taking advantage of their proposed CBGPS method. Their empirical results suggest, contrary to Urban and Niebler (2014), that advertising in non-competitive states did *not* have a significant impact on the level of campaign contributions (cf. Fong, Hazlett, and Imai, 2018, Table 2).

As shown in Section 9, our generalized optimization estimator has a better performance than Fong, Hazlett, and Imai's (2018) parametric CBGPS estimator. Our estimator exhibits a solid performance even if a DGP of treatment T_i or outcome Y_i is nonlinear in covariate $\boldsymbol{X}_i$. It is thus of interest to apply our approach to the continuous version of the treatment variable in order to see how the results change.

10.1 Fong, Hazlett, and Imai's (2018) CBGPS approach

We begin with Fong, Hazlett, and Imai's (2018) parametric CBGPS estimator as a benchmark. It requires a choice of pre-treatment covariates $\boldsymbol{X}_i$ in a generalized propensity score

model. There are eight covariates

$$\boldsymbol{X}_1 = \begin{bmatrix} \log(\text{Population}) \\ \%\text{Over 65} \\ \log(\text{Income}+1) \\ \%\text{Hispanic} \\ \%\text{Black} \\ \text{Population Density} \\ \%\text{College Graduates} \\ \text{Can Commute} \end{bmatrix}. \tag{10.1}$$

Subscript i is omitted for brevity, but (10.1) is defined for each zip code $i \in \{1, \ldots, N\}$. The definition of each covariate is almost self-explanatory (see Fong, Hazlett, and Imai, 2018, Sec. 5 for more details). Following Fong, Hazlett, and Imai (2018, Table 1), we add squared terms to construct a 15×1 vector of pre-treatment covariates:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \{\log(\text{Population})\}^2 \\ \{\%\text{Over 65}\}^2 \\ \{\log(\text{Income}+1)\}^2 \\ \{\%\text{Hispanic}\}^2 \\ \{\%\text{Black}\}^2 \\ \{\text{Population Density}\}^2 \\ \{\%\text{College Graduates}\}^2 \end{bmatrix}. \tag{10.2}$$

The square of "Can Commute" is not added since it is a binary indicator of whether it is possible to commute to zip code i from a competitive state so that $\text{Can Commute} = \{\text{Can Commute}\}^2$.

Let T_i be the treatment of interest (i.e. the number of political advertisements aired in each zip code). The CBGPS approach assumes that the standardized treatment variable

$$T_i^* = s_T^{-1/2}(T_i - \bar{T}) \tag{10.3}$$

follows the standard normal distribution, where $\bar{T} = (1/N)\sum_{i=1}^N T_i$ and $s_T = (1/(N-1))\sum_{i=1}^N (T_i - \bar{T})^2$. Given the data of political advertisements, the normality assumption is far from satisfied (see Panel 1 of Figure 2). Fong, Hazlett, and Imai (2018) therefore run a Box-Cox transformation $T_i' = \{(T_i+1)^\lambda - 1\}/\lambda$ with $\lambda = -0.16$ and then standardize

T_i' according to (10.3). They choose $\lambda = -0.16$, since it yields the greatest correlation between the sample quantiles of the standardized treatment and the corresponding theoretical quantiles of the standard normal distribution. As Fong, Hazlett, and Imai (2018, p.15) admit, the Gaussian approximation is very poor even after running the Box-Cox transformation (see Panels 2-3 of Figure 2). This result suggests that the normality of a standardized treatment is often a too strong assumption to make in practice.

For an outcome model, we consider four cases for covariates $\boldsymbol{Z}_i$:

Case #1. $\boldsymbol{Z}_i = (T_i, T_i^2, 1)^\top$.

Case #2. $\boldsymbol{Z}_i = (T_i, T_i^2, \boldsymbol{SD}_i^\top)^\top$.

Case #3. $\boldsymbol{Z}_i = (T_i, T_i^2, 1, \boldsymbol{X}_{1i}^\top)^\top$.

Case #4. $\boldsymbol{Z}_i = (T_i, T_i^2, \boldsymbol{SD}_i^\top, \boldsymbol{X}_{1i}^\top)^\top$.

Note that $\boldsymbol{SD}_i = (SD_{1i}, SD_{2i}, \ldots, SD_{24i})^\top$, where SD_{ji} is a binary indicator that equals 1 if zip code i belongs to state j and equals 0 otherwise. Any zip code contained in the dataset belongs to one and only one of 24 states (e.g., Alabama, Arkansas, ..., Wyoming).

For each of Cases #1–#4, we compute the parametric CBGPS estimator and its asymptotic 95% confidence bands (see Fong, Hazlett, and Imai, 2018, Sec. 3.2 for procedures). Our main interest lies in the parameters of (T_i, T_i^2) and their statistical significance. See Table 9 for results. It is evident that the empirical results depend critically on a specification of $\boldsymbol{Z}_i$. In Case #2, T_i has a significantly positive impact on Y_i and T_i^2 has a significantly negative impact on Y_i. In the other three cases, both T_i and T_i^2 have *insignificant* impacts on Y_i.

10.2 Generalized optimization approach

A practical advantage of our proposed approach over the CBGPS approach is that we do not require the normality assumption for the treatment variable T. As indicated in Figure 2, the normality assumption is too strong for the number of political advertisements aired in each zip code whether or not the Box-Cox transformation is implemented. The generalized optimization approach allows us to work with the original treatment variable (Panel 1 of Figure 2).

We assume that the link function is quadratic with $p = 3$, i.e.,

$$g(T, \boldsymbol{\beta}) = \beta_1 + \beta_2 T + \beta_3 T^2.$$

Our covariates $\boldsymbol{X}$ are chosen to be identical to Eq. (10.2). Given that the dimension of $\boldsymbol{X}$ is as large as 15, we use simple polynomials with $K_1 = 3$ and $K_2 = 16$ to compute $\hat{\pi}_K(T, \boldsymbol{X})$ and $\boldsymbol{\beta}$:

$$u_{K_1}(T) = (1, T, T^2)^\top, \;\; v_{K_2}(\boldsymbol{X}) = (1, \boldsymbol{X}^\top)^\top.$$

To compute the variance estimator $\hat{V}_{eff}$, we use the same propensity score $\hat{\pi}_K(T, \boldsymbol{X})$ and variance-specific polynomials with $M_0 = 3$ and $K_0 = 17$:

$$v_{M_0}(\boldsymbol{X}) = (1, \boldsymbol{X}^\top)^\top, \; w_{K_0}(T, \boldsymbol{X}) = (1, T, \boldsymbol{X}^\top)^\top.$$

See Table 10 for results. Neither $\hat{\beta}_2$ nor $\hat{\beta}_3$ is different from 0 at the 5% level. Hence there do not exist statistically significant impacts of the political advertisements on the level of campaign contributions Y.

11 Concluding Remarks

The weighted optimization framework provides a unified approach towards estimation of treatment effects, under the condition of unconfounded treatment assignment. We established the semiparametric efficiency of our methodology, but perhaps the main advantage is its relatively simple form and good finite sample properties.

There are several extensions worth pursuing in future projects. First, estimation of the nonparametric causal effect function under general loss function has not been completely dealt with in this paper. But this is an important extension since it removes the burden of parameterizing the causal effect. Second, the extension of the current setting to allow for high dimensional covariates is also an important project. Third, panel data are common in the empirical literature. Our approach is readily applicable to those data, although the efficiency issue is more difficult. All these extensions shall be taken up in future studies.

Acknowledgement

The first author, Chunrong Ai, acknowledges financial support from National Natural Science Foundation of China through project 71873138. The second author, Oliver Linton, acknowledges Cambridge INET for financial support. The third author, Kaiji Motegi, is grateful for the financial support of JSPS KAKENHI Grant Number 19K13670. The last author, Zheng Zhang, acknowledges the financial support from Renmin University of China through the project 297517501221, and the fund for building world-class universities (disciplines) of Renmin University of China.

References

ABADIE, A. (2005): "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 72(1), 1–19.

ABADIE, A., J. D. ANGRIST, AND G. W. IMBENS (1998): "Instrumental Variables Estimation of Quantile Treatment Effects," *NBER Technical Working Paper No. 229*.

ABADIE, A., AND G. W. IMBENS (2006): "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 74(1), 235–267.

——— (2011): "Bias-corrected matching estimators for average treatment effects," *Journal of Business & Economic Statistics*, 29(1), 1–11.

——— (2012): "A martingale representation for matching estimators," *Journal of the American Statistical Association*, 107(498), 833–843.

——— (2016): "Matching on the estimated propensity score," *Econometrica*, 84(2), 781–807.

AI, C., AND X. CHEN (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

AI, C., O. LINTON, K. MOTEGI, AND Z. ZHANG (2019): "Supplemental material for 'A Unified Framework for Efficient Estimation of General Treatment Models'," Discussion paper, University of Florida.

ANDREWS, D. W. K. (1991): "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," *Econometrica*, 59(2), 307–345.

——— (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4, chap. 37, pp. 2247–2294. Citeseer.

ANGRIST, J. D., AND J.-S. PISCHKE (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

ATHEY, S., G. IMBENS, T. PHAM, AND S. WAGER (2017): "Estimating average treatment effects: Supplementary analyses and remaining challenges," *American Economic Review*, 107(5), 278–81.

ATHEY, S., G. W. IMBENS, AND S. WAGER (2018): "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4), 597–623.

BANG, H., AND J. M. ROBINS (2005): "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61(4), 962–973.

BICKEL, P. J., C. A. J. KLAASSEN, Y. RITOV, AND J. A. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press.

BONEVA, L., D. ELLIOTT, I. KAMINSKA, O. LINTON, N. MCLAREN, AND B. MORLEY (2018): "The Impact of QE on liquidity: Evidence from the UK Corporate Bond Purchase Scheme," *Bank of England Working Paper*, Staff Working Paper No.782.

BUCHINSKY, M. (1995): "Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963–1987," *Journal of Econometrics*, 65(1), 109–154.

BUSSO, M., J. DINARDO, AND J. MCCRARY (2014): "New evidence on the finite sample properties of propensity score reweighting and matching estimators," *Review of Economics and Statistics*, 96(5), 885–897.

CAO, W., A. A. TSIATIS, AND M. DAVIDIAN (2009): "Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data," *Biometrika*, 96(3), 723–734.

CATTANEO, M. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138–154.

CATTANEO, M. D., AND M. H. FARRELL (2011): "Efficient estimation of the dose-response function under ignorability using subclassification on the covariates," in *Missing Data Methods: Cross-sectional Methods and Applications*, vol. 27A, pp. 93–127. Emerald Group Publishing Limited.

CHAN, K. C. G., S. C. P. YAM, AND Z. ZHANG (2016): "Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3), 673–700.

CHEN, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6(B), 5549–5632.

CHEN, X., O. LINTON, AND I. VAN KEILEGOM (2003): "Estimation of Semiparametric Models When the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21(1), C1–C68.

CHERNOZHUKOV, V., J. C. ESCANCIANO, H. ICHIMURA, W. K. NEWEY, AND J. M. ROBINS (2018): "Locally robust semiparametric estimation," *arXiv preprint arXiv:1608.00033*.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND B. MELLY (2013): "Inference on counterfactual distributions," *Econometrica*, 81(6), 2205–2268.

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.

DOKSUM, K. (1974): "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *Annals of Statistics*, 2(2), 267–277.

DONALD, S. G., AND Y.-C. HSU (2014): "Estimation and inference for distribution functions and quantile functions in treatment effect models," *Journal of Econometrics*, 178(3), 383–397.

DONALD, S. G., AND W. K. NEWEY (2001): "Choosing the number of instruments," *Econometrica*, 69(5), 1161–1191.

FAN, Y., AND S. S. PARK (2010): "Sharp bounds on the distribution of treatment effects and their statistical inference," *Econometric Theory*, 26(3), 931–951.

FARRELL, M. H. (2015): "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189(1), 1–23.

FIRPO, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259–276.

FLORENS, J. P., J. J. HECKMAN, C. MEGHIR, AND E. VYTLACIL (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76(5), 1191–1206.

FONG, C., C. HAZLETT, AND K. IMAI (2018): "Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements," *Annals of Applied Statistics*, 12(1), 156–177.

FRÖLICH, M., AND B. MELLY (2013): "Unconditional Quantile Treatment Effects Under Endogeneity," *Journal of Business & Economic Statistics*, 31(3), 346–357.

GALVAO, A. F., AND L. WANG (2015): "Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment," *Journal of the American Statistical Association*, 110(512), 1528–1542.

GRAHAM, B. S., C. C. D. X. PINTO, AND D. EGEL (2012): "Inverse probability tilting for moment condition models with missing data," *The Review of Economic Studies*, 79(3), 1053–1079.

HAHN, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

HECKMAN, J. J., H. ICHIMURA, AND P. TODD (1998): "Matching as an econometric evaluation estimator," *The Review of Economic Studies*, 65(2), 261–294.

HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73(3), 669–738.

HIRANO, K., AND G. W. IMBENS (2004): "The propensity score with continuous treatments," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. by A. Gelman, and X.-L. Meng, chap. 7, pp. 73–84. John Wiley & Sons Ltd.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

IMAI, K., AND M. RATKOVIC (2014): "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 243–263.

IMAI, K., AND D. A. VAN DYK (2004): "Causal inference with general treatment regimes: Generalizing the propensity score," *Journal of the American Statistical Association*, 99(467), 854–866.

IMBENS, G., R. SPADY, AND P. JOHNSON (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66(2), 333–357.

IMBENS, G. W. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87(3), 706–710.

——— (2002): "Generalized method of moments and empirical likelihood," *Journal of Business & Economic Statistics*, 20(4), 493–506.

——— (2004): "Nonparametric estimation of average treatment effects under exogeneity: A review," *The Review of Economics and Statistics*, 86(1), 4–29.

IMBENS, G. W., AND J. M. WOOLDRIDGE (2009): "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47(1), 5–86.

KENNEDY, E. H., Z. MA, M. D. MCHUGH, AND D. S. SMALL (2017): "Non-parametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4), 1229–1245.

KITAMURA, Y., AND M. STUTZER (1997): "An information-theoretic alternative to generalized method of moments estimation," *Econometrica*, 65(4), 861–874.

LEHMANN, E. L. (1975): *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

LI, K.-C. (1987): "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set," *The Annals of Statistics*, 15(3), 958–975.

LI, Q., AND J. S. RACINE (2007): *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

LINTON, O. (1995): "Second order approximation in the partially linear regression model," *Econometrica*, 63(5), 1079–1112.

NEWEY, W., AND D. MCFADDEN (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

NEWEY, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382.

——— (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79(1), 147–168.

NEWEY, W. K., AND J. L. POWELL (1987): "Asymmetric least squares estimation and testing," *Econometrica*, 55(4), 819–847.

NIELSEN, J. P., AND O. B. LINTON (1998): "An optimization interpretation of integration and back-fitting estimators for separable nonparametric models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1), 217–222.

PAKES, A., AND D. POLLARD (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, 57(5), 1027–1057.

QIN, J., AND B. ZHANG (2007): "Empirical-likelihood-based inference in missing response problems and its application in observational studies," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1), 101–122.

ROBINS, J. M., M. A. HERNÁN, AND B. BRUMBACK (2000): "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 11(5), 550–560.

ROBINS, J. M., A. ROTNITZKY, AND L. P. ZHAO (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., AND D. B. RUBIN (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55.

ROTHE, C. (2017): "Robust confidence intervals for average treatment effects under limited overlap," *Econometrica*, 85(2), 645–660.

RUBIN, D. B. (1977): "Assignment to treatment group on the basis of a covariate," *Journal of Educational Statistics*, 2(1), 1–26.

SŁOCZYŃSKI, T., AND J. M. WOOLDRIDGE (2018): "A general double robustness result for estimating average treatment effects," *Econometric Theory*, 34(1), 112–133.

SMITH, R. J. (1997): "Alternative semi-parametric likelihood approaches to generalised method of moments estimation," *The Economic Journal*, 107(441), 503–519.

STONE, M. (1974): "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111–133.

TAN, Z. (2010): "Bounded, efficient and doubly robust estimation with inverse weighting," *Biometrika*, 97(3), 661–682.

TSENG, P., AND D. P. BERTSEKAS (1991): "Relaxation methods for problems with strictly convex costs and linear constraints," *Mathematics of Operations Research*, 16(3), 462–481.

URBAN, C., AND S. NIEBLER (2014): "Dollars on the Sidewalk: Should U.S. Presidential Candidates Advertise in Uncontested States?," *American Journal of Political Science*, 58(2), 322–336.

VANSTEELANDT, S., M. BEKAERT, AND G. CLAESKENS (2010): "On model selection and model misspecification in causal inference," *Statistical Methods in Medical Research*, 21(1), 7–30.

WAGER, S., AND S. ATHEY (2018): "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

YIU, S., AND L. SU (2018): "Covariate association eliminating weights: A unified weighting framework for causal effect estimation," *Biometrika*, 105(3), 709–722.

Appendix

A Proof of (2.2)

Using the law of iterated expectation and Assumption 1, we can deduce that

$$\begin{aligned}
&\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L\left(Y-g(T;\boldsymbol{\beta})\right)\right]\\
=&\mathbb{E}\left[\mathbb{E}[\pi(T,\boldsymbol{X})L(Y^*(T)-g(T;\boldsymbol{\beta}))|T,\boldsymbol{X}]\right]\\
=&\int \pi_0(t,\boldsymbol{x})\cdot\mathbb{E}[L(Y^*(T)-g(T;\boldsymbol{\beta}))|T=t,\boldsymbol{X}=\boldsymbol{x}]\,dF_{T|X}(t|\boldsymbol{x})dF_X(\boldsymbol{x})\\
=&\int \mathbb{E}\left[L(Y^*(t)-g(t;\boldsymbol{\beta}))|T=t,\boldsymbol{X}=\boldsymbol{x}\right]dF_T(t)dF_X(\boldsymbol{x})
\end{aligned}$$

$$= \int \mathbb{E}\left[L(Y^*(t) - g(t;\boldsymbol{\beta}))|\boldsymbol{X} = \boldsymbol{x}\right] dF_T(t) dF_X(\boldsymbol{x}) \text{ (using Assumption 1)}$$
$$= \int \mathbb{E}\left[L(Y^*(t) - g(t;\boldsymbol{\beta}))\right] dF_T(t).$$

B Proof of Theorem 2

The sufficient part is obvious. We prove the necessary part. Let $u(T) = \exp(a \cdot T \cdot i)$ and $v(\boldsymbol{X}) = \exp(b^\top \boldsymbol{X} \cdot i)$ be the test functions, where $a \in \mathbb{R}$ and $b \in \mathbb{R}^r$. By assumption,

$$\mathbb{E}\left[\{\pi(T,\boldsymbol{X}) - \pi_0(T,\boldsymbol{X})\}\exp\left\{a \cdot T \cdot i + b^\top \boldsymbol{X} \cdot i\right\}\right] + \mathbb{E}\left[\pi_0(T,\boldsymbol{X})\exp\left\{a \cdot T \cdot i + b^\top \boldsymbol{X} \cdot i\right\}\right]$$
$$= \mathbb{E}\left[\exp(a \cdot T \cdot i)\right] \cdot \mathbb{E}\left[\exp(b^\top \boldsymbol{X} \cdot i)\right].$$

By definition $\mathbb{E}\left[\pi_0(T,\boldsymbol{X})\exp\left\{a \cdot T \cdot i + b^\top \boldsymbol{X} \cdot i\right\}\right] = \mathbb{E}\left[\exp(a \cdot T \cdot i)\right] \cdot \mathbb{E}\left[\exp(b^\top \boldsymbol{X} \cdot i)\right]$. Then $\mathbb{E}\left[\{\pi(T,\boldsymbol{X}) - \pi_0(T,\boldsymbol{X})\}\exp\left\{a \cdot T \cdot i + b^\top \boldsymbol{X} \cdot i\right\}\right] = 0$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}^r$. Dues to the uniqueness of Fourier transform we can obtain $\pi(T,\boldsymbol{X}) = \pi_0(T,\boldsymbol{X})$ a.s..

C Asymptotic result when $\pi_0(T, \mathbf{X})$ is known

Suppose the stabilized weight function $\pi_0(T, \mathbf{X})$ is known, the weighted optimization estimator of $\boldsymbol{\beta}^*$, denoted by $\widehat{\boldsymbol{\beta}}_{known}$, is

$$\widehat{\boldsymbol{\beta}}_{known} = \min_{\beta} \sum_{i=1}^{N} \pi_0(T_i, \boldsymbol{X}_i) L(Y_i - g(T_i;\beta)).$$

We also assume the asymptotic first order condition

$$\frac{1}{N}\sum_{i=1}^{N} \pi_0(T_i, \boldsymbol{X}_i) L'(Y_i - g(T_i;\widehat{\boldsymbol{\beta}}_{known}))m(T_i;\widehat{\boldsymbol{\beta}}_{known}) = o_P(N^{-1/2}) \tag{C.1}$$

holds with probability approaching to one.

Proposition B.1 Suppose Assumptions 5, 6 (i-ii), and 7 hold, and (C.1) holds, then we have

1. $\widehat{\boldsymbol{\beta}}_{known} \xrightarrow{p} \boldsymbol{\beta}^*$;
2. $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known} - \boldsymbol{\beta}^*) \xrightarrow{d} \mathcal{N}(0, V_{ineff})$, where
$$V_{ineff} := H_0^{-1} \cdot \mathbb{E}\left[\pi_0(T,\boldsymbol{X})^2 L'(Y - g(T;\boldsymbol{\beta}^*))^2 m(T;\boldsymbol{\beta}^*) m(T;\boldsymbol{\beta}^*)^\top\right] \cdot H_0^{-1};$$
3. furthermore, if $\mathbb{E}\left[L'(Y(t) - g(t;\boldsymbol{\beta}^*))\right] = 0$ holds for all $t \in \mathcal{T}$, then $V_{ineff} \geq V_{eff}$ in the sense of that $c^\top \cdot V_{ineff} \cdot c \geq c^\top \cdot V_{eff} \cdot c$ for any vector $c \in \mathbb{R}^p$.

Proof. By Assumption 5 and the uniform law of large number, we obtain

$$\frac{1}{N}\sum_{i=1}^{N}\pi_0(T_i,\boldsymbol{X}_i)L\left\{Y_i-g(T_i;\boldsymbol{\beta})\right\}\to\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L\left\{Y-g(T;\boldsymbol{\beta})\right\}\right]\text{ in probability uniformly over }\boldsymbol{\beta},$$

which implies the consistency result $\|\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*\|\xrightarrow{p}0$.

The first order condition (C.1) holds with probability approaching to one. Note that $L'(\cdot)$ may not be a differentiable function, e.g. $L'(v)=\tau-I(v<0)$ in quantile regression, we cannot simply apply Mean Value Theorem on (C.1) to obtain the expression for $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*)$. To solve this problem, we resort to the empirical process theory in Andrews (1994). Define

$$f(\boldsymbol{\beta}):=\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\boldsymbol{\beta}))m(T;\boldsymbol{\beta})\right],$$

which is a differentiable function in $\boldsymbol{\beta}$ and by (2.3) $f(\boldsymbol{\beta}^*)=0$. Using Mean Value Theorem, we can obtain

$$0=\sqrt{N}f(\boldsymbol{\beta}^*)=\sqrt{N}f(\widehat{\boldsymbol{\beta}}_{known})-\nabla_\beta f(\bar{\boldsymbol{\beta}})\cdot\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*),$$

where $\bar{\boldsymbol{\beta}}$ lies on the line joining $\widehat{\boldsymbol{\beta}}_{known}$ and $\boldsymbol{\beta}^*$. Because $\nabla_\beta f(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^*$, and $\|\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*\|\xrightarrow{p}0$, then we have

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*)=[\nabla_\beta f(\boldsymbol{\beta}^*)]^{-1}\cdot\sqrt{N}f(\widehat{\boldsymbol{\beta}}_{known})+o_P(1).$$

Define the empirical process

$$\nu_N(\boldsymbol{\beta})=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\pi_0(T_i,\boldsymbol{X}_i)L'(Y_i-g(T_i;\boldsymbol{\beta}))m(T_i;\boldsymbol{\beta})-\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\boldsymbol{\beta}))m(T;\boldsymbol{\beta})\right]\right\}.$$

By (C.1) and the definition of $\nu_N(\boldsymbol{\beta})$, we have

$$\begin{aligned}
&\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*)\\
=&\nabla_\beta f(\boldsymbol{\beta}^*)^{-1}\cdot\left\{\sqrt{N}f(\widehat{\boldsymbol{\beta}}_{known})-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\pi_0(T_i,\boldsymbol{X}_i)L'(Y_i-g(T_i;\widehat{\boldsymbol{\beta}}_{known}))m(T_i;\widehat{\boldsymbol{\beta}}_{known})\right.\\
&\qquad\left.+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\pi_0(T_i,\boldsymbol{X}_i)L'(Y_i-g(T_i;\widehat{\boldsymbol{\beta}}_{known}))m(T_i;\widehat{\boldsymbol{\beta}}_{known})\right\}\\
=&-\nabla_\beta f(\boldsymbol{\beta}^*)^{-1}\cdot\nu_N(\widehat{\boldsymbol{\beta}}_{known})+o_p(1)\\
=&H_0^{-1}\cdot\left\{\left(\nu_N(\widehat{\boldsymbol{\beta}}_{known})-\nu_N(\boldsymbol{\beta}^*)\right)+\nu_N(\boldsymbol{\beta}^*)\right\}+o_p(1).
\end{aligned}$$

By Assumptions 6, 7, Theorems 4 and 5 of Andrews (1994), we have that $\nu_N(\cdot)$ is stochastically equicontinuous, which implies $\nu_N(\widehat{\boldsymbol{\beta}}_{known}) - \nu_N(\boldsymbol{\beta}^*) \xrightarrow{p} 0$. Therefore,

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known} - \boldsymbol{\beta}^*) = H_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \pi_0(T_i, \boldsymbol{X}_i) L'(Y_i - g(T_i; \boldsymbol{\beta}^*)) m(T_i; \boldsymbol{\beta}^*) + o_p(1) \ ,$$

then we can conclude that the asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known} - \boldsymbol{\beta}^*)$ is V_{ineff}.

We next show $V_{ineff} \geq V_{eff}$. From Theorem 1, we have

$$\begin{aligned}
V_{eff} = H_0^{-1} \cdot \Big\{ & \mathbb{E}\left[\pi_0(T, \boldsymbol{X})^2 L'(Y - g(T; \boldsymbol{\beta}^*))^2 m(T; \boldsymbol{\beta}^*) m(T; \boldsymbol{\beta}^*)^\top\right] \\
& + \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}]^\top\right] \\
& + \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \\
& + \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T]^\top\right] \\
& - 2 \cdot \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*)^\top\right] \\
& - 2 \cdot \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \\
& - 2 \cdot \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T]^\top\right] \\
& + 2 \cdot \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \\
& + 2 \cdot \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}_0)) m(T; \boldsymbol{\beta}^*) | T]^\top\right] \\
& + 2 \cdot \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \Big\} H_0^{-1} \\
= H_0^{-1} \Big\{ & \mathbb{E}\left[\pi_0(T, \boldsymbol{X})^2 L'(Y - g(T; \boldsymbol{\beta}^*))^2 m(T; \boldsymbol{\beta}^*) m(T; \boldsymbol{\beta}^*)^\top\right] \\
& - \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}]^\top\right] \\
& + \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \Big\} H_0^{-1},
\end{aligned}$$

where the last equality holds by noting

$$\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T = t] = \mathbb{E}\left[L'(Y^*(t) - g(t; \boldsymbol{\beta}^*))\right] \cdot m(t; \boldsymbol{\beta}^*) = 0 \ ,$$

since the model is correctly specified, i.e. $\mathbb{E}\left[L'(Y^*(t) - g(t; \boldsymbol{\beta}_0))\right] = 0$ for $t \in \mathcal{T}$. Therefore,

$$\begin{aligned}
& V_{ineff} - V_{eff} \\
= & H_0^{-1} \Big\{ \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}]^\top\right] \\
& - \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \Big\} H_0^{-1} \geq 0 \ ,
\end{aligned}$$

where the last inequality holds by using Jensen's inequality:

$$\mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X})(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X})(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right]$$

$$=\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}[\pi_0(T,\boldsymbol{X})(Y-g(T;\boldsymbol{\beta}^*))m(T;\boldsymbol{\beta}^*)|T,\mathbf{X}]|\boldsymbol{X}\right]\cdot\mathbb{E}\left[\mathbb{E}[\pi_0(T,\boldsymbol{X})(Y-g(T;\boldsymbol{\beta}^*))m(T;\boldsymbol{\beta}^*)|T,\mathbf{X}]|\boldsymbol{X}\right]^\top\right]$$
$$<\mathbb{E}\left[\mathbb{E}[\pi_0(T,\boldsymbol{X})(Y-g(T;\boldsymbol{\beta}^*))m(T;\boldsymbol{\beta}^*)|T,\mathbf{X}]\cdot\mathbb{E}[\pi_0(T,\boldsymbol{X})(Y-g(T;\boldsymbol{\beta}^*))m(T;\boldsymbol{\beta}^*)|T,\mathbf{X}]^\top\right].$$

□

D Duality of primal problem (4.3)

We first introduce some notation:

- Let $m_K(T,\boldsymbol{X}) = \text{vec}\left(u_{K_1}(T)v_{K_2}^\top(\boldsymbol{X})\right)$ denote a K-dimensional column vector formed by the elements of the matrix $u_{K_1}(T)v_{K_2}^\top(\boldsymbol{X})$. Let $M_{K\times N}=(m_K(T_1,\boldsymbol{X}_1),\ldots,m_K(T_N,\boldsymbol{X}_N))$, which is a $K\times N$ matrix.
- Let $u_{K_1,k}(T)$ (resp. $v_{K_2,k'}(\boldsymbol{X})$) denote the k^{th} (resp. k'^{th}) component of $u_{K_1}(T)$ (resp. $v_{K_2}(\boldsymbol{X})$), and denote

$$\overline{u}_{K_1,k}=\frac{1}{N}\sum_{i=1}^N u_{K_1,k}(T_i) \text{ and } \overline{v}_{K_2,k'}=\frac{1}{N}\sum_{i=1}^N v_{K_2,k'}(\boldsymbol{X}_i).$$

  Let b_K be a K dimensional column vector whose elements are formed by $\{\overline{u}_{K_1,k}\overline{v}_{K_2,k'};k=1,...,K_1,k'=1,...,K_2\}$.
- Denote $\boldsymbol{\pi}=(\pi_1,...,\pi_N)$ and $F(\boldsymbol{\pi})=\sum_{i=1}^N\pi_i\log\pi_i$.

The primal optimization problem (4.3) can be written as

$$\begin{cases}\min_{\boldsymbol{\pi}} F(\boldsymbol{\pi})\\ \text{subject to}\quad M_{K\times N}\cdot\boldsymbol{\pi}=N\cdot b_K\end{cases}\tag{D.1}$$

By Tseng and Bertsekas (1991), the conjugate convex function of $F(\cdot)$ is

$$F^*(\mathbf{z})=\sup_{\boldsymbol{\pi}}\sum_{i=1}^N\{z_i\pi_i-\pi_i\log\pi_i\}=\sum_{i=1}^N\{z_i\pi_i^*-\pi_i^*\log\pi_i^*\},$$

where π_j^* satisfies the first order condition:

$$z_j=\log\pi_j^*+1\Rightarrow\pi_j^*=e^{z_j-1}=\rho'(z_i).$$

By substitution, we obtain

$$F^*(\mathbf{z})=\sum_{i=1}^N\left\{z_ie^{z_i-1}-e^{z_i-1}(z_i-1)\right\}=\sum_{i=1}^N e^{z_i-1}=\sum_{i=1}^N-\rho(-z_i).$$

By Tseng and Bertsekas (1991), the dual problem of (D.1) is

$$\begin{aligned}
&\max_{\lambda\in\mathbb{R}^K}\left\{\lambda^\top(N\cdot b_K)-F^*\left(\lambda^\top M_{K\times N}\right)\right\}\\
=&\max_{\Lambda\in\mathbb{R}^{K_1}\times\mathbb{R}^{K_2}}\sum_{i=1}^N\left\{\overline{u}_{K_1}^\top\Lambda\overline{v}_{K_2}+\rho\left(-u_{K_1}(T_i)^\top\Lambda v_{K_2}(\boldsymbol{X}_i)\right)\right\}\\
=&\max_{\Lambda\in\mathbb{R}^{K_1}\times\mathbb{R}^{K_2}}\sum_{i=1}^N\left\{\rho\left(u_{K_1}(T_i)^\top\Lambda v_K(\boldsymbol{X}_i)\right)-\overline{u}_{K_1}^\top\Lambda\overline{v}_{K_2}\right\}\\
=&\max_{\Lambda\in\mathbb{R}^{K_1}\times\mathbb{R}^{K_2}}\hat{G}_{K_1\times K_2}(\Lambda).
\end{aligned}\tag{D.2}$$

Therefore, the dual solution of (4.3) is given by

$$\hat{\pi}_K(T_i,\boldsymbol{X}_i)=\rho'\left(u_{K_1}(T_i)^\top\hat{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)\ ,$$

where $\hat{\Lambda}_{K_1\times K_2}$ is the maximizer of the strictly concave objective function $\hat{G}_{K_1\times K_2}$.

E Proof of (6.1)

$$\begin{aligned}
&\nabla_\beta\mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}))|T=t,\boldsymbol{X}=\boldsymbol{x}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\\
=&\nabla_\beta\left[\int_{\mathbb{R}}L'(y-g(t;\boldsymbol{\beta}))f_{Y|T,X}(y|t,\boldsymbol{x})dy\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\\
=&\nabla_\beta\left[\int_{\mathbb{R}}L'(z)f_{Y|T,X}(z+g(t;\boldsymbol{\beta})|t,\boldsymbol{x})dz\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\ (\text{use } z=y-g(t;\boldsymbol{\beta}))\\
=&\int_{\mathbb{R}}L'(z)\cdot\frac{\partial}{\partial y}f_{Y|T,X}(z+g(t;\boldsymbol{\beta}^*)|t,\boldsymbol{x})dz\cdot m(t;\boldsymbol{\beta}^*)\\
=&\int_{\mathbb{R}}L'(y-g(t;\boldsymbol{\beta}^*))\cdot\frac{\partial}{\partial y}f_{Y|T,X}(y|t,\boldsymbol{x})dy\cdot m(t;\boldsymbol{\beta}^*)\\
=&\int_{\mathbb{R}}L'(y-g(t;\boldsymbol{\beta}^*))\cdot\frac{\frac{\partial}{\partial y}f_{Y|T,X}(y|t,\boldsymbol{x})}{f_{Y|T,X}(y|t,\boldsymbol{x})}f_{Y|T,X}(y|t,\boldsymbol{x})dy\cdot m(t;\boldsymbol{\beta}^*)\\
=&\int_{\mathbb{R}}L'(y-g(t;\boldsymbol{\beta}^*))\cdot\frac{\frac{\partial}{\partial y}f_{Y,T,X}(y,t,\boldsymbol{x})}{f_{Y,T,X}(y,t,\boldsymbol{x})}f_{Y|T,X}(y|t,\boldsymbol{x})dy\cdot m(t;\boldsymbol{\beta}^*)\\
=&\mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}^*))\frac{\frac{\partial}{\partial y}f_{Y,T,\boldsymbol{X}}(Y,T,\boldsymbol{X})}{f_{Y,T,\boldsymbol{X}}(Y,T,\boldsymbol{X})}\middle|T=t,\boldsymbol{X}=\boldsymbol{x}\right]m(t;\boldsymbol{\beta}^*).
\end{aligned}$$

Table 1: Simulation results on point estimation of slope β_2 under DGP-L1 ($\beta_2^* = 1$)

		$N = 100$			$N = 500$			$N = 1000$		
	(K_1, K_2)	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	$(2,2)$	-0.002	0.185	0.185	-0.000	0.081	0.081	-0.001	0.056	0.056
GOE	$(2,3)$	0.010	0.178	0.178	-0.001	0.080	0.080	0.004	0.057	0.057
GOE	$(2,4)$	-0.000	0.196	0.196	-0.005	0.083	0.083	-0.001	0.057	0.057
GOE	$(3,2)$	-0.002	0.185	0.185	-0.001	0.081	0.081	0.002	0.057	0.057
GOE	$(3,3)$	-0.001	0.190	0.190	0.000	0.080	0.080	-0.003	0.057	0.057
GOE	$(3,4)$	-0.007	0.201	0.201	-0.011	0.085	0.086	-0.011	0.060	0.061
GOE	$(4,2)$	-0.005	0.184	0.185	-0.002	0.080	0.080	-0.000	0.055	0.055
GOE	$(4,3)$	-0.007	0.205	0.205	-0.006	0.083	0.084	-0.011	0.060	0.061
GOE	$(4,4)$	-0.020	0.207	0.208	-0.012	0.084	0.084	-0.013	0.062	0.064
GOE	MSE (none)	0.002	0.171	0.171	-0.008	0.079	0.080	-0.006	0.057	0.058
GOE	MSE (add)	-0.013	0.169	0.170	-0.005	0.076	0.076	-0.002	0.057	0.057
GOE	MSE (multi)	0.003	0.165	0.165	-0.001	0.079	0.079	-0.003	0.056	0.056
GOE	CV ($J = 5$)	0.006	0.191	0.191	0.004	0.080	0.080	0.001	0.058	0.058
GOE	CV ($J = 10$)	0.005	0.182	0.182	0.001	0.079	0.079	0.001	0.057	0.057
CBGPS	-	-0.002	0.102	0.103	-0.002	0.045	0.046	-0.002	0.032	0.032

DGP-L1: $T = 1 + 0.2X_1 + \xi$ and $Y = 1 + X_1 + T + \epsilon$, where $X_1 \sim N(0,1)$. "GOE" is the proposed generalized optimization estimator. K_1 and K_2 are the dimensions of the polynomials of T and X_1, respectively. "MSE (none)" signifies that we pick (K_1, K_2) that minimizes $MSE(K_1, K_2) = N^{-1}\sum_{i=1}^{N} \hat{\pi}_K(T_i, X_{1i})(Y_i - \hat{\beta}_1 - \hat{\beta}_2 T_i)^2$. "MSE (add)" signifies that we pick (K_1, K_2) that minimizes $(1 + 2(K_1 + K_2)/N) \times MSE(K_1, K_2)$. "MSE (multi)" signifies that we pick (K_1, K_2) that minimizes $(1 + 2K_1K_2/N) \times MSE(K_1, K_2)$. "CV" signifies that we pick (K_1, K_2) that minimizes the loss function of the J-folder cross validation with $J \in \{5, 10\}$. The choice set of (K_1, K_2) is the nine pairs listed in the table. "CBGPS" is Fong, Hazlett, and Imai's (2018) parametric covariate balancing generalized propensity score estimator. The number of Monte Carlo iterations is $M = 1000$.

Table 2: Simulation results on point estimation of slope β_2 under DGP-NL1 ($\beta_2^* = 1$)

		$N = 100$			$N = 500$			$N = 1000$		
	(K_1, K_2)	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	$(2,2)$	-0.029	0.171	0.174	-0.020	0.075	0.078	-0.021	0.055	0.059
GOE	$(2,3)$	0.001	0.175	0.175	-0.003	0.074	0.074	-0.000	0.054	0.054
GOE	$(2,4)$	-0.004	0.177	0.177	-0.001	0.080	0.080	-0.000	0.057	0.057
GOE	$(3,2)$	-0.042	0.168	0.173	-0.021	0.075	0.077	-0.021	0.054	0.058
GOE	$(3,3)$	-0.017	0.186	0.187	-0.001	0.081	0.081	0.002	0.056	0.056
GOE	$(3,4)$	-0.025	0.180	0.182	-0.004	0.080	0.080	-0.000	0.058	0.058
GOE	$(4,2)$	-0.063	0.170	0.181	-0.028	0.076	0.081	-0.023	0.053	0.058
GOE	$(4,3)$	-0.015	0.197	0.197	-0.002	0.087	0.087	0.001	0.058	0.058
GOE	$(4,4)$	-0.044	0.187	0.192	-0.011	0.081	0.082	-0.003	0.058	0.058
GOE	MSE (none)	-0.061	0.175	0.185	-0.021	0.078	0.081	-0.013	0.057	0.058
GOE	MSE (add)	-0.075	0.164	0.180	-0.021	0.079	0.081	-0.015	0.054	0.056
GOE	MSE (multi)	-0.057	0.171	0.181	-0.017	0.077	0.079	-0.013	0.055	0.056
GOE	CV ($J = 5$)	-0.035	0.174	0.177	-0.010	0.079	0.079	-0.006	0.055	0.056
GOE	CV ($J = 10$)	-0.026	0.171	0.173	-0.013	0.077	0.078	-0.006	0.055	0.055
CBGPS	-	0.189	0.186	0.266	0.190	0.080	0.206	0.195	0.055	0.203

DGP-NL1: $T = 0.1X_1^2 + \xi$ and $Y = X_1^2 + T + \epsilon$, where $X_1 \sim N(0,1)$. "GOE" is the proposed generalized optimization estimator. K_1 and K_2 are the dimensions of the polynomials of T and X_1, respectively. "MSE (none)" signifies that we pick (K_1, K_2) that minimizes $MSE(K_1, K_2) = N^{-1} \sum_{i=1}^{N} \hat{\pi}_K(T_i, X_{1i})(Y_i - \hat{\beta}_1 - \hat{\beta}_2 T_i)^2$. "MSE (add)" signifies that we pick (K_1, K_2) that minimizes $(1 + 2(K_1 + K_2)/N) \times MSE(K_1, K_2)$. "MSE (multi)" signifies that we pick (K_1, K_2) that minimizes $(1 + 2K_1K_2/N) \times MSE(K_1, K_2)$. "CV" signifies that we pick (K_1, K_2) that minimizes the loss function of the J-folder cross validation with $J \in \{5, 10\}$. The choice set of (K_1, K_2) is the nine pairs listed in the table. "CBGPS" is Fong, Hazlett, and Imai's (2018) parametric covariate balancing generalized propensity score estimator. The number of Monte Carlo iterations is $M = 1000$.

Table 3: Simulation results on point estimation of slope β_2 under DGP-L2 ($\beta_2^* = 1$)

		$N = 100$			$N = 500$			$N = 1000$		
	(K_1, K_2)	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	$(2, 3)$	-0.004	0.163	0.163	-0.003	0.075	0.075	0.003	0.053	0.053
GOE	$(2, 6)$	-0.019	0.178	0.179	-0.013	0.078	0.079	-0.010	0.056	0.057
GOE	$(2, 10)$	-0.036	0.196	0.199	-0.031	0.076	0.082	-0.030	0.056	0.064
GOE	$(3, 3)$	-0.005	0.178	0.178	-0.004	0.078	0.079	-0.001	0.054	0.054
GOE	$(3, 6)$	-0.038	0.190	0.194	-0.027	0.084	0.088	-0.025	0.060	0.065
GOE	$(3, 10)$	-0.036	0.207	0.210	-0.033	0.082	0.088	-0.028	0.058	0.065
GOE	$(4, 3)$	-0.014	0.188	0.188	-0.006	0.081	0.081	-0.007	0.058	0.058
GOE	$(4, 6)$	-0.037	0.202	0.205	-0.034	0.082	0.089	-0.028	0.058	0.065
GOE	$(4, 10)$	-0.026	0.213	0.215	-0.025	0.083	0.086	-0.027	0.058	0.065
GOE	MSE (none)	-0.028	0.162	0.165	-0.019	0.072	0.075	-0.014	0.052	0.054
GOE	MSE (add)	-0.009	0.160	0.161	-0.014	0.072	0.073	-0.010	0.052	0.053
GOE	MSE (multi)	-0.006	0.163	0.163	-0.002	0.073	0.073	-0.006	0.052	0.052
GOE	CV ($J = 5$)	0.003	0.161	0.161	0.001	0.075	0.075	0.001	0.052	0.052
GOE	CV ($J = 10$)	0.003	0.164	0.164	-0.001	0.071	0.071	-0.002	0.053	0.053
CBGPS	-	-0.003	0.114	0.114	-0.001	0.050	0.050	-0.001	0.036	0.036

DGP-L2: $T = 1 + 0.2\sum_{j=1}^{2} X_j + \xi$ and $Y = 1 + (1/2)\sum_{j=1}^{2} X_j + T + \epsilon$, where $X_1, X_2 \overset{i.i.d.}{\sim} N(0, 1)$. "GOE" is the proposed generalized optimization estimator. K_1 and K_2 are the dimensions of the polynomials of T and $\boldsymbol{X} = (X_1, X_2)^\top$, respectively. "MSE (none)" signifies that we pick (K_1, K_2) that minimizes $MSE(K_1, K_2) = N^{-1}\sum_{i=1}^{N} \hat{\pi}_K(T_i, \boldsymbol{X}_i)(Y_i - \hat{\beta}_1 - \hat{\beta}_2 T_i)^2$. "MSE (add)" signifies that we pick (K_1, K_2) that minimizes $(1 + 2(K_1 + K_2)/N) \times MSE(K_1, K_2)$. "MSE (multi)" signifies that we pick (K_1, K_2) that minimizes $(1 + 2K_1K_2/N) \times MSE(K_1, K_2)$. "CV" signifies that we pick (K_1, K_2) that minimizes the loss function of the J-folder cross validation with $J \in \{5, 10\}$. The choice set of (K_1, K_2) is the nine pairs listed in the table. "CBGPS" is Fong, Hazlett, and Imai's (2018) parametric covariate balancing generalized propensity score estimator. The number of Monte Carlo iterations is $M = 1000$.

Table 4: Simulation results on point estimation of slope β_2 under DGP-NL2 ($\beta_2^* = 1$)

		$N = 100$			$N = 500$			$N = 1000$		
	(K_1, K_2)	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	$(2, 3)$	-0.037	0.125	0.130	-0.036	0.054	0.065	-0.036	0.037	0.052
GOE	$(2, 6)$	-0.008	0.141	0.141	0.005	0.062	0.062	0.007	0.045	0.045
GOE	$(2, 10)$	-0.022	0.136	0.138	-0.007	0.059	0.060	-0.007	0.043	0.044
GOE	$(3, 3)$	-0.045	0.123	0.131	-0.036	0.052	0.063	-0.037	0.037	0.052
GOE	$(3, 6)$	-0.031	0.132	0.135	-0.012	0.060	0.061	-0.005	0.045	0.045
GOE	$(3, 10)$	-0.031	0.147	0.151	-0.016	0.061	0.063	-0.014	0.043	0.045
GOE	$(4, 3)$	-0.049	0.128	0.137	-0.039	0.055	0.068	-0.037	0.038	0.053
GOE	$(4, 6)$	-0.032	0.148	0.151	-0.014	0.060	0.061	-0.009	0.044	0.045
GOE	$(4, 10)$	-0.046	0.155	0.162	-0.016	0.059	0.061	-0.016	0.044	0.047
GOE	MSE (none)	-0.056	0.134	0.146	-0.023	0.057	0.061	-0.018	0.042	0.045
GOE	MSE (add)	-0.044	0.128	0.136	-0.022	0.056	0.060	-0.021	0.041	0.047
GOE	MSE (multi)	-0.048	0.121	0.130	-0.022	0.054	0.058	-0.017	0.040	0.043
GOE	CV ($J = 5$)	-0.027	0.123	0.125	-0.013	0.056	0.058	-0.007	0.043	0.044
GOE	CV ($J = 10$)	-0.030	0.125	0.129	-0.013	0.058	0.059	-0.009	0.044	0.044
CBGPS	-	0.168	0.139	0.218	0.177	0.058	0.186	0.183	0.041	0.188

DGP-NL2: $T = 0.1(\sum_{j=1}^{2} X_j)^2 + \xi$ and $Y = 1/2 + [(1/2)\sum_{j=1}^{2} X_j]^2 + T + \epsilon$, where $X_1, X_2 \overset{i.i.d.}{\sim} N(0, 1)$. "GOE" is the proposed generalized optimization estimator. K_1 and K_2 are the dimensions of the polynomials of T and $\boldsymbol{X} = (X_1, X_2)^\top$, respectively. "MSE (none)" signifies that we pick (K_1, K_2) that minimizes $MSE(K_1, K_2) = N^{-1}\sum_{i=1}^{N} \hat{\pi}_K(T_i, \boldsymbol{X}_i)(Y_i - \hat{\beta}_1 - \hat{\beta}_2 T_i)^2$. "MSE (add)" signifies that we pick (K_1, K_2) that minimizes $(1 + 2(K_1 + K_2)/N) \times MSE(K_1, K_2)$. "MSE (multi)" signifies that we pick (K_1, K_2) that minimizes $(1 + 2K_1K_2/N) \times MSE(K_1, K_2)$. "CV" signifies that we pick (K_1, K_2) that minimizes the loss function of the J-folder cross validation with $J \in \{5, 10\}$. The choice set of (K_1, K_2) is the nine pairs listed in the table. "CBGPS" is Fong, Hazlett, and Imai's (2018) parametric covariate balancing generalized propensity score estimator. The number of Monte Carlo iterations is $M = 1000$.

Table 5: Simulation results on variance estimation under DGP-L1

$N = 100$

	V_{11} (truth: 3.142)			V_{12} (truth: -1.097)			V_{22} (truth: 1.097)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	0.172	1.273	1.285	-0.117	0.725	0.735	0.109	0.573	0.584
CBGPS	-1.109	1.298	1.707	0.113	0.421	0.436	-0.124	0.365	0.385

$N = 500$

	V_{11} (truth: 3.142)			V_{12} (truth: -1.097)			V_{22} (truth: 1.097)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	0.025	0.458	0.458	-0.021	0.283	0.284	0.037	0.253	0.256
CBGPS	-1.043	0.318	1.091	0.038	0.212	0.215	-0.037	0.187	0.190

$N = 1000$

	V_{11} (truth: 3.142)			V_{12} (truth: -1.097)			V_{22} (truth: 1.097)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	0.026	0.333	0.334	-0.007	0.201	0.201	0.018	0.164	0.165
CBGPS	-1.013	0.244	1.042	0.013	0.173	0.174	-0.012	0.162	0.162

DGP-L1: $T = 1 + 0.2X_1 + \xi$ and $Y = 1 + X_1 + T + \epsilon$, where $X_1 \sim N(0, 1)$. "GOE" is the proposed generalized optimization estimator. $(K_1, K_2) = (2, 2)$ is used to compute $\hat{\pi}_K(T, X_1)$, which is used to estimate $\boldsymbol{\beta}$. $(K_1', K_2', M_0, K_0) = (3, 3, 3, 5)$ is used to compute $\hat{\pi}_{K'}(T, X_1)$, which is used to estimate V_{eff}. "CBGPS" is Fong, Hazlett, and Imai's (2018) parametric covariate balancing generalized propensity score estimator. We report the bias, standard deviation, and RMSE of each element of the variance estimator $\hat{V}_{eff}$ across $M = 1000$ Monte Carlo samples.

Table 6: Simulation results on variance estimation under DGP-NL1

$N = 100$

	V_{11} (truth: 3.043)			V_{12} (truth: -0.118)			V_{22} (truth: 1.074)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	-0.427	0.812	0.917	0.054	0.582	0.584	0.214	0.615	0.651
CBGPS	-0.273	1.149	1.181	0.381	0.713	0.809	1.644	1.252	2.067

$N = 500$

	V_{11} (truth: 3.043)			V_{12} (truth: -0.118)			V_{22} (truth: 1.074)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	-0.249	0.439	0.505	0.078	0.260	0.271	0.027	0.188	0.190
CBGPS	-0.205	0.338	0.395	0.501	0.387	0.633	2.147	0.885	2.323

$N = 1000$

	V_{11} (truth: 3.043)			V_{12} (truth: -0.118)			V_{22} (truth: 1.074)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	-0.235	0.309	0.389	0.071	0.191	0.204	0.005	0.136	0.136
CBGPS	-0.205	0.229	0.307	0.507	0.268	0.574	2.172	0.660	2.270

DGP-NL1: $T = 0.1X_1^2 + \xi$ and $Y = X_1^2 + T + \epsilon$, where $X_1 \sim N(0,1)$. "GOE" is the proposed generalized optimization estimator. $(K_1, K_2) = (2,3)$ is used to compute $\hat{\pi}_K(T, X_1)$, which is used to estimate $\boldsymbol{\beta}$. $(K_1', K_2', M_0, K_0) = (3,3,3,5)$ is used to compute $\hat{\pi}_{K'}(T, X_1)$, which is used to estimate V_{eff}. "CBGPS" is Fong, Hazlett, and Imai's (2018) parametric covariate balancing generalized propensity score estimator. We report the bias, standard deviation, and RMSE of each element of the variance estimator $\hat{V}_{eff}$ across $M = 1000$ Monte Carlo samples.

Table 7: Simulation results on variance estimation under DGP-L2

$N = 100$

	V_{11} (truth: 2.840)			V_{12} (truth: -1.236)			V_{22} (truth: 1.236)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	-0.098	1.016	1.021	0.037	0.619	0.620	0.002	0.544	0.544
CBGPS	0.091	16.173	16.173	-0.159	7.307	7.308	-0.039	3.464	3.464

$N = 500$

	V_{11} (truth: 2.840)			V_{12} (truth: -1.236)			V_{22} (truth: 1.236)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	-0.096	0.533	0.541	0.035	0.346	0.348	-0.021	0.300	0.301
CBGPS	-0.652	0.429	0.780	0.124	0.296	0.320	-0.122	0.259	0.287

$N = 1000$

	V_{11} (truth: 2.840)			V_{12} (truth: -1.236)			V_{22} (truth: 1.236)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	-0.098	0.429	0.440	0.029	0.285	0.287	-0.012	0.240	0.241
CBGPS	-0.582	0.423	0.720	0.072	0.283	0.292	-0.071	0.271	0.280

DGP-L2: $T = 1 + 0.2\sum_{j=1}^{2} X_j + \xi$ and $Y = 1 + (1/2)\sum_{j=1}^{2} X_j + T + \epsilon$, where $X_1, X_2 \overset{i.i.d.}{\sim} N(0,1)$. "GOE" is the proposed generalized optimization estimator. $(K_1, K_2) = (2, 3)$ is used to compute $\hat{\pi}_K(T, \boldsymbol{X})$, which is used to estimate $\boldsymbol{\beta}$. $(K_1', K_2', M_0, K_0) = (3, 3, 6, 8)$ is used to compute $\hat{\pi}_{K'}(T, \boldsymbol{X})$, which is used to estimate V_{eff}. "CBGPS" is Fong, Hazlett, and Imai's (2018) parametric covariate balancing generalized propensity score estimator. We report the bias, standard deviation, and RMSE of each element of the variance estimator $\hat{V}_{eff}$ across $M = 1000$ Monte Carlo samples.

Table 8: Simulation results on variance estimation under DGP-NL2

$N = 100$

	V_{11} (truth: 1.867)			V_{12} (truth: -0.476)			V_{22} (truth: 1.458)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	-0.104	0.602	0.610	0.309	0.474	0.566	-0.056	0.730	0.732
CBGPS	-0.499	0.284	0.574	0.410	0.272	0.492	-0.104	0.496	0.507

$N = 500$

	V_{11} (truth: 1.867)			V_{12} (truth: -0.476)			V_{22} (truth: 1.458)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	-0.058	0.326	0.331	0.132	0.304	0.331	0.048	0.466	0.468
CBGPS	-0.426	1.294	1.363	0.434	0.175	0.468	0.155	0.643	0.662

$N = 1000$

	V_{11} (truth: 1.867)			V_{12} (truth: -0.476)			V_{22} (truth: 1.458)		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
GOE	-0.015	0.301	0.301	0.085	0.288	0.300	0.094	0.376	0.388
CBGPS	-0.467	0.134	0.486	0.438	0.123	0.455	0.185	0.307	0.358

DGP-NL2: $T = 0.1(\sum_{j=1}^{2} X_j)^2 + \xi$ and $Y = 1/2 + [(1/2)\sum_{j=1}^{2} X_j]^2 + T + \epsilon$, where $X_1, X_2 \overset{i.i.d.}{\sim} N(0,1)$. "GOE" is the proposed generalized optimization estimator. $(K_1, K_2) = (2, 6)$ is used to compute $\hat{\pi}_K(T, \boldsymbol{X})$, which is used to estimate $\boldsymbol{\beta}$. $(K_1', K_2', M_0, K_0) = (2, 10, 3, 4)$ is used to compute $\hat{\pi}_{K'}(T, \boldsymbol{X})$, which is used to estimate V_{eff}. "CBGPS" is Fong, Hazlett, and Imai's (2018) parametric covariate balancing generalized propensity score estimator. We report the bias, standard deviation, and RMSE of each element of the variance estimator $\hat{V}_{eff}$ across $M = 1000$ Monte Carlo samples.

Table 9: Empirical results of Fong, Hazlett, and Imai's (2018) CBGPS approach

	Covariates	Parameter of T_i	Parameter of T_i^2
Case #1	$\boldsymbol{Z}_i = (T_i, T_i^2, 1)^\top$	0.088 (0.456) $[-0.804, 0.981]$	-8.5×10^{-6} (2.3×10^{-5}) $[-5.4 \times 10^{-5}, 3.7 \times 10^{-5}]$
Case #2	$\boldsymbol{Z}_i = (T_i, T_i^2, \boldsymbol{SD}_i^\top)^\top$	1.333 (0.444) $[0.462, 2.204]$	-8.6×10^{-5} (2.0×10^{-5}) $[-1.3 \times 10^{-4}, -4.6 \times 10^{-5}]$
Case #3	$\boldsymbol{Z}_i = (T_i, T_i^2, 1, \boldsymbol{X}_{1i}^\top)^\top$	-0.545 (0.423) $[-1.373, 0.284]$	-2.2×10^{-5} (2.2×10^{-5}) $[-6.6 \times 10^{-5}, 2.1 \times 10^{-5}]$
Case #4	$\boldsymbol{Z}_i = (T_i, T_i^2, \boldsymbol{SD}_i^\top, \boldsymbol{X}_{1i}^\top)^\top$	-0.216 (0.422) $[-1.044, 0.611]$	2.7×10^{-5} (2.1×10^{-5}) $[-1.4 \times 10^{-5}, 6.8 \times 10^{-5}]$

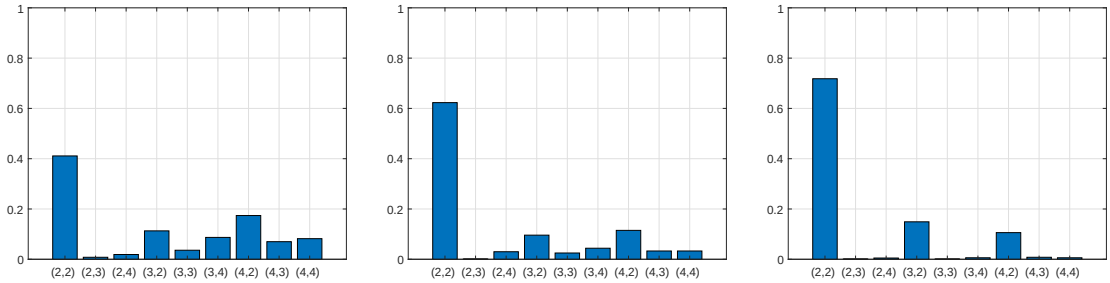
$\boldsymbol{X}_{1i}$ is a vector of eight covariates used in the generalized propensity score model (cf. Eq. (10.1)). $\boldsymbol{SD}_i = (SD_{1i}, SD_{2i}, \ldots, SD_{24i})^\top$, where SD_{ji} is a binary indicator that equals 1 if zip code i belongs to state j and equals 0 otherwise. Any zip code contained in the dataset belongs to one and only one of 24 states. In this table we report the CBGPS estimates for the parameters of T_i and T_i^2 as well as their standard errors in round brackets and 95% confidence bands in square brackets.

Table 10: Empirical results of the generalized optimization approach

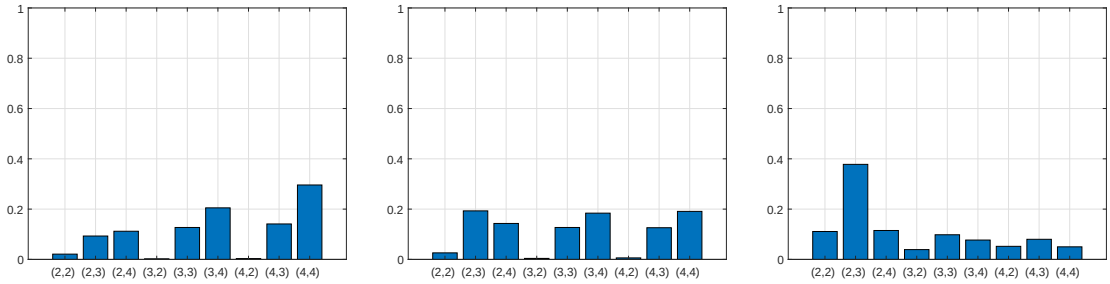
	β_1	β_2	β_3
Point estimate	22.09	-4.7×10^{-4}	1.5×10^{-8}
Standard error	1.214	0.001	4.3×10^{-8}
95% confidence band	$[19.71, 24.47]$	$[-0.002, 0.001]$	$[-7.0 \times 10^{-8}, 1.0 \times 10^{-7}]$

The link function is $g(T, \boldsymbol{\beta}) = \beta_1 + \beta_2 T + \beta_3 T^2$. Covariates $\boldsymbol{X}$ are defined in Eq. (10.2).

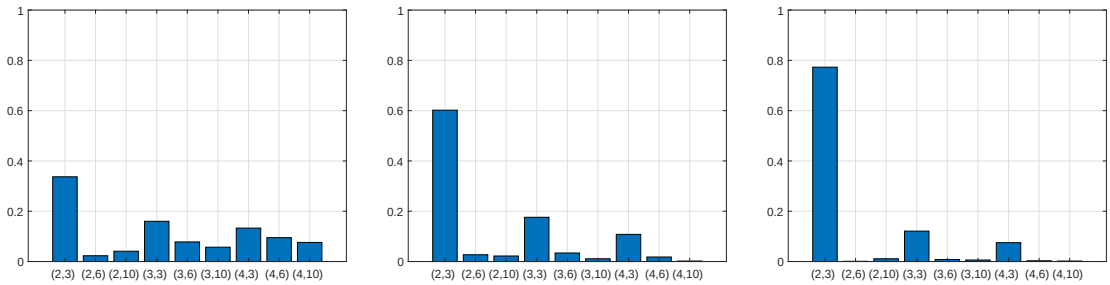
Figure 1: Share of (K_1, K_2) selected via data-driven methods ($N = 1000$)



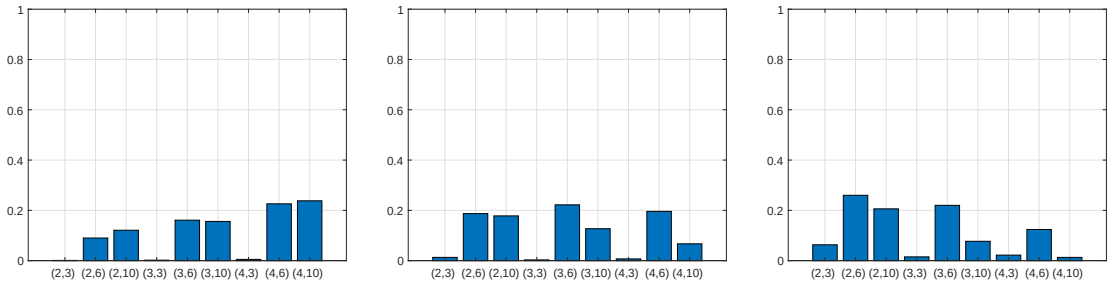
DGP-L1, additive | DGP-L1, multiplicative | DGP-L1, CV5



DGP-NL1, additive | DGP-NL1, multiplicative | DGP-NL1, CV5



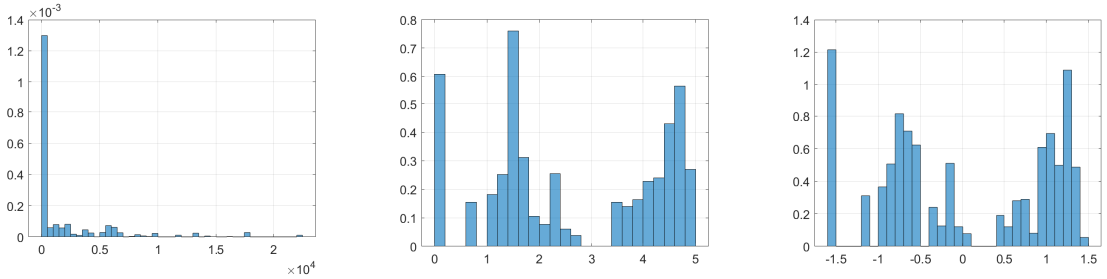
DGP-L2, additive | DGP-L2, multiplicative | DGP-L2, CV5



DGP-NL2, additive | DGP-NL2, multiplicative | DGP-NL2, CV5

Results of the MSE-minimization with additive and multiplicative penalties as well as the 5-folder cross validation are presented. The choice set of (K_1, K_2) is the nine pairs put on the horizontal axis.

Figure 2: Empirical densities of political advertisements



1. Original T 2. Box-Cox 3. Box-Cox & Std.

In this figure, we draw empirical densities of the treatment variable studied in Fong, Hazlett, and Imai (2018) (i.e., the number of political advertisements aired in each zip code). Panel 1 plots the original treatment T; Panel 2 plots T', namely the treatment after running the Box-Cox transformation with $\lambda = -0.16$; Panel 3 plots T^*, namely the standardized version of T'. The vertical axis of each panel is normalized so that each empirical density integrates to 1.

Supplemental Material for "*A Unified Framework for Efficient Estimation of General Treatment Models* "

Chunrong Ai*– University of Florida
Oliver Linton†– University of Cambridge
Kaiji Motegi‡– Kobe University
Zheng Zhang§– Renmin University of China

This draft: November 23, 2019

*Department of Economics, University of Florida. E-mail: tsinghua@ufl.edu
†Faculty of Economics, University of Cambridge. E-mail: obl20@cam.ac.uk
‡Graduate School of Economics, Kobe University. E-mail: motegi@econ.kobe-u.ac.jp
§Institute of Statistics and Big Data, Renmin University of China. E-mail: zhengzhang@ruc.edu.cn

Contents

1 Assumptions **3**

2 Efficiency Bound **4**

2.1 Proof of Theorem 1 4

2.2 Particular Case I: Binary Average Treatment Effects 8

2.3 Particular Case II: Multiple Average Treatment Effects 11

2.4 Particular Case III: Binary Quantile Treatment Effects 13

3 Convergence Rate of Estimated Stabilized Weights **15**

3.1 Lemma 3.1 16

3.2 Lemma 3.2 23

3.3 Corollary 3.3 31

4 Efficient Estimation **33**

4.1 Proof of Theorem 4 33

4.2 Proof of Theorem 5 34

4.3 Proof of (40) 36

5 Some Extensions **46**

5.1 Proof of Theorem 7 46

5.2 Proof of Theorem 9 52

6 Variance Estimation in Monte Carlo Simulations **57**

6.1 Proposed Variance Estimator 57

6.2 True Values of V_{eff} in Monte Carlo Simulations 58

6.2.1 DGP-L1 59

6.2.2 DGP-NL1 60

1 Assumptions

Assumption 1.1 (*Unconfounded Treatment Assignment*) *For all $t \in \mathcal{T}$, given $\boldsymbol{X}$, T is independent of $Y^*(t)$, i.e., $Y^*(t) \perp T|\boldsymbol{X}$, for all $t \in \mathcal{T}$.*

Assumption 1.2 *(i) The support $\mathcal{X}$ of $\boldsymbol{X}$ is a compact subset of $\mathbb{R}^r$. The support $\mathcal{T}$ of the treatment variable T is a compact subset of $\mathbb{R}$. (ii) There exist two positive constants η_1 and η_2 such that*

$$0 < \eta_1 \leq \pi_0(t, \boldsymbol{x}) \leq \eta_2 < \infty \ , \ \forall (t, \boldsymbol{x}) \in \mathcal{T} \times \mathcal{X} \ .$$

Assumption 1.3 *There exist $\Lambda_{K_1 \times K_2} \in \mathbb{R}^{K_1 \times K_2}$ and a positive constant $\alpha > 0$ such that*

$$\sup_{(t,\boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}} \left| \left(\rho'^{-1}\left(\pi_0(t, \boldsymbol{x})\right) - u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right| = O(K^{-\alpha}).$$

Assumption 1.4 *(i) For every K_1 and K_2, the smallest eigenvalues of $\mathbb{E}\left[u_{K_1}(T)u_{K_1}(T)^\top\right]$ and $\mathbb{E}\left[v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\right]$ are bounded away from zero uniformly in K_1 and K_2. (ii) There are two sequences of constants $\zeta_1(K_1)$ and $\zeta_2(K_2)$ satisfying $\sup_{t \in \mathcal{T}} \|u_{K_1}(t)\| \leq \zeta_1(K_1)$ and $\sup_{\boldsymbol{x} \in \mathcal{X}} \|v_{K_2}(\boldsymbol{x})\| \leq \zeta_2(K_2)$, $K = K_1(N)K_2(N)$ and $\zeta(K) := \zeta_1(K_1)\zeta_2(K_2)$, such that $\zeta(K)K^{-\alpha} \to 0$ and $\zeta(K)\sqrt{K/N} \to 0$ as $N \to \infty$.*

Assumption 1.5 *(i) The parameter space $\Theta \subset \mathbb{R}^p$ is a compact set and the true parameter $\boldsymbol{\beta}^*$ is in the interior of Θ , where $p \in \mathbb{N}$. (ii) $L\left(Y - g(T; \boldsymbol{\beta})\right)$ is continuous in $\boldsymbol{\beta}$, $\sup_{\boldsymbol{\beta} \in \Theta} \mathbb{E}[|L(Y - g(T; \boldsymbol{\beta}))|^2] < \infty$ and $\mathbb{E}[\sup_{\boldsymbol{\beta} \in \Theta} |L\left(Y - g(T; \boldsymbol{\beta})\right)|] < \infty$.*

Assumption 1.6

(i) The loss function $L(v)$ is differentiable almost everywhere, $g(t; \boldsymbol{\beta})$ is twice continuously differentiable in $\boldsymbol{\beta} \in \Theta$ and we denote its first derivative by $m(t; \boldsymbol{\beta}) := \nabla_\beta g(t; \boldsymbol{\beta})$;

(ii) $\mathbb{E}\left[\pi_0(T, \boldsymbol{X})L'(Y - g(T; \boldsymbol{\beta}))m(T; \boldsymbol{\beta})\right]$ is differentiable with respect to $\boldsymbol{\beta}$ and $H_0 := -\nabla_\beta \mathbb{E}[\pi_0(T, \boldsymbol{X})$ $L'(Y - g(T; \boldsymbol{\beta}))m(T;\ \boldsymbol{\beta})]\Big|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0}$ is nonsingular;

(iii) $\varepsilon(t, \boldsymbol{x}; \boldsymbol{\beta}_0) := \mathbb{E}[L'(Y - g(T; \boldsymbol{\beta}_0))|T = t, \boldsymbol{X} = \boldsymbol{x}]$ is continuously differentiable in $(t, \boldsymbol{x})$;

(iv) Suppose that $N^{-1}\sum_{i=1}^N \hat{\pi}_K(T_i, \boldsymbol{X}_i)L'\left(Y_i - g(T_i; \hat{\boldsymbol{\beta}})\right) m(T_i; \hat{\boldsymbol{\beta}}) = o_p(N^{-1/2})$ holds with probability approaching one.

Assumption 1.7 *(i)* $\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in\Theta}|L'(Y-g(T;\boldsymbol{\beta}))|^{2+\delta}\right]<\infty$ *for some* $\delta>0$*; (ii) The function class* $\{L'(y-g(t;\boldsymbol{\beta})):\boldsymbol{\beta}\in\Theta\}$ *satisfies:*

$$\mathbb{E}\left[\sup_{\boldsymbol{\beta}_1:\|\boldsymbol{\beta}_1-\boldsymbol{\beta}\|<\delta}\left|L'(Y-g(T;\boldsymbol{\beta}_1))-L'(Y-g(T;\boldsymbol{\beta}))\right|^2\right]^{1/2}\leq a\cdot\delta^b$$

for any $\boldsymbol{\beta}\in\Theta$ *and any small* $\delta>0$ *and for some finite positive constants* a *and* b*.*

Assumption 1.8 $\zeta(K)\sqrt{K^2/N}\to 0$ *and* $\sqrt{N}K^{-\alpha}\to 0$*.*

2 Efficiency Bound

2.1 Proof of Theorem 1

Without loss of generality, we only consider the distribution of $(T,\boldsymbol{X},Y)$ to be absolutely continuous with respect to Lebesgue measure, i.e., there exists a density function $f_{T,X,Y}(t,\boldsymbol{x},y)$ such that $dF_{T,X,Y}(t,\boldsymbol{x},y)=f_{T,X,Y}(t,\boldsymbol{x},y)dtd\boldsymbol{x}dy$. For discrete cases, the proof can be established by using a similar argument.

We follow the approach of Bickel, Klaassen, Ritov, and Wellner (1993, Section 3.3) to derive the variance bound of $\boldsymbol{\beta}^*$, see also Tchetgen Tchetgen and Shpitser (2012). Let $\left\{f^{\alpha}_{Y,T,X}(y,t,\boldsymbol{x})\right\}_{\alpha\in\mathbb{R}}$ denote a one dimensional regular parametric submodel with $f^{\alpha=0}_{Y,T,X}(y,t,\boldsymbol{x})=f_{Y,T,X}(y,t,\boldsymbol{x})$. By definition, $\boldsymbol{\beta}^*$ solves following equation:

$$\int_{\mathcal{T}}\mathbb{E}\left[m(t;\boldsymbol{\beta}^*)L'\left(Y^*(t)-g\left(t;\boldsymbol{\beta}^*\right)\right)\right]f_T(t)dt=0\ . \tag{1}$$

By Assumption 1.1, (1) is equivalent to

$$\int_{\mathcal{T}}\int_{\mathcal{X}}\mathbb{E}\left[m(T;\boldsymbol{\beta}^*)L'\left(Y-g\left(T;\boldsymbol{\beta}^*\right)\right)|T=t,\boldsymbol{X}=\boldsymbol{x}\right]f_X(\boldsymbol{x})f_T(t)d\boldsymbol{x}dt=0\ .$$

Therefore, the parameter $\boldsymbol{\beta}(\alpha)$ induced by the submodel $f^{\alpha}_{Y,T,X}(y,t,\boldsymbol{x})$ satisfies:

$$\int_{\mathcal{T}}\int_{\mathcal{X}}m(t;\boldsymbol{\beta}(\alpha))\cdot\mathbb{E}^{\alpha}\left[L'\left(Y-g\left(t;\boldsymbol{\beta}(\alpha)\right)\right)|T=t,\boldsymbol{X}=\boldsymbol{x}\right]f^{\alpha}_T(t)f^{\alpha}_X(\boldsymbol{x})d\boldsymbol{x}dt=0\ , \tag{2}$$

where $\mathbb{E}^{\alpha}\left[\cdot|T=t,\boldsymbol{X}=\boldsymbol{x}\right]$ denotes taking expectation with respect to the submodel $f^{\alpha}_{Y|T,X}(\cdot|t,\boldsymbol{x})$.

Differentiating both sides of (2) with respect to α, evaluating at $\alpha=0$ and using the condition

$Y^*(t) \perp T|\boldsymbol{X}$, we can deduce that

$$
\begin{aligned}
0 =& \int_{\mathcal{T}}\int_{\mathcal{X}} \frac{\partial}{\partial\alpha}\Big|_{\alpha=0} \left\{m(t;\boldsymbol{\beta}(\alpha))\mathbb{E}^{\alpha}\left[L'(Y-g(t;\boldsymbol{\beta}(\alpha)))|T=t,\boldsymbol{X}=\boldsymbol{x}\right] f_T^{\alpha}(t) f_X^{\alpha}(\boldsymbol{x})\right\} d\boldsymbol{x}dt \\
=& \int_{\mathcal{T}}\int_{\mathcal{X}} \mathbb{E}\left[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}\right] f_T(t) f_X(\boldsymbol{x}) \nabla_{\boldsymbol{\beta}} m(t;\boldsymbol{\beta}^*) d\boldsymbol{x}dt \cdot \frac{\partial}{\partial\alpha}\Big|_{\alpha=0} \boldsymbol{\beta}(\alpha) \\
&+ \int_{\mathcal{X}\times\mathcal{T}} \mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}] m(t;\boldsymbol{\beta}^*) \cdot \frac{\partial}{\partial\alpha} f_X^{\alpha}(\boldsymbol{x})\Big|_{\alpha=0} f_T(t) d\boldsymbol{x}dt \\
&+ \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{T}} m(t;\boldsymbol{\beta}^*) L'(y-g(t;\boldsymbol{\beta}^*)) \cdot \frac{\partial}{\partial\alpha} f_{Y|T,X}^{\alpha}(y|t,\boldsymbol{x})\Big|_{\alpha=0} f_X(\boldsymbol{x}) f_T(t) dy d\boldsymbol{x}dt \\
&+ \int_{\mathcal{X}\times\mathcal{T}} m(t;\boldsymbol{\beta}^*) \cdot \nabla_{\beta}\mathbb{E}[L'(Y^*(t)-g(t;\boldsymbol{\beta}))|T=t,\boldsymbol{X}=\boldsymbol{x}]\Big|_{\beta=\boldsymbol{\beta}^*} \cdot \frac{\partial}{\partial\alpha}\Big|_{\alpha=0} \boldsymbol{\beta}(\alpha) \cdot f_T(t) f_X(\boldsymbol{x}) d\boldsymbol{x}dt \\
&+ \int_{\mathcal{X}\times\mathcal{T}} \mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}] m(t;\boldsymbol{\beta}^*) \cdot \frac{\partial}{\partial\alpha} f_T^{\alpha}(t)\Big|_{\alpha=0} f_X(\boldsymbol{x}) d\boldsymbol{x}dt \\
=& \int_{\mathcal{T}}\int_{\mathcal{X}} \mathbb{E}\left[L'(Y^*(t)-g(t;\boldsymbol{\beta}^*))|\boldsymbol{X}=\boldsymbol{x}\right] f_T(t) f_X(\boldsymbol{x}) \nabla_{\boldsymbol{\beta}} m(t;\boldsymbol{\beta}^*) d\boldsymbol{x}dt \cdot \frac{\partial}{\partial\alpha}\Big|_{\alpha=0} \boldsymbol{\beta}(\alpha) \\
&+ \int_{\mathcal{X}\times\mathcal{T}} \mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}] m(t;\boldsymbol{\beta}^*) \cdot \frac{\partial}{\partial\alpha} f_X^{\alpha}(\boldsymbol{x})\Big|_{\alpha=0} f_T(t) d\boldsymbol{x}dt \\
&+ \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{T}} m(t;\boldsymbol{\beta}^*) L'(y-g(t;\boldsymbol{\beta}^*)) \cdot \frac{\partial}{\partial\alpha} f_{Y|T,X}^{\alpha}(y|t,\boldsymbol{x})\Big|_{\alpha=0} f_X(\boldsymbol{x}) f_T(t) dy d\boldsymbol{x}dt \\
&+ \int_{\mathcal{X}\times\mathcal{T}} m(t;\boldsymbol{\beta}^*) \cdot \nabla_{\beta}\mathbb{E}[L'(Y^*(t)-g(t;\boldsymbol{\beta}))|\boldsymbol{X}=\boldsymbol{x}]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \cdot f_T(t) f_X(\boldsymbol{x}) d\boldsymbol{x}dt \cdot \frac{\partial}{\partial\alpha}\Big|_{\alpha=0} \boldsymbol{\beta}(\alpha) \\
&+ \int_{\mathcal{X}\times\mathcal{T}} \mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}] m(t;\boldsymbol{\beta}^*) \cdot \frac{\partial}{\partial\alpha} f_T^{\alpha}(t)\Big|_{\alpha=0} f_X(\boldsymbol{x}) d\boldsymbol{x}dt \\
=& \int_{\mathcal{T}} \mathbb{E}[L'(Y^*(t)-g(t;\boldsymbol{\beta}^*))] \cdot f_T(t) \nabla_{\boldsymbol{\beta}} m(t;\boldsymbol{\beta}^*) dt \cdot \frac{\partial}{\partial\alpha}\Big|_{\alpha=0} \boldsymbol{\beta}(\alpha) \\
&+ \int_{\mathcal{X}\times\mathcal{T}} \mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}] m(t;\boldsymbol{\beta}^*) \cdot \frac{\partial}{\partial\alpha} f_X^{\alpha}(\boldsymbol{x})\Big|_{\alpha=0} f_T(t) d\boldsymbol{x}dt \\
&+ \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{T}} m(t;\boldsymbol{\beta}^*) \cdot L'(y-g(t;\boldsymbol{\beta}^*)) \cdot \frac{\partial}{\partial\alpha} f_{Y|T,X}^{\alpha}(y|t,\boldsymbol{x})\Big|_{\alpha=0} f_X(\boldsymbol{x}) f_T(t) dy d\boldsymbol{x}dt \\
&+ \int_{\mathcal{T}} \nabla_{\beta}\mathbb{E}[L'(Y^*(t)-g(t;\boldsymbol{\beta}))]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} m(t;\boldsymbol{\beta}^*) \cdot f_T(t) dt \cdot \frac{\partial}{\partial\alpha}\Big|_{\alpha=0} \boldsymbol{\beta}(\alpha) \\
&+ \int_{\mathcal{X}\times\mathcal{T}} \mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}] m(t;\boldsymbol{\beta}^*) \cdot \frac{\partial}{\partial\alpha} f_T^{\alpha}(t)\Big|_{\alpha=0} f_X(\boldsymbol{x}) d\boldsymbol{x}dt \\
=& \nabla_{\beta}\left\{\int_{\mathcal{T}} \mathbb{E}[L'(Y^*(t)-g(t;\boldsymbol{\beta}))] \cdot m(t;\boldsymbol{\beta}) f_T(t) dt\right\}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \cdot \frac{\partial}{\partial\alpha}\Big|_{\alpha=0} \boldsymbol{\beta}(\alpha)
\end{aligned}
$$

$$+\int_{\mathcal{X}\times\mathcal{T}}\mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}]m(t;\boldsymbol{\beta}^*)\cdot\frac{\partial}{\partial\alpha}f_X^\alpha(\boldsymbol{x})\Big|_{\alpha=0}f_T(t)d\boldsymbol{x}dt$$
$$+\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{T}}m(t;\boldsymbol{\beta}^*)\cdot L'(y-g(t;\boldsymbol{\beta}^*))\cdot\frac{\partial}{\partial\alpha}f_{Y|T,X}^\alpha(y|t,\boldsymbol{x})\Big|_{\alpha=0}f_X(\boldsymbol{x})f_T(t)dyd\boldsymbol{x}dt$$
$$+\int_{\mathcal{X}\times\mathcal{T}}\mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}]m(t;\boldsymbol{\beta}^*)\cdot\frac{\partial}{\partial\alpha}f_T^\alpha(t)\Big|_{\alpha=0}f_X(\boldsymbol{x})d\boldsymbol{x}dt.$$

Since $H_0=-\nabla_\beta\left\{\int_{\mathcal{T}}\mathbb{E}[L'(Y^*(t)-g(t;\boldsymbol{\beta}))]\cdot m(t;\boldsymbol{\beta})f_T(t)dt\right\}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ is invertible, we get

$$\begin{aligned}\frac{\partial}{\partial\alpha}\Big|_{\alpha=0}\boldsymbol{\beta}(\alpha)=H_0^{-1}\cdot\Bigg\{&\int_{\mathcal{X}\times\mathcal{T}}\mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}]m(t;\boldsymbol{\beta}^*)\cdot\frac{\partial}{\partial\alpha}f_X^\alpha(\boldsymbol{x})\Big|_{\alpha=0}f_T(t)d\boldsymbol{x}dt\\&+\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{T}}m(t;\boldsymbol{\beta}^*)\cdot L'(y-g(t;\boldsymbol{\beta}^*))\cdot\frac{\partial}{\partial\alpha}f_{Y|T,X}^\alpha(y|t,\boldsymbol{x})\Big|_{\alpha=0}f_X(\boldsymbol{x})f_T(t)dyd\boldsymbol{x}dt\\&+\int_{\mathcal{X}\times\mathcal{T}}\mathbb{E}[L'(Y-g(t;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}]m(t;\boldsymbol{\beta}^*)\cdot\frac{\partial}{\partial\alpha}f_T^\alpha(t)\Big|_{\alpha=0}f_X(\boldsymbol{x})d\boldsymbol{x}dt\Bigg\}.\end{aligned}$$

The efficient influence function of $\boldsymbol{\beta}^*$, denoted by $S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)$, is a unique function satisfying the following equation:

$$\frac{\partial}{\partial\alpha}\Big|_{\alpha=0}\boldsymbol{\beta}(\alpha)=\mathbb{E}\left[S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)\frac{\partial}{\partial\alpha}\Big|_{\alpha=0}\log f_{Y,X,T}^\alpha(Y,\boldsymbol{X},T)\right]. \tag{3}$$

Therefore, to justify our theorem, it suffices to substitute $S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)=H_0^{-1}\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)$ into (3) and check the validity. Note that

$$\begin{aligned}&\mathbb{E}\left[S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)\frac{\partial}{\partial\alpha}\Big|_{\alpha=0}\log f_{Y,X,T}^\alpha(Y,\boldsymbol{X},T)\right]\\=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}\times\mathcal{Y}}\psi(y,t,\boldsymbol{x};\boldsymbol{\beta}^*)\frac{\partial}{\partial\alpha}\Big|_{\alpha=0}f_{Y|X,T}^\alpha(y|\boldsymbol{x},t)f_{T,X}(t,\boldsymbol{x})dyd\boldsymbol{x}dt &(4)\\+&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}\times\mathcal{Y}}\psi(y,t,\boldsymbol{x};\boldsymbol{\beta}^*)f_{Y|X,T}(y|\boldsymbol{x},t)\frac{\partial}{\partial\alpha}\Big|_{\alpha=0}f_{T|X}^\alpha(t|\boldsymbol{x})f_X(\boldsymbol{x})dyd\boldsymbol{x}dt &(5)\\+&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}\times\mathcal{Y}}\psi(y,t,\boldsymbol{x};\boldsymbol{\beta}^*)f_{Y|X,T}(y|\boldsymbol{x},t)f_{T|X}(t|\boldsymbol{x})\frac{\partial}{\partial\alpha}\Big|_{\alpha=0}f_X^\alpha(\boldsymbol{x})dyd\boldsymbol{x}dt. &(6)\end{aligned}$$

For the term (4), we have

$$(4)=H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}\times\mathcal{Y}}\left\{\frac{f_T(t)}{f_{T|X}(t|\boldsymbol{x})}m(t;\boldsymbol{\beta}^*)\cdot L'(y-g(t;\boldsymbol{\beta}^*))-\frac{f_T(t)}{f_{T|X}(t|\boldsymbol{x})}m(t;\boldsymbol{\beta}^*)\cdot\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\right.$$

$$
\begin{aligned}
&+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{x}\right]+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|T=t\right]\Big\}\\
&\times\frac{\partial}{\partial\alpha}\Big|_{\alpha=0} f^{\alpha}_{Y|X,T}(y|\boldsymbol{x},t)f_{T,X}(t,\boldsymbol{x})dyd\boldsymbol{x}dt\\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}\times\mathcal{Y}}\frac{f_T(t)}{f_{T|X}(t|\boldsymbol{x})}m(t;\boldsymbol{\beta}^*)\cdot L'(y-g(t;\boldsymbol{\beta}^*))\cdot\frac{\partial}{\partial\alpha}\Big|_{\alpha=0} f^{\alpha}_{Y|X,T}(y|\boldsymbol{x},t)f_{T,X}(t,\boldsymbol{x})dyd\boldsymbol{x}dt\\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}\times\mathcal{Y}}m(t;\boldsymbol{\beta}^*)\cdot L'(y-g(t;\boldsymbol{\beta}^*))\cdot\frac{\partial}{\partial\alpha}\Big|_{\alpha=0} f^{\alpha}_{Y|X,T}(y|\boldsymbol{x},t)f_T(t)f_X(\boldsymbol{x})dyd\boldsymbol{x}dt.
\end{aligned}
$$

For the term (5), we have

$$
\begin{aligned}
(5)=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}\times\mathcal{Y}}\Big\{\frac{f_T(t)}{f_{T|X}(t|\boldsymbol{x})}m(t;\boldsymbol{\beta}^*)\cdot L'(y-g(t;\boldsymbol{\beta}^*))-\frac{f_T(t)}{f_{T|X}(t|\boldsymbol{x})}m(t;\boldsymbol{\beta}^*)\cdot\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\\
&+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{x}\right]+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|T=t\right]\Big\}\\
&\times f_{Y|X,T}(y|\boldsymbol{x},t)\frac{\partial}{\partial\alpha}\Big|_{\alpha=0} f^{\alpha}_{T|X}(t|\boldsymbol{x})f_X(\boldsymbol{x})dyd\boldsymbol{x}dt\\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}}\Big\{\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{x}\right]+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|T=t\right]\Big\}\\
&\cdot\frac{\partial}{\partial\alpha}\Big|_{\alpha=0} f^{\alpha}_{T|X}(t|\boldsymbol{x})f_X(\boldsymbol{x})d\boldsymbol{x}dt\\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}}\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|T=t\right]\cdot\frac{\partial}{\partial\alpha}\Big|_{\alpha=0} f^{\alpha}_{T|X}(t|\boldsymbol{x})f_X(\boldsymbol{x})d\boldsymbol{x}dt\\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}}\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|T=t\right]\cdot\frac{\partial}{\partial\alpha}\Big|_{\alpha=0} f^{\alpha}_{T}(t)dt\\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}}\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\frac{f_T(t)}{f_{T|X}(t|\boldsymbol{x})}m(t;\boldsymbol{\beta}^*)\cdot\frac{\partial}{\partial\alpha}\Big|_{\alpha=0} f^{\alpha}_{T}(t)\cdot f_{X|T}(\boldsymbol{x}|t)d\boldsymbol{x}dt\\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}}\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)m(t;\boldsymbol{\beta}^*)\cdot\frac{\partial}{\partial\alpha}\Big|_{\alpha=0} f^{\alpha}_{T}(t)\cdot f_X(\boldsymbol{x})d\boldsymbol{x}dt,
\end{aligned}
$$

where the first equality holds in accordance with the definition of $\int_{\mathcal{Y}} L'(y-g(t;\boldsymbol{\beta}^*))f_{Y|X,T}(y|\boldsymbol{x},t)dy =: \varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)$.

For the term (6), we have

$$
\begin{aligned}
(6)=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}\times\mathcal{Y}}\Big\{\frac{f_T(t)}{f_{T|X}(t|\boldsymbol{x})}m(t;\boldsymbol{\beta}^*)\cdot L'(y-g(t;\boldsymbol{\beta}^*))-\frac{f_T(t)}{f_{T|X}(t|\boldsymbol{x})}m(t;\boldsymbol{\beta}^*)\cdot\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\\
&+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{x}\right]+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|T=t\right]\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&\times f_{Y|X,T}(y|\boldsymbol{x},t)f_{T|X}(t|\boldsymbol{x})\frac{\partial}{\partial\alpha}\bigg|_{\alpha=0} f_X^{\alpha}(\boldsymbol{x})dyd\boldsymbol{x}dt \\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}}\left\{\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{x}\right]+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|T=t\right]\right\} \\
&\times f_{T|X}(t|\boldsymbol{x})\cdot\frac{\partial}{\partial\alpha}\bigg|_{\alpha=0} f_X^{\alpha}(\boldsymbol{x})d\boldsymbol{x}dt \\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}}\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{x}\right]\cdot f_{T|X}(t|\boldsymbol{x})\cdot\frac{\partial}{\partial\alpha}\bigg|_{\alpha=0} f_X^{\alpha}(\boldsymbol{x})d\boldsymbol{x}dt \\
=&H_0^{-1}\int_{\mathcal{X}}\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{x}\right]\cdot\frac{\partial}{\partial\alpha}\bigg|_{\alpha=0} f_X^{\alpha}(\boldsymbol{x})d\boldsymbol{x} \\
=&H_0^{-1}\int_{\mathcal{X}\times\mathcal{T}}\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)m(t;\boldsymbol{\beta}^*)\cdot f_T(t)\cdot\frac{\partial}{\partial\alpha}\bigg|_{\alpha=0} f_X^{\alpha}(\boldsymbol{x})d\boldsymbol{x}dt.
\end{aligned}
$$

We have proved (3) holds, hence S_{eff} is the efficient influence function of $\boldsymbol{\beta}^*$.

2.2 Particular Case I: Binary Average Treatment Effects

In this section, we show that when $T\in\{0,1\}$, $g(t;\boldsymbol{\beta})=\beta_0+\beta_1\cdot t$ and $L(v)=v^2$, our general efficiency bound derived in Theorem 1 reduces to the well-known efficiency bound for average treatment effects in Robins, Rotnitzky, and Zhao (1994) and Hahn (1998). In accordance with our identification condition, β_0^* and β_1^* are identified by minimizing the following loss function

$$
\sum_{t\in\{0,1\}}\mathbb{E}[(Y^*(t)-\beta_0-\beta_1\cdot t)^2]\cdot\mathbb{P}(T=t).
$$

The solutions are given by

$$
\beta_0^*=\mathbb{E}[Y^*(0)],\ \ \beta_1^*=\mathbb{E}[Y^*(1)-Y^*(0)].
$$

Here β_1^* is the average treatment effects.

Corollary 2.1 *Suppose $T\in\{0,1\}$, $L(v)=v^2$, $g(t;\boldsymbol{\beta})=\beta_0+\beta_1\cdot t$ and the conditions in Theorem 1 hold, the efficient influence functions of β_0^* and β_1^* given by Theorem 1 reduce to*

$$
\begin{aligned}
&S_{eff}(T,\boldsymbol{X},Y;\beta_0^*)=\phi_2(T,\mathbf{X},Y;\beta_0^*),\\
&S_{eff}(T,\boldsymbol{X},Y;\beta_1^*,\beta_0^*)=\phi_2(T,\mathbf{X},Y;\beta_0^*)-\phi_1(T,\mathbf{X},Y;\beta_1^*,\beta_0^*),
\end{aligned}
$$

where

$$\phi_1(T, \boldsymbol{X}, Y; \beta_1^*, \beta_0^*) = \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot Y^*(1) - \left\{ \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})} - 1 \right\} \cdot \mathbb{E}[Y^*(1)|\boldsymbol{X}] - \beta_0^* - \beta_1^*,$$
$$\phi_2(T, \boldsymbol{X}, Y; \beta_0^*) = \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot Y^*(0) - \left\{ \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})} - 1 \right\} \cdot \mathbb{E}[Y^*(0)|\boldsymbol{X}] - \beta_0^*,$$

and they are the same as the efficient influence functions given in Robins, Rotnitzky, and Zhao (1994) and Hahn (1998).

Proof. Using our notation, we have

$$\boldsymbol{\beta}^* = (\beta_0, \beta_1)^\top, \ g(t; \boldsymbol{\beta}^*) = \beta_0^* + \beta_1^* \cdot t, \quad m(t; \boldsymbol{\beta}^*) = \begin{bmatrix} 1 \\ t \end{bmatrix}, \ H_0 = \mathbb{E}\left[m(T; \boldsymbol{\beta}^*) m(T; \boldsymbol{\beta}^*)^\top \right] \ ,$$
$$\varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}^*) = T \cdot \{ \mathbb{E}[Y^*(1) - Y^*(0)|\boldsymbol{X}] - \beta_1^* \} + \mathbb{E}[Y^*(0)|\boldsymbol{X}] - \beta_0^*,$$
$$\pi_0(T, \boldsymbol{X}) = \frac{T \cdot p + (1-T) \cdot q}{T \cdot \mathbb{P}(T=1|\boldsymbol{X}) + T \cdot \mathbb{P}(T=0|\boldsymbol{X})} = \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot p + \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot q,$$

where $p = \mathbb{P}(T=1)$ and $q = \mathbb{P}(T=0)$. In accordance with our Theorem 1, the efficient influence function of (β_0, β_1) is

$$H_0^{-1} \left\{ \pi_0(T, \boldsymbol{X}) m(T; \boldsymbol{\beta}^*) \left\{ Y - \mathbb{E}[Y|\boldsymbol{X}, T] \right\} + \mathbb{E}\left[\varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}^*) \pi_0(T, \boldsymbol{X}) m(T; \boldsymbol{\beta}^*) | \boldsymbol{X} \right] \right\} .$$

With some computation, we have

$$H_0^{-1} = \begin{bmatrix} 1 & p \\ p & p \end{bmatrix}^{-1} = \frac{1}{pq} \cdot \begin{bmatrix} p & -p \\ -p & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{q} & -\frac{1}{q} \\ -\frac{1}{q} & \frac{1}{pq} \end{bmatrix} \ . \tag{7}$$

and

$$\begin{aligned}
&\pi_0(T, \boldsymbol{X}) m(T; \boldsymbol{\beta}^*) \left\{ Y - \mathbb{E}[Y|\boldsymbol{X}, T] \right\} \\
=&\frac{T}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot p \cdot \begin{bmatrix} 1 \\ T \end{bmatrix} \cdot \left\{ Y - T \cdot \mathbb{E}[Y^*(1)|\boldsymbol{X}] - (1-T) \cdot \mathbb{E}[Y^*(0)|\boldsymbol{X}] \right\} \\
&+ \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot q \cdot \begin{bmatrix} 1 \\ T \end{bmatrix} \cdot \left\{ Y - T \cdot \mathbb{E}[Y^*(1)|\boldsymbol{X}] - (1-T) \cdot \mathbb{E}[Y^*(0)|\boldsymbol{X}] \right\} \\
=&\frac{T}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot p \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \left\{ Y^*(1) - \mathbb{E}[Y^*(1)|\boldsymbol{X}] \right\} + \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot q \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \left\{ Y^*(0) - \mathbb{E}[Y^*(0)|\boldsymbol{X}] \right\}
\end{aligned}$$

$$= \begin{bmatrix} \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot \{Y^*(1) - \mathbb{E}[Y^*(1)|\boldsymbol{X}]\} \cdot p + \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot \{Y^*(0) - \mathbb{E}[Y^*(0)|\boldsymbol{X}]\} \cdot q \\ \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot \{Y^*(1) - \mathbb{E}[Y^*(1)|\boldsymbol{X}]\} \cdot p \end{bmatrix} \tag{8}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}^*)\pi_0(T, \boldsymbol{X})m(T; \boldsymbol{\beta}^*)|\boldsymbol{X}\right] \\
=&\mathbb{E}\left[\left(T \cdot \{\mathbb{E}[Y^*(1) - Y^*(0)|\boldsymbol{X}] - \beta_1^*\} + \mathbb{E}[Y^*(0)|\boldsymbol{X}] - \beta_0^*\right) \cdot \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot p \cdot \begin{bmatrix} 1 \\ T \end{bmatrix} \middle| \boldsymbol{X}\right] \\
&+ \mathbb{E}\left[\left(T \cdot \{\mathbb{E}[Y^*(1) - Y^*(0)|\boldsymbol{X}] - \beta_1^*\} + \mathbb{E}[Y^*(0)|\boldsymbol{X}] - \beta_0^*\right) \cdot \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot q \cdot \begin{bmatrix} 1 \\ T \end{bmatrix} \middle| \boldsymbol{X}\right] \\
=&\mathbb{E}\left[\left(\mathbb{E}[Y^*(1)|\boldsymbol{X}] - \beta_1^* - \beta_0^*\right) \cdot \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot p \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \middle| \boldsymbol{X}\right] \\
&+ \mathbb{E}\left[\left(\mathbb{E}[Y^*(0)|\boldsymbol{X}] - \beta_0^*\right) \cdot \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot q \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \middle| \boldsymbol{X}\right] \\
=&\begin{bmatrix} \left(\mathbb{E}[Y^*(1)|\boldsymbol{X}] - \beta_1^* - \beta_0^*\right) \cdot p + \left(\mathbb{E}[Y^*(0)|\boldsymbol{X}] - \beta_1^*\right) \cdot q \\ \left(\mathbb{E}[Y^*(1)|\boldsymbol{X}] - \beta_1^* - \beta_0^*\right) \cdot p \end{bmatrix}.
\end{aligned} \tag{9}$$

Therefore, with (7), (8), and (9) we can obtain that

$$\begin{aligned}
&\pi_0(T, \boldsymbol{X})m(T; \boldsymbol{\beta}^*)\{Y - \mathbb{E}[Y|\boldsymbol{X}, T]\} + \mathbb{E}\left[\varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}^*)\pi_0(T, \boldsymbol{X})m(T; \boldsymbol{\beta}^*)|\boldsymbol{X}\right] \\
=&\begin{pmatrix} p \cdot \phi_1(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*) + q \cdot \phi_2(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*) \\ p \cdot \phi_1(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*) \end{pmatrix},
\end{aligned}$$

and the efficient influence functions of β_0^* and β_1^* are given by

$$\begin{bmatrix} \frac{1}{q} & -\frac{1}{q} \\ -\frac{1}{q} & \frac{1}{pq} \end{bmatrix} \cdot \begin{pmatrix} p \cdot \phi_1(T, \boldsymbol{X}, Y; \boldsymbol{\beta}) + q \cdot \phi_2(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*) \\ p \cdot \phi_1(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*) \end{pmatrix} = \begin{pmatrix} \phi_2(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*) \\ \phi_1(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*) - \phi_2(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*) \end{pmatrix}.$$

■

2.3 Particular Case II: Multiple Average Treatment Effects

In this section, we show that when $T \in \{0, 1, ..., J\}$, $J \in \mathbb{N}$, $g(t; \boldsymbol{\beta}) = \sum_{j=0}^{J} \beta_j \cdot I(t = j)$ and $L(v) = v^2$, our general efficiency bound derived in Theorem 1 reduces to the efficiency bound of multi-level treatment effects given in Cattaneo (2010). In accordance with our proposed identification condition, $\{\beta_j^*\}_{j=0}^J$ are identified by minimizing the following loss function

$$\sum_{j=0}^{J} \mathbb{E}\left[(Y^*(j) - \beta_j)^2\right] \cdot \mathbb{P}(T = j).$$

The solutions are $\beta_j^* = \mathbb{E}[Y^*(j)]$ for $j \in \{0, ..., J\}$.

Corollary 2.2 *Suppose $T \in \{0, 1, ..., J\}$, $J \in \mathbb{N}$, $g(t; \boldsymbol{\beta}) = \sum_{j=0}^{J} \beta_j \cdot I(t = j)$, $L(v) = v^2$, and the conditions in Theorem 1 hold, the efficient influence functions of $\{\beta_j^*\}_{j=0}^J$ given by Theorem 1 reduce to*

$$S_{eff}(T, \boldsymbol{X}, Y; \beta_j^*) = \frac{I(T = j)}{\mathbb{P}(T = j|\boldsymbol{X})} \cdot \{Y^*(j) - \mathbb{E}[Y^*(j)|\boldsymbol{X}]\} + \mathbb{E}[Y^*(j)|X] - \beta_j^*, \ j \in \{0, ..., J\},$$

and they are the same as the efficient influence functions given in Cattaneo (2010).

Proof. Using our notation, we have

$$\boldsymbol{\beta}^* = (\beta_0^*, ..., \beta_J^*)^\top, \ g(t; \boldsymbol{\beta}^*) = \sum_{j=0}^{J} \beta_j \cdot I(t = j), \ m(t; \boldsymbol{\beta}^*) = \begin{bmatrix} I(t = 0) \\ I(t = 1) \\ \vdots \\ I(t = J) \end{bmatrix}, \ H_0 = \mathbb{E}\left[m(T; \boldsymbol{\beta}^*) m(T; \boldsymbol{\beta}^*)^\top\right].$$

Then

$$\begin{aligned} \varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}^*) =& \mathbb{E}[Y|T, X] - g(T; \boldsymbol{\beta}^*) \\ =& \sum_{j=0}^{J} \mathbb{E}[Y^*(j)|X] \cdot I(t = j) - \sum_{j=0}^{J} \beta_j^* \cdot I(T = j) \\ =& \sum_{j=0}^{J} \left(\mathbb{E}[Y^*(j)|X] - \beta_j^*\right) \cdot I(T = j) \end{aligned}$$

and

$$\pi_0(T, X) = \sum_{j=0}^{J} \frac{I(T=j)}{\mathbb{P}(T=j|\boldsymbol{X})} \cdot p_j, \text{ where } p_j = \mathbb{P}(T=j).$$

Then we have

$$H_0^{-1} = \mathbb{E}\left[m(T;\boldsymbol{\beta}^*)m(T;\boldsymbol{\beta}^*)^\top\right]^{-1} = \begin{bmatrix} p_0^{-1} & & & \\ & p_1^{-1} & & \\ & & \cdots & \\ & & & p_J^{-1} \end{bmatrix},$$

and

$$\begin{aligned}
&\pi_0(T, \boldsymbol{X})m(T;\boldsymbol{\beta}^*)\left\{Y - \mathbb{E}[Y|\boldsymbol{X}, T]\right\} \\
&= \left\{\sum_{j=0}^{J} \frac{I(T=j)}{\mathbb{P}(T=j|\boldsymbol{X})} \cdot p_j\right\} \cdot \begin{bmatrix} I(T=0) \\ I(T=1) \\ \vdots \\ I(T=J) \end{bmatrix} \cdot \left\{Y - \sum_{j=0}^{J} I(T=j) \cdot \mathbb{E}[Y^*(j)|\boldsymbol{X}]\right\} \\
&= \begin{bmatrix} I(T=0) \\ I(T=1) \\ \vdots \\ I(T=J) \end{bmatrix} \left\{\sum_{j=0}^{J} \frac{I(T=j)}{\mathbb{P}(T=j|\boldsymbol{X})} \cdot p_j \cdot Y^*(j) - \sum_{j=0}^{J} \frac{I(T=j)}{\mathbb{P}(T=j|\boldsymbol{X})} \cdot p_j \cdot \mathbb{E}[Y^*(j)|\boldsymbol{X}]\right\} \\
&= \begin{bmatrix} \frac{I(T=0)}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot p_0 \cdot \{Y^*(0) - \mathbb{E}[Y^*(0)|\boldsymbol{X}]\} \\ \frac{I(T=1)}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot p_1 \cdot \{Y^*(1) - \mathbb{E}[Y^*(1)|\boldsymbol{X}]\} \\ \vdots \\ \frac{I(T=J)}{\mathbb{P}(T=J|\boldsymbol{X})} \cdot p_J \cdot \{Y^*(j) - \mathbb{E}[Y^*(j)|\boldsymbol{X}]\} \end{bmatrix}
\end{aligned} \tag{10}$$

and

$$\varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}^*)\pi_0(T, \boldsymbol{X})m(T;\boldsymbol{\beta}^*)$$

$$= \left\{ \sum_{j=0}^{J} \left(\mathbb{E}[Y^*(j)|X] - \beta_j^* \right) \cdot I(T=j) \right\} \left\{ \sum_{j=0}^{J} \frac{I(T=j)}{\mathbb{P}(T=j|\boldsymbol{X})} \cdot p_j \right\} \begin{bmatrix} I(T=0) \\ I(T=1) \\ \vdots \\ I(T=J) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{I(T=0)}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot p_0 \cdot \{\mathbb{E}[Y^*(0)|X] - \beta_0^*\} \\ \frac{I(T=1)}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot p_1 \cdot \{\mathbb{E}[Y^*(1)|X] - \beta_1^*\} \\ \vdots \\ \frac{I(T=J)}{\mathbb{P}(T=J|\boldsymbol{X})} \cdot p_J \cdot \{\mathbb{E}[Y^*(j)|X] - \beta_J^*\} \end{bmatrix}$$

and

$$\mathbb{E}\left[\varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}^*)\pi_0(T, \boldsymbol{X})m(T; \boldsymbol{\beta}^*)|\boldsymbol{X}\right] = \begin{bmatrix} p_0 \cdot \{\mathbb{E}[Y^*(0)|X] - \beta_0^*\} \\ p_1 \cdot \{\mathbb{E}[Y^*(1)|X] - \beta_1^*\} \\ \vdots \\ p_J \cdot \{\mathbb{E}[Y^*(j)|X] - \beta_J^*\} \end{bmatrix}. \tag{11}$$

From Theorem 1, the efficient influence function of $\boldsymbol{\beta}^* = (\beta_0^*, ..., \beta_J^*)$ is given by

$$H_0^{-1}\left\{\pi_0(T, \boldsymbol{X})m(T; \boldsymbol{\beta}^*)\left\{Y - \mathbb{E}[Y|\boldsymbol{X}, T]\right\} + \mathbb{E}\left[\varepsilon(T, \boldsymbol{X}; \boldsymbol{\beta}_0)\pi_0(T, \boldsymbol{X})m(T; \boldsymbol{\beta}^*)|\boldsymbol{X}\right]\right\}$$

$$= \begin{bmatrix} \frac{I(T=0)}{\mathbb{P}(T=0|\boldsymbol{X})} \cdot \{Y^*(0) - \mathbb{E}[Y^*(0)|\boldsymbol{X}]\} + \mathbb{E}[Y^*(0)|X] - \beta_0^* \\ \frac{I(T=1)}{\mathbb{P}(T=1|\boldsymbol{X})} \cdot \{Y^*(1) - \mathbb{E}[Y^*(1)|\boldsymbol{X}]\} + \mathbb{E}[Y^*(1)|X] - \beta_1^* \\ \vdots \\ \frac{I(T=J)}{\mathbb{P}(T=J|\boldsymbol{X})} \cdot \{Y^*(j) - \mathbb{E}[Y^*(j)|\boldsymbol{X}]\} + \mathbb{E}[Y^*(j)|X] - \beta_J^* \end{bmatrix},$$

which is the same as the efficient influence function developed in Corollary 1 of Cattaneo (2010). ■

2.4 Particular Case III: Binary Quantile Treatment Effects

In this section, we show that when $T \in \{0, 1\}$ is a binary treatment variable, $L(v) = v(\tau - I(v \leq 0))$ is the check function with $\tau \in (0, 1)$, and $g(t; \boldsymbol{\beta}^*) = \beta_0^* \cdot (1-t) + \beta_1^* \cdot t$, where $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)$, our general efficiency bound derived in Theorem 1 reduces to the efficiency bound of quantile treatment effects given in Firpo (2007). In accordance with our identification condition, β_0^* and

β_1^* are identified by minimizing the following loss function

$$\sum_{j\in\{0,1\}} \mathbb{P}(T=j)\cdot\mathbb{E}\left[(Y^*(j)-\beta_j)\left\{\tau-I(Y^*(j)\leq\beta_j)\right\}\right].$$

The solutions are $\beta_0^*=\inf\{q:\mathbb{P}(Y^*(0)\leq q)\geq\tau\}$ and $\beta_1^*=\inf\{q:\mathbb{P}(Y^*(1)\leq q)\geq\tau\}$, which are the τ^{th} quantiles of potential outcomes.

Corollary 2.3 *Let $T\in\{0,1\}$, $f_{Y^*(1)}$ and $f_{Y^*(0)}$ be the probability densities of the potential outcomes $Y^*(1)$ and $Y^*(0)$ respectively, $g(t;\boldsymbol{\beta}^*)=\beta_0^*\cdot(1-t)+\beta_1^*\cdot t$, $L(v)=v(\tau-I(v\leq 0))$, and the conditions in Theorem 1 hold, then the efficient influence function of $\boldsymbol{\beta}^*$ given by Theorem 1 reduces to*

$$S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)=\begin{bmatrix}\frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}\cdot\left\{\frac{\tau-I(Y^*(0)\leq\beta_0^*)}{f_{Y^*(0)}(\beta_0^*)}\right\}-\left(\frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}-1\right)\cdot\mathbb{E}\left[\frac{\tau-I(Y^*(0)\leq\beta_0^*)}{f_{Y^*(0)}(\beta_0^*)}\Big|\boldsymbol{X}\right]\\ \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}\cdot\left\{\frac{\tau-I(Y^*(1)\leq\beta_1^*)}{f_{Y^*(1)}(\beta_1^*)}\right\}-\left(\frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}-1\right)\cdot\mathbb{E}\left[\frac{\tau-I(Y^*(1)\leq\beta_1^*)}{f_{Y^*(1)}(\beta_1^*)}\Big|\boldsymbol{X}\right]\end{bmatrix},$$

which is the same as the efficient influence function given in Firpo (2007).

Proof. Using our notation, we have

$$\begin{aligned}
&\boldsymbol{\beta}^*=(\beta_0^*,\beta_1^*)^\top,\ g(t;\boldsymbol{\beta}^*)=\beta_0^*\cdot(1-t)+\beta_1^*\cdot t,\ m(t;\boldsymbol{\beta}^*)=\begin{bmatrix}1-t\\ t\end{bmatrix},\\
&L(v)=v(\tau-I(v\leq 0)),\ L'(v)=\tau-I(v\leq 0)\text{ a.s.},\\
&\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)=T\cdot\mathbb{E}[\tau-I(Y^*(1)\leq\beta_1^*)|\boldsymbol{X}]+(1-T)\cdot\mathbb{E}[\tau-I(Y^*(0)\leq\beta_0^*)|\boldsymbol{X}],\\
&\pi_0(T,\boldsymbol{X})=\frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}\cdot p+\frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}\cdot q\ ,\ p=\mathbb{P}(T=1),\ q=\mathbb{P}(T=0).
\end{aligned}$$

Direct computation yields

$$\begin{aligned}
&\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)L'(Y-g(T;\boldsymbol{\beta}^*))\\
=&\left\{\frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}\cdot p+\frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}\cdot q\right\}\cdot\begin{bmatrix}1-T\\ T\end{bmatrix}\cdot\left\{\tau-I(Y\leq\beta_0^*\cdot(1-T)+\beta_1^*\cdot T)\right\}\\
=&\begin{bmatrix}\frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}\cdot q\cdot\{\tau-I(Y^*(0)\leq\beta_0^*)\}\\ \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}\cdot p\cdot\{\tau-I(Y^*(1)\leq\beta_1^*)\}\end{bmatrix}
\end{aligned}$$

and

$$\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*) = \begin{bmatrix} \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}\cdot q\cdot \mathbb{E}\left[\tau - I(Y^*(0)\le \beta_0^*)|\boldsymbol{X}\right] \\ \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}\cdot p\cdot \mathbb{E}\left[\tau - I(Y^*(1)\le \beta_1^*)|\boldsymbol{X}\right] \end{bmatrix}$$

and

$$\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|\boldsymbol{X}\right] = \begin{bmatrix} q\cdot \mathbb{E}\left[\tau - I(Y^*(0)\le \beta_0^*)|\boldsymbol{X}\right] \\ p\cdot \mathbb{E}\left[\tau - I(Y^*(1)\le \beta_1^*)|\boldsymbol{X}\right] \end{bmatrix}$$

and

$$H_0 = \nabla_{\boldsymbol{\beta}}\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)L'(Y-g(T;\boldsymbol{\beta}^*))\right] = \begin{bmatrix} -q\cdot f_{Y^*(0)}(\beta_0^*) & 0 \\ 0 & -p\cdot f_{Y^*(1)}(\beta_1^*) \end{bmatrix}.$$

Therefore, by Theorem 1, the efficient influence function of $\boldsymbol{\beta}^*$ is

$$\begin{aligned}
&S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*) \\
=&H_0^{-1}\cdot\Big\{\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)L'(Y-g(T;\boldsymbol{\beta}^*)) - \pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*) \\
&\qquad + \mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}\right]\Big\} \\
=&\begin{bmatrix} q^{-1}\cdot\frac{1}{f_{Y^*(0)}(\beta_0^*)} & 0 \\ 0 & p^{-1}\cdot\frac{1}{f_{Y^*(1)}(\beta_1^*)} \end{bmatrix} \\
&\times\begin{bmatrix} \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}\cdot q\cdot\{\tau - I(Y^*(0)\le\beta_0^*)\} - q\cdot\left(\frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}-1\right)\cdot\mathbb{E}\left[\tau - I(Y^*(0)\le\beta_0^*)|\boldsymbol{X}\right] \\ \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}\cdot p\cdot\{\tau - I(Y^*(1)\le\beta_1^*)\} - p\cdot\left(\frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}-1\right)\cdot\mathbb{E}\left[\tau - I(Y^*(1)\le\beta_1^*)|\boldsymbol{X}\right] \end{bmatrix} \\
=&\begin{bmatrix} \frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}\cdot\left\{\frac{\tau - I(Y^*(0)\le\beta_0^*)}{f_{Y^*(0)}(\beta_0^*)}\right\} - \left(\frac{1-T}{\mathbb{P}(T=0|\boldsymbol{X})}-1\right)\cdot\mathbb{E}\left[\frac{\tau - I(Y^*(0)\le\beta_0^*)}{f_{Y^*(0)}(\beta_0^*)}\middle|\boldsymbol{X}\right] \\ \frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}\cdot\left\{\frac{\tau - I(Y^*(1)\le\beta_1^*)}{f_{Y^*(1)}(\beta_1^*)}\right\} - \left(\frac{T}{\mathbb{P}(T=1|\boldsymbol{X})}-1\right)\cdot\mathbb{E}\left[\frac{\tau - I(Y^*(1)\le\beta_1^*)}{f_{Y^*(1)}(\beta_1^*)}\middle|\boldsymbol{X}\right] \end{bmatrix},
\end{aligned}$$

which coincides with efficiency bound derived in Firpo (2007). ■

3 Convergence Rate of Estimated Stabilized Weights

In this section, we establish the convergence rate of estimated stabilized weights $\hat{\pi}_K(T,\boldsymbol{X})$. Let $G^*_{K_1\times K_2}$, $\Lambda^*_{K_1\times K_2}$ and $\pi^*_K(t,\boldsymbol{x})$ be the theoretical counterparts of $\hat{G}_{K_1\times K_2}$, $\hat{\Lambda}_{K_1\times K_2}$ and $\hat{\pi}_K(t,\boldsymbol{x})$

respectively:

$$
\begin{aligned}
G^*_{K_1\times K_2}(\Lambda) :=& \mathbb{E}[\hat{G}_{K_1\times K_2}(\Lambda)] = \mathbb{E}\left[\rho\left(u_{K_1}(T)^\top \Lambda v_{K_2}(\boldsymbol{X})\right)\right] - \mathbb{E}[u_{K_1}(T)^\top]\cdot\Lambda\cdot\mathbb{E}[v_{K_2}(\boldsymbol{X})],\\
\Lambda^*_{K_1\times K_2} :=& \arg\max G^*_{K_1\times K_2}(\Lambda),\\
\pi^*_K(t,\boldsymbol{x}) :=& \rho'\left(u_{K_1}(t)^\top \Lambda^*_{K_1\times K_2} v_{K_2}(\boldsymbol{x})\right).
\end{aligned}
$$

Because of Assumption 1.4, without loss of generality, we can assume the sieve bases $u_{K_1}(T)$ and $v_{K_2}(\boldsymbol{X})$ are orthonormalized, i.e.,

$$
\mathbb{E}\left[u_{K_1}(T)u^\top_{K_1}(T)\right] = I_{K_1\times K_1}, \quad \mathbb{E}\left[v_{K_2}(\boldsymbol{X})v^\top_{K_2}(\boldsymbol{X})\right] = I_{K_2\times K_2}. \tag{12}
$$

Let

$$
\zeta_1(K_1) := \sup_{t\in\mathcal{T}}\|u_{K_1}(t)\|\ ,\ \zeta_2(K_2) := \sup_{\boldsymbol{x}\in\mathcal{X}}\|v_{K_2}(\boldsymbol{x})\|\ ,\ K = K_1\cdot K_2\ ,\ \zeta(K) = \zeta_1(K_1)\zeta_2(K_2).
$$

We also recall the following property satisfied by $\pi_0(T,\boldsymbol{X})$: for any integrable functions $u(t)$ and $v(\boldsymbol{X})$,

$$
\mathbb{E}\left[\pi_0(T,\boldsymbol{X})u(T)v(\boldsymbol{X})\right] = \mathbb{E}[u(T)]\cdot\mathbb{E}[v(\boldsymbol{X})]. \tag{13}
$$

3.1 Lemma 3.1

The first lemma states that $\pi^*_K(t,\boldsymbol{x})$ is arbitrarily close to the true stabilized weights $\pi_0(t,\boldsymbol{x})$.

Lemma 3.1 *Under Assumption 1.2-1.4, we have*

$$
\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}|\pi_0(t,\boldsymbol{x}) - \pi^*_K(t,\boldsymbol{x})| = O(\zeta(K)K^{-\alpha}),
$$

and

$$
\mathbb{E}\left[|\pi_0(T,\boldsymbol{X}) - \pi^*_K(T,\boldsymbol{X})|^2\right] = O\left(K^{-2\alpha}\right),
$$

and

$$
\frac{1}{N}\sum_{i=1}^N|\pi_0(T_i,\boldsymbol{X}_i) - \pi^*_K(T_i,\boldsymbol{X}_i)|^2 = O_p\left(K^{-2\alpha}\right).
$$

Proof. By Assumption 1.2, $\pi_0(t,\boldsymbol{x})\in[\eta_1,\eta_2],\ \forall(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}$ and $(\rho')^{-1}$ is strictly decreasing. Define

$$
\overline{\gamma} := \sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}(\rho')^{-1}\left(\pi_0(t,\boldsymbol{x})\right) \le (\rho')^{-1}(\eta_1) \ \text{ and } \ \underline{\gamma} := \inf_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}(\rho')^{-1}\left(\pi_0(t,\boldsymbol{x})\right) \ge (\rho')^{-1}(\eta_2),
$$

which are two finite constants. By Assumptions 1.3, there exist a constant $C > 0$ and a $K_1 \times K_2$ matrix $\Lambda_{K_1 \times K_2} \in \mathbb{R}^{K_1 \times K_2}$ such that

$$\sup_{(t,\boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}} \left| (\rho')^{-1} \left(\pi_0(t, \boldsymbol{x}) \right) - u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right| < C K^{-\alpha},$$

which implies

$$\begin{aligned} u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) &\in \left((\rho')^{-1} \left(\pi_0(t, \boldsymbol{x}) \right) - C K^{-\alpha}, (\rho')^{-1} \left(\pi_0(t, \boldsymbol{x}) \right) + C K^{-\alpha} \right) \\ &\subset \left[\underline{\gamma} - C K^{-\alpha}, \overline{\gamma} + C K^{-\alpha} \right], \ \forall (t, \boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}, \end{aligned} \tag{14}$$

and

$$\begin{aligned} &\rho' \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) + C K^{-\alpha} \right) - \rho' (u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{X})) \\ <& \pi_0(t, \boldsymbol{x}) - \rho' \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right) \\ <& \rho' \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) - C K^{-\alpha} \right) - \rho' (u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x})) \ , \forall (t, \boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}. \end{aligned}$$

Let $\Gamma_1 := [\underline{\gamma} - 1, \overline{\gamma} + 1]$, by Mean Value Theorem, for large enough K, there exist

$$\begin{aligned} \xi_1(t, \boldsymbol{x}) &\in \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}), u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) + C K^{-\alpha} \right) \\ &\subset \left[\underline{\gamma} - C K^{-\alpha}, \overline{\gamma} + 2 C K^{-\alpha} \right] \subset \Gamma_1 \ , \\ \xi_2(t, \boldsymbol{x}) &\in \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) - C K^{-\alpha}, u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right) \\ &\subset \left[\underline{\gamma} - 2 C K^{-\alpha}, \overline{\gamma} + C K^{-\alpha} \right] \subset \Gamma_1, \end{aligned}$$

such that

$$\rho' \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) + C K^{-\alpha} \right) - \rho' \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right) = \rho''(\xi_1(t, x)) C K^{-\alpha} \geq -a_1 C K^{-\alpha}$$

and

$$\rho' \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) - C K^{-\alpha} \right) - \rho' \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right) = -\rho''(\xi_2(t, \boldsymbol{x})) C K^{-\alpha} \leq a_2 C K^{-\alpha},$$

where $-a_1 := \inf_{\gamma \in \Gamma_1} \rho''(\gamma)$ and $a_2 := \sup_{\gamma \in \Gamma_1} (-\rho''(\gamma))$. Let $a := \max\{a_1, a_2\}$, we have

$$\sup_{(t,\boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}} \left| \pi_0(t, \boldsymbol{x}) - \rho' \left(u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right) \right| < a C K^{-\alpha}. \tag{15}$$

For some fixed $C_2 > 0$ (to be chosen later), define

$$\Upsilon_{K_1\times K_2} := \left\{\Lambda \in \mathbb{R}^{K_1\times K_2} : \|\Lambda - \Lambda_{K_1\times K_2}\| \le C_2 K^{-\alpha}\right\}.$$

For sufficiently large K_1 and K_2, we have that $\forall \Lambda \in \Upsilon_{K_1\times K_2}$, $\forall (t,\boldsymbol{x}) \in \mathcal{T}\times\mathcal{X}$,

$$\begin{aligned}&\left|u_{K_1}(t)^\top \Lambda v_{K_2}(\boldsymbol{x}) - u_{K_1}(t)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{x})\right| \\ \le &\|\Lambda - \Lambda_{K_1\times K_2}\| \cdot \sup_{\boldsymbol{x}\in\mathcal{X}} \|v_{K_2}(\boldsymbol{x})\| \cdot \sup_{t\in\mathcal{T}} \|u_{K_1}(t)\| \le C_2 K^{-\alpha}\zeta_1(K_1)\zeta_2(K_2).\end{aligned}$$

Then in light of (14) and Assumption 1.4, for large enough K_1 and K_2, $\forall \Lambda \in \Upsilon_{K_1\times K_2}$ and $\forall (t,\boldsymbol{x}) \in \mathcal{T}\times\mathcal{X}$, we can deduce that

$$\begin{aligned}u_{K_1}(t)^\top \Lambda v_{K_2}(\boldsymbol{x}) \in &\big(u_{K_1}(t)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{x}) - C_2 K^{-\alpha}\zeta_1(K_1)\zeta_2(K_2), \\ &\quad u_{K_1}(t)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{x}) + C_2 K^{-\alpha}\zeta_1(K_1)\zeta_2(K_2)\big) \\ \subset &\left[\underline{\gamma} - CK^{-\alpha} - C_2 K^{-\alpha}\zeta_1(K_1)\zeta_2(K_2),\right. \\ &\quad \left.\overline{\gamma} + CK^{-\alpha} + C_2 K^{-\alpha}\zeta_1(K_1)\zeta_2(K_2)\right] \subset \Gamma_1.\end{aligned} \tag{16}$$

By definition

$$G^*_{K_1\times K_2}(\Lambda) = \mathbb{E}\left[\rho\left(u_{K_1}(T)^\top \Lambda v_{K_2}(\boldsymbol{X})\right)\right] - \mathbb{E}[u_{K_1}(T)]^\top \Lambda \mathbb{E}[v_{K_2}(\boldsymbol{X})],$$

is a strictly concave function of Λ. By (13), the formula $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for matrices A and B, the facts $\mathbb{E}\left[v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\right] = I_{K_2\times K_2}$ and $\mathbb{E}\left[u_{K_1}(T)u_{K_1}(T)^\top\right] = I_{K_1\times K_1}$, we can deduce that

$$\begin{aligned}&\|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\|^2 \\ =&\left\|\mathbb{E}\left[\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right) u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right] - \mathbb{E}[u_{K_1}(T)]\mathbb{E}[v_{K_2}(\boldsymbol{X})]^\top\right\|^2 \\ =&\left\|\mathbb{E}\left[\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right) u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right] - \mathbb{E}[\pi_0(T,\boldsymbol{X}) u_{K_1}(T) v_{K_2}(\boldsymbol{X})]^\top\right\|^2 \quad \text{(by (13))} \\ =&\left\|\mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right) - \pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right]\right\|^2 \\ =&\mathrm{tr}\left\{\mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right) - \pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right]\right. \\ &\left.\times \mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right) - \pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} v_{K_2}(\boldsymbol{X}) u_{K_1}(T)^\top\right]\right\} \\ =&\mathrm{tr}\left\{\mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right) - \pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right]\right. \\ &\left.\cdot \mathbb{E}\left[u_{K_2}(\boldsymbol{X}) u_{K_2}(\boldsymbol{X})^\top\right]\right\}\end{aligned}$$

$$
\begin{aligned}
&\times \mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} v_{K_2}(\boldsymbol{X}) u_{K_1}(T)^\top\right]\cdot \mathbb{E}\left[u_{K_1}(T)u_{K_1}(T)^\top\right]\Bigg\}\\
=&\mathbb{E}\Bigg[\mathrm{tr}\Bigg\{u_{K_1}(T)^\top\cdot \mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right]\cdot \mathbb{E}\left[u_{K_2}(\boldsymbol{X})u_{K_2}(\boldsymbol{X})^\top\right]\\
&\times \mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} v_{K_2}(\boldsymbol{X}) u_{K_1}(T)^\top\right]\cdot u_{K_1}(T)\Bigg\}\Bigg]\\
=&\mathbb{E}\Bigg[\pi_0(T,\boldsymbol{X})\cdot u_{K_1}(T)^\top\cdot \mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right]\cdot u_{K_2}(\boldsymbol{X})\\
&\times \cdot u_{K_2}(\boldsymbol{X})^\top \mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} v_{K_2}(\boldsymbol{X}) u_{K_1}(T)^\top\right]\cdot u_{K_1}(T)\Bigg] \quad \text{(by (13))}\\
=&\mathbb{E}\left[\left|\pi_0(T,\boldsymbol{X})^{\frac{1}{4}} u_{K_1}(T)\mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right]\pi_0(T,\boldsymbol{X})^{\frac{1}{4}} v_{K_2}(\boldsymbol{X})\right|^2\right].
\end{aligned}
\tag{17}
$$

Note that the term in the last expression

$$
\pi_0(T,\boldsymbol{X})^{\frac{1}{4}} u_{K_1}(T)\cdot \mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right]\pi_0(T,\boldsymbol{X})^{\frac{1}{4}} v_{K_2}(\boldsymbol{X})
$$

is the $L^2(dF_{T,X})$-projection of $\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}}$ on the space spanned by $\{\pi_0(T,\boldsymbol{X})^{\frac{1}{4}} u_{K_1}(T)$, $\pi_0(T,\boldsymbol{X})^{\frac{1}{4}} v_{K_2}(\boldsymbol{X})\}$, which implies that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\pi_0(T,\boldsymbol{X})^{\frac{1}{4}} u_{K_1}(T)\mathbb{E}\left[\sqrt{\pi_0(T,\boldsymbol{X})}\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}} u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right]\pi_0(T,\boldsymbol{X})^{\frac{1}{4}} v_{K_2}(\boldsymbol{X})\right|^2\right]\\
\le&\mathbb{E}\left[\left|\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}}\right|^2\right].
\end{aligned}
\tag{18}
$$

Now, with (17), (18), we can obtain that

$$
\begin{aligned}
&\|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\|\\
\le&\mathbb{E}\left[\left|\frac{\left\{\rho'\left(u_{K_1}(T)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)-\pi_0(T,\boldsymbol{X})\right\}}{\sqrt{\pi_0(T,\boldsymbol{X})}}\right|^2\right]^{\frac{1}{2}}\\
\le&\frac{1}{\sqrt{\eta_1}}\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}\left|\rho'\left(u_{K_1}(t)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{x})\right)-\pi_0(t,\boldsymbol{x})\right| \qquad \text{(by Assumption 1.2)}\\
\le&\frac{aC}{\sqrt{\eta_1}}\cdot K^{-\alpha} \qquad \text{(by (15))}.
\end{aligned}
\tag{19}
$$

Note that for any $\Lambda\in\partial\Upsilon_{K_1\times K_2}$, i.e. $\|\Lambda-\Lambda_{K_1\times K_2}\|=C_2K^{-\alpha}$, by Mean Value Theorem and the

fact $\rho''(y) = -\rho'(y)$, we can deduce that

$$
\begin{aligned}
& G^*_{K_1\times K_2}(\Lambda) - G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2}) \\
&= \sum_{j=1}^{K_2}(\lambda_j - \lambda_j^K)^\top \frac{\partial}{\partial \lambda_i} G^*_{K_1\times K_2}(\lambda_1^K, \ldots, \lambda_{K_2}^K) \\
&\quad + \sum_{l=1}^{K_2}\sum_{j=1}^{K_2} \frac{1}{2}(\lambda_j - \lambda_j^K)^\top \frac{\partial^2}{\partial \lambda_i \partial \lambda_l} G^*_{K_1\times K_2}(\bar{\lambda}_1^K, \ldots, \bar{\lambda}_{K_2}^K)(\lambda_l - \lambda_l^K) \\
&\le \|\Lambda - \Lambda_{K_1\times K_2}\| \, \|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\| \\
&+ \frac{1}{2}\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j - \lambda_j^K)^\top \mathbb{E}\left[\rho''\left(u_{K_1}^\top(T)\bar{\Lambda}_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right) u_{K_1}(T)u_{K_1}(T)^\top v_{K_2,j}(\boldsymbol{X}) v_{K_2,l}(\boldsymbol{X})\right](\lambda_l - \lambda_l^K) \\
&= \|\Lambda - \Lambda_{K_1\times K_2}\| \, \|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\| \\
&\quad - \frac{1}{2}\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j - \lambda_j^K)^\top \mathbb{E}\left[\frac{\rho'\left(u_{K_1}^\top(T)\bar{\Lambda}_{K_1\times K_2} v_{K_2}(\boldsymbol{X})\right)}{\pi_0(T,\boldsymbol{X})}\pi_0(T,\boldsymbol{X}) u_{K_1}(T)u_{K_1}(T)^\top v_{K_2,j}(\boldsymbol{X}) v_{K_2,l}(\boldsymbol{X})\right](\lambda_l - \lambda_l^K) \\
&\le \|\Lambda - \Lambda_{K_1\times K_2}\| \, \|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\| \\
&\quad - \frac{a_3}{2\eta_2}\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j - \lambda_j^K)^\top \mathbb{E}\left[\pi_0(T,\boldsymbol{X}) u_{K_1}(T)u_{K_1}(T)^\top v_{K_2,j}(\boldsymbol{X}) v_{K_2,l}(\boldsymbol{X})\right](\lambda_l - \lambda_l^K) \quad (\text{by } a_3 = \inf_{y\in\Gamma_1}\{\rho'(y)\}) \\
&= \|\Lambda - \Lambda_{K_1\times K_2}\| \, \|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\| \\
&\quad - \frac{a_3}{2\eta_2}\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j - \lambda_j^K)^\top \mathbb{E}\left[u_{K_1}(T)u_{K_1}(T)^\top\right]\mathbb{E}\left[v_{K_2,j}(\boldsymbol{X}) v_{K_2,l}(\boldsymbol{X})\right](\lambda_l - \lambda_l^K) \quad (\text{by } (13)) \\
&= \|\Lambda - \Lambda_{K_1\times K_2}\| \, \|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\| - \frac{a_3}{2\eta_2}\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j - \lambda_j^K)^\top \mathbb{E}\left[v_{K_2,j}(\boldsymbol{X}) v_{K_2,l}(\boldsymbol{X})\right](\lambda_l - \lambda_l^K) \\
&= \|\Lambda - \Lambda_{K_1\times K_2}\| \, \|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\| - \frac{a_3}{2\eta_2}\sum_{j=1}^{K_2}(\lambda_j - \lambda_j^K)^\top(\lambda_j - \lambda_j^K) \\
&= \|\Lambda - \Lambda_{K_1\times K_2}\| \, \|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\| - \frac{a_3}{2\eta_2}\|\Lambda - \Lambda_{K_1\times K_2}\|^2 \\
&= \|\Lambda - \Lambda_{K_1\times K_2}\| \left(\|\nabla G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2})\| - \frac{a_3}{2\eta_2}\|\Lambda - \Lambda_{K_1\times K_2}\|\right) \\
&\le \|\Lambda - \Lambda_{K_1\times K_2}\| \left(\frac{aC}{\sqrt{\eta_1}}K^{-\alpha} - \frac{a_3}{2\eta_2}\cdot C_2 K^{-\alpha}\right), \quad (\text{by } (19))
\end{aligned}
$$

where $\bar{\Lambda}_{K_1\times K_2} = (\bar{\lambda}_1^K, ..., \bar{\lambda}_{K_2}^K)$ lies on the line joining $\Lambda = (\lambda_1, ..., \lambda_{K_2})$ and $\Lambda_{K_1\times K_2} = (\lambda_1^K, ..., \lambda_{K_2}^K)$, which implies $u_{K_1}^\top(t)\bar{\Lambda}_{K_1\times K_2} v_{K_2}(\boldsymbol{x}) \in \Gamma_1$ by (16); $a_3 = \inf_{y\in\Gamma_1}\{\rho'(y)\} > 0$ is a finite positive constant; the fourth and fifth equalities follow from $\mathbb{E}\left[u_{K_1}(T)u_{K_1}(T)^\top\right] = I_{K_1\times K_1}$ and

$\mathbb{E}\left[v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\right] = I_{K_2\times K_2}$ respectively. Therefore, by choosing

$$C_2 > \frac{2\eta_2}{a_3}\cdot\frac{aC}{\sqrt{\eta_1}},$$

we can obtain the following conclusion:

$$G^*_{K_1\times K_2}(\Lambda_{K_1\times K_2}) > G^*_{K_1\times K_2}(\Lambda)\ ,\quad \forall \Lambda\in\partial\Upsilon_{K_1\times K_2}\ . \tag{20}$$

Since $G^*_{K_1\times K_2}$ is continuous, (20) implies that there exists a local maximum of $G^*_{K_1\times K_2}$ in the interior of $\Upsilon_{K_1\times K_2}$. Note that $G^*_{K_1\times K_2}$ is strictly concave with a unique global maximum point $\Lambda^*_{K_1\times K_2}$, therefore we can claim that

$$\Lambda^*_{K_1\times K_2}\in\Upsilon^\circ_{K_1\times K_2},\ \text{i.e. } \|\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\| = O(K^{-\alpha})\ . \tag{21}$$

By Mean Value Theorem, (16) and (21), we can deduce that

$$\begin{aligned}
&\left|\rho'\left(u_{K_1}(t)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)-\rho'\left(u_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)\right|\\
=&\left|\rho''(\xi^*(t,\boldsymbol{x}))\right|\left|u_{K_1}(t)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{x})-u_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right|\\
\le&-\rho''(\xi^*(t,\boldsymbol{x}))\times\|\Lambda_{K_1\times K_2}-\Lambda^*_{K_1\times K_2}\|\times\sup_{t\in\mathcal{T}}\|u_{K_1}(t)\|\times\sup_{\boldsymbol{x}\in\mathcal{X}}\|v_{K_2}(\boldsymbol{x})\|\\
\le&a_2C_2K^{-\alpha}\zeta_1(K_1)\zeta_2(K_2)\ ,
\end{aligned}$$

where $a_2=\sup_{\gamma\in\Gamma_1}\{-\rho''(\gamma)\}<\infty$ is a finite positive constant, and $\xi^*(t,\boldsymbol{x})$ lies between the point $u_{K_1}(t)^\top\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})$ and $u_{K_1}(t)^\top\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{x})$ (note (16) implies $\xi^*(t,\boldsymbol{x})\in\Gamma_1$ for all $(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}$ and large enough K). Therefore, using the triangle inequality, and Assumption 1.4, we can have

$$\begin{aligned}
&\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}\left|\pi_0(t,\boldsymbol{x})-\pi^*_K(t,\boldsymbol{x})\right|\\
\le&\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}\left|\pi_0(t,\boldsymbol{x})-\rho'\left(u_{K_1}(t)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)\right|\\
&+\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}\left|\rho'\left(u_{K_1}(t)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)-\rho'\left(u_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)\right|\\
\le&\ aCK^{-\alpha}+a_2C_2K^{-\alpha}\zeta_1(K_1)\zeta_2(K_2)=O\left(K^{-\alpha}\zeta(K)\right),
\end{aligned}$$

where $\zeta(K)=\zeta_1(K_1)\zeta_2(K_2)$.

We next prove $\mathbb{E}\left[\left|\pi_0(T,\boldsymbol{X})-\pi_K^*(T,\boldsymbol{X})\right|^2\right]=O\left(K^{-2\alpha}\right)$. By Assumption 1.4, we can deduce that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\pi_0(T,\boldsymbol{X})-\pi_K^*(T,\boldsymbol{X})\right|^2\right]\\
\leq&2\cdot\mathbb{E}\left[\left|\pi_0(T,\boldsymbol{X})-\rho'\left(u_{K_1}(T)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{X})\right)\right|^2\right]+2\cdot\mathbb{E}\left[\left|\rho'\left(u_{K_1}(T)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X})\right)-\rho'\left(u_{K_1}(T)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{X})\right)\right|^2\right]\\
\leq&2\cdot\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}\left|\pi_0(t,\boldsymbol{x})-\rho'\left(u_{K_1}(t)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)\right|^2+2\sup_{\gamma\in\Gamma_1}|\rho''(\gamma)|^2\cdot\mathbb{E}\left[\left|u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X})\right|^2\right]\\
\leq&O(K^{-2\alpha})+O(1)\cdot\mathbb{E}\left[\left|u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X})\right|^2\right].
\end{aligned}
$$

We next compute the order of $\mathbb{E}\left[\left|u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X})\right|^2\right]$. Note that $\mathbb{E}[u_{K_1}(T)u_{K_1}(T)^\top]=I_{K_1\times K_1}$, $\mathbb{E}[v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top]=I_{K_2\times K_2}$, (13), (21) and Assumption 1.2, we can deduce that

$$
\begin{aligned}
&\mathbb{E}\left[\left|u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X})\right|^2\right]\\
=&\mathbb{E}\left[u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}^\top u_{K_1}(T)\right]\\
=&\mathbb{E}\left[\frac{1}{\pi_0(T,\boldsymbol{X})}\pi_0(T,\boldsymbol{X})u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}^\top u_{K_1}(T)\right]\\
\leq&\frac{1}{\eta_1}\cdot\mathbb{E}\left[\pi_0(T,\boldsymbol{X})u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}^\top u_{K_1}(T)\right]\\
=&\frac{1}{\eta_1}\cdot\int_{\mathcal{T}}u_{K_1}^\top(t)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}\mathbb{E}\left[v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\right]\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}^\top u_{K_1}(t)dF_T(t)\quad\text{(by (13))}\\
=&\frac{1}{\eta_1}\cdot\int_{\mathcal{T}}u_{K_1}^\top(t)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}\cdot\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}^\top u_{K_1}(t)dF_T(t)\\
=&\frac{1}{\eta_1}\cdot\int_{\mathcal{T}}\mathrm{tr}\left(\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}\cdot\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}^\top u_{K_1}(t)u_{K_1}^\top(t)\right)dF_T(t)\\
=&\frac{1}{\eta_1}\cdot\mathrm{tr}\left(\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}\cdot\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}^\top\right)\\
\leq&\frac{1}{\eta_1}\cdot\|\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\|^2=O(K^{-2\alpha}).\quad\text{(by (21))}
\end{aligned}
\tag{22}
$$

Therefore, we can obtain

$$
\mathbb{E}\left[\left|\pi_0(T,\boldsymbol{X})-\pi_K^*(T,\boldsymbol{X})\right|^2\right]=O\left(K^{-2\alpha}\right).
$$

We finally prove $N^{-1}\sum_{i=1}^N\left|\pi_0(T_i,\boldsymbol{X}_i)-\pi_K^*(T_i,\boldsymbol{X}_i)\right|^2=O_p\left(K^{-2\alpha}\right)$. Note that by (22), we can have

$$
\begin{aligned}
&\mathbb{E}\left[\left\{\frac{1}{N}\sum_{i=1}^N\left|u_{K_1}^\top(T_i)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X}_i)\right|^2-\mathbb{E}\left[\left|u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X})\right|^2\right]\right\}^2\right]\\
\leq&\frac{1}{N}\cdot\mathbb{E}\left[\left|u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2}-\Lambda_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X})\right|^4\right]
\end{aligned}
$$

$$\leq \frac{1}{N} \cdot \mathbb{E}\left[\left|u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2} - \Lambda_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{X})\right|^2\right] \cdot \sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} \left|u_{K_1}^\top(t)\left\{\Lambda^*_{K_1\times K_2} - \Lambda_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{x})\right|^2$$
$$\leq \frac{1}{N} \cdot O(K^{-2\alpha}) \cdot \zeta_1(K_1)^2\zeta_2(K_2)^2 \cdot \left\|\Lambda^*_{K_1\times K_2} - \Lambda_{K_1\times K_2}\right\|^2 \leq \frac{1}{N} \cdot \zeta(K)^2 \cdot O(K^{-4\alpha}),$$

then in light of Chebyshev's inequality and Assumption 1.4, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \left|u_{K_1}^\top(T_i)\left\{\Lambda^*_{K_1\times K_2} - \Lambda_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{X}_i)\right|^2 - \mathbb{E}\left[\left|u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2} - \Lambda_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{X})\right|^2\right] \\
=&O_p\left(\frac{\zeta(K)}{\sqrt{N}}K^{-2\alpha}\right) = o_p\left(K^{-2\alpha}\right). \qquad (23)
\end{aligned}$$

With (21), (22), (23), and Assumption 1.2, we can deduce that

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \left|\pi_0(T_i,\boldsymbol{X}_i) - \pi^*_K(T_i,\boldsymbol{X}_i)\right|^2 \\
\leq&\frac{2}{N}\sum_{i=1}^N \left|\pi_0(T_i,\boldsymbol{X}_i) - \rho'\left(u_{K_1}(T_i)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)\right|^2 \\
&+\frac{2}{N}\sum_{i=1}^N \left|\rho'\left(u_{K_1}(T_i)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right) - \rho'\left(u_{K_1}(T_i)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)\right|^2 \\
\leq&2\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} \left|\pi_0(t,\boldsymbol{x}) - \rho'\left(u_{K_1}(t)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)\right|^2 \\
&+\sup_{\gamma\in\Gamma_1}|\rho''(\gamma)|^2 \cdot \frac{2}{N}\sum_{i=1}^N \left|u_{K_1}^\top(T_i)\left\{\Lambda^*_{K_1\times K_2} - \Lambda_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{X}_i)\right|^2 \\
\leq&2\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} \left|\pi_0(t,\boldsymbol{x}) - \rho'\left(u_{K_1}(t)\Lambda_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)\right|^2 \\
&+2\cdot\sup_{\gamma\in\Gamma_1}|\rho''(\gamma)|^2 \cdot \mathbb{E}\left[\left|u_{K_1}^\top(T)\left\{\Lambda^*_{K_1\times K_2} - \Lambda_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{X})\right|^2\right] + o_p\left(K^{-2\alpha}\right) \\
=&O(K^{-2\alpha}) + O(K^{-2\alpha}) + o_p\left(K^{-2\alpha}\right) = O_p\left(K^{-2\alpha}\right). \quad \text{(by (21))}
\end{aligned}$$

■

3.2 Lemma 3.2

Lemma 3.2 *Under Assumption 1.2-1.4, we have*

$$\left\|\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\| = O_p\left(\sqrt{\frac{K}{N}}\right).$$

Proof. Define

$$\hat{S}_N := \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top \pi_0(T_i,\boldsymbol{X}_i)u_{K_1}(T_i)u_{K_1}(T_i)^\top(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X}_i)v_{K_2,l}(\boldsymbol{X}_i),$$

where λ_j and λ_j^* are the j-th column of Λ and $\Lambda^*_{K_1\times K_2}$ respectively. Since $\hat{S}_N$ is symmetric, using (13) and the facts that $\mathbb{E}\left[u_{K_1}(T)u_{K_1}(T)^\top\right]=I_{K_1\times K_1}$ and $\mathbb{E}\left[v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\right]=I_{K_2\times K_2}$, we can have

$$\begin{aligned}
\mathbb{E}\left[\hat{S}_N\right] =& \sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top \mathbb{E}\left[\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top v_{K_2,j}(\boldsymbol{X})v_{K_2,l}(\boldsymbol{X})\right](\lambda_l-\lambda_l^*)\\
=& \sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top \mathbb{E}\left[u_{K_1}(T)u_{K_1}(T)^\top\right]\mathbb{E}[v_{K_2,j}(\boldsymbol{X})v_{K_2,l}(\boldsymbol{X})](\lambda_l-\lambda_l^*)\\
=& \sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top(\lambda_j-\lambda_j^*) = \left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|.
\end{aligned}$$

Then we can further deduce that

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{S}_N-\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|\right|^2\right]\\
=&\mathbb{E}[\hat{S}_N^2]-2\mathbb{E}[\hat{S}_N]\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|+\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^2\\
=&\frac{N}{N^2}\cdot\mathbb{E}\left[\left(\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X})v_{K_2,l}(\boldsymbol{X})\right)^2\right]\\
&+2\cdot\frac{1}{N^2}\cdot\binom{N}{2}\cdot\mathbb{E}\left[\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X})v_{K_2,l}(\boldsymbol{X})\right]^2\\
&-\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^2\\
=&\frac{1}{N}\mathbb{E}\left[\left(\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X})v_{K_2,l}(\boldsymbol{X})\right)^2\right]\\
&+\frac{N(N-1)}{N^2}\cdot\mathbb{E}\left[\hat{S}_N\right]^2-\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^2\\
=&\frac{1}{N}\mathbb{E}\left[\left(\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X})v_{K_2,l}(\boldsymbol{X})\right)^2\right]\\
&-\frac{1}{N}\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^2
\end{aligned}$$

$$<\frac{1}{N}\mathbb{E}\left[\left(\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X})v_{K_2,l}(\boldsymbol{X})\right)^2\right].$$

In light of the fact that

$$0\le y^\top\left\{\pi_0(t,\boldsymbol{x})u_{K_1}(t)u_{K_1}(t)^\top\right\}y\le\eta_2\zeta_1(K_1)^2y^\top y\ ,\ \forall y\in\mathbb{R}^{K_1},\ \forall(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}\ ,$$

we can deduce that

$$\begin{aligned}
&\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j-\lambda_j^*)^\top\left\{\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top\right\}(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X})v_{K_2,l}(\boldsymbol{X})\\
=&\left[\sum_{j=1}^{K_2}v_{K_2,j}(\boldsymbol{X})(\lambda_j-\lambda_j^*)^\top\right]\cdot\left\{\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top\right\}\cdot\left[\sum_{l=1}^{K_2}(\lambda_l-\lambda_l^*)v_{K_2,l}(\boldsymbol{X})\right]\\
\le&\eta_2\cdot\|u_{K_1}(T)\|^2\cdot\left\|\sum_{i=1}^{K_2}(\lambda_i-\lambda_i^*)^\top v_{K_2,i}(\boldsymbol{X})\right\|^2\\
\le&\eta_2\cdot\|u_{K_1}(T)\|^2\cdot\left(\sum_{i=1}^{K_2}\|\lambda_i-\lambda_i^*\|^2\right)\left(\sum_{i=1}^{K_2}v_{K_2,i}(\boldsymbol{X})^2\right)\\
=&\eta_2\cdot\|u_{K_1}(T)\|^2\cdot\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^2\|v_{K_2}(\boldsymbol{X})\|^2.
\end{aligned}$$

Therefore, we can obtain that

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{S}_N-\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|\right|^2\right]\\
\le&\frac{1}{N}\eta_2^2\cdot\mathbb{E}\left[\|u_{K_1}(T)\|^4\cdot\|v_{K_2}(\boldsymbol{X})\|^4\right]\cdot\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^4\\
\le&\frac{1}{N}\eta_2^2\cdot\zeta_1(K_1)^2\cdot\zeta_2(K_2)^2\cdot\mathbb{E}\left[\|u_{K_1}(T)\|^2\cdot\|v_{K_2}(\boldsymbol{X})\|^2\right]\cdot\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^4\\
=&\frac{1}{N}\eta_2^2\cdot\zeta_1(K_1)^2\cdot\zeta_2(K_2)^2\cdot\mathbb{E}\left[\frac{1}{\pi_0(T,\boldsymbol{X})}\cdot\pi_0(T,\boldsymbol{X})\|u_{K_1}(T)\|^2\cdot\|v_{K_2}(\boldsymbol{X})\|^2\right]\cdot\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^4\\
\le&\frac{1}{N}\frac{\eta_2^2}{\eta_1}\cdot\zeta_1(K_1)^2\cdot\zeta_2(K_2)^2\cdot\mathbb{E}\left[\pi_0(T,\boldsymbol{X})\|u_{K_1}(T)\|^2\cdot\|v_{K_2}(\boldsymbol{X})\|^2\right]\cdot\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^4\quad\text{(by Assumption 1.2)}\\
=&\frac{1}{N}\frac{\eta_2^2}{\eta_1}\cdot\zeta_1(K_1)^2\cdot\zeta_2(K_2)^2\cdot\mathbb{E}\left[\|u_{K_1}(T)\|^2\right]\cdot\mathbb{E}\left[\|v_{K_2}(\boldsymbol{X})\|^2\right]\cdot\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^4\quad\text{(by (13))}\\
=&\frac{1}{N}\frac{\eta_2^2}{\eta_1}\cdot\zeta_1(K_1)^2\cdot\zeta_2(K_2)^2\cdot K_1\cdot K_2\cdot\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^4\text{(since }\mathbb{E}[\|u_{K_1}(T)\|^2]=K_1\text{ and }\mathbb{E}[\|v_{K_2}(\boldsymbol{X})\|^2]=K_2)\\
=&\frac{1}{N}\frac{\eta_2^2}{\eta_1}\cdot\zeta(K)^2\cdot K\cdot\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^4\ .\quad\text{(since }\zeta(K)=\zeta_1(K_1)\zeta_2(K_2)\text{ and }K=K_1\cdot K_2)\qquad(24)
\end{aligned}$$

Considering the event set

$$E_N := \left\{ \hat{S}_N > \frac{1}{2} \left\| \Lambda - \Lambda^*_{K_1 \times K_2} \right\|^2, \ \Lambda \neq \Lambda^*_{K_1 \times K_2} \right\} ,$$

by Chebyshev's inequality, (24), and Assumption 1.4 we can get

$$\begin{aligned}
&\mathbb{P}\left(\left| \hat{S}_N - \left\| \Lambda - \Lambda^*_{K_1 \times K_2} \right\|^2 \right| \geq \frac{1}{2} \left\| \Lambda - \Lambda^*_{K_1 \times K_2} \right\|^2, \ \Lambda \neq \Lambda^*_{K_1 \times K_2} \right) \\
&\leq \frac{4\mathbb{E}\left[\left| \hat{S}_N - \left\| \Lambda - \Lambda^*_{K_1 \times K_2} \right\| \right|^2 \right]}{\left\| \Lambda - \Lambda^*_{K_1 \times K_2} \right\|^4} \\
&\leq \frac{4}{N} \frac{\eta_2^2}{\eta_1} \cdot \zeta(K)^2 \cdot K \leq O\left(\frac{\zeta(K)^2 K}{N} \right) = o(1).
\end{aligned} \tag{25}$$

Note that

$$\begin{aligned}
\nabla \hat{G}_{K_1 \times K_2}(\Lambda) = &\frac{1}{N} \sum_{i=1}^{N} \left\{ \rho' \left(u_{K_1}^\top(T_i) \Lambda v_{K_2}(\boldsymbol{X}_i) \right) u_{K_1}(T_i) v_{K_2}^\top(\boldsymbol{X}_i) - u_{K_1}(T_i) \cdot \mathbb{E}[v_{K_2}^\top(\boldsymbol{X})] \right\} \\
&- \mathbb{E}[u_{K_1}(T)] \cdot \left\{ \frac{1}{N} \sum_{l=1}^{N} v_{K_2}^\top(\boldsymbol{X}_l) - \mathbb{E}[v_{K_2}^\top(\boldsymbol{X})] \right\} \\
&- \left\{ \frac{1}{N} \sum_{i=1}^{N} u_{K_1}(T_i) - \mathbb{E}[u_{K_1}(T)] \right\} \left\{ \frac{1}{N} \sum_{l=1}^{N} v_{K_2}^\top(\boldsymbol{X}_l) - \mathbb{E}[v_{K_2}^\top(\boldsymbol{X})] \right\} \\
= &\nabla \widehat{H}_{K_1 \times K_2}(\Lambda) - \left\{ \frac{1}{N} \sum_{i=1}^{N} u_{K_1}(T_i) - \mathbb{E}[u_{K_1}(T)] \right\} \left\{ \frac{1}{N} \sum_{l=1}^{N} v_{K_2}^\top(\boldsymbol{X}_l) - \mathbb{E}[v_{K_2}^\top(\boldsymbol{X})] \right\},
\end{aligned} \tag{26}$$

where

$$\begin{aligned}
\widehat{H}_{K_1 \times K_2}(\Lambda) := &\frac{1}{N} \sum_{i=1}^{N} \left\{ \rho \left(u_{K_1}^\top(T_i) \Lambda v_{K_2}(\boldsymbol{X}_i) \right) - u_{K_1}(T_i)^\top \Lambda \mathbb{E}[v_{K_2}(\boldsymbol{X})] \right\} \\
&- \mathbb{E}[u_{K_1}^\top(T)] \Lambda \left\{ \frac{1}{N} \sum_{l=1}^{N} v_{K_2}(\boldsymbol{X}_l) - \mathbb{E}[v_{K_2}(\boldsymbol{X})] \right\}.
\end{aligned}$$

Since $\Lambda^*_{K_1 \times K_2}$ is a unique maximizer of $G^*_{K_1 \times K_2}(\cdot)$, then for each $j \in \{1, \ldots, K_2\}$,

$$\begin{aligned}
&\frac{\partial}{\partial \lambda_j} G^*_{K_1 \times K_2}(\lambda_1^*, \ldots, \lambda_{K_2}^*) \\
= &\mathbb{E}\left[\rho' \left(u_{K_1}^\top(T) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{X}) \right) u_{K_1}(X) v_{K_2,j}(Y) \right] - \mathbb{E}[u_{K_1}(T)] \mathbb{E}[v_{K_2,j}(\boldsymbol{X})] = 0 .
\end{aligned}$$

Therefore, for large enough K, we can deduce that

$$\mathbb{E}\left[\|\nabla \hat{H}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2})\|^2\right] = \sum_{j=1}^{K_2}\mathbb{E}\left[\left\|\frac{\partial}{\partial\lambda_j}\hat{H}_{K_1\times K_2}(\lambda^*_1,\ldots,\lambda^*_{K_2})\right\|^2\right] \tag{27}$$

$$\begin{aligned}
\leq& 2\sum_{j=1}^{K_2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left\{\rho'\left(u_{K_1}^\top(T_i)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)u_{K_1}(T_i)v_{K_2,j}(\boldsymbol{X}_i) - \mathbb{E}[v_{K_2,j}(\boldsymbol{X})]u_{K_1}(T_i)\right\}\right\|^2\right] \\
&+ 2\sum_{j=1}^{K_2}\mathbb{E}\left[\left\|\mathbb{E}\left[u_{K_1}(T)\right]\cdot\left\{\frac{1}{N}\sum_{l=1}^{N}v_{K_2,j}(\boldsymbol{X}_l) - \mathbb{E}[v_{K_2,j}(\boldsymbol{X})]\right\}\right\|^2\right] \\
\leq& \frac{4}{N}\sum_{j=1}^{K_2}\left\{\mathbb{E}\left[\left\|\rho'\left(u_{K_1}^\top(T)\Lambda^*_{K_1\times K_2}v_{K_2}(Y)\right)u_{K_1}(T)v_{K_2,j}(\boldsymbol{X})\right\|^2\right] + \mathbb{E}[v_{K_2,j}(\boldsymbol{X})^2]\mathbb{E}\left[\|u_{K_1}(T)\|^2\right]\right\} \\
&+ \frac{2}{N}\sum_{j=1}^{K_2}\mathbb{E}\left[v_{K_2,j}(\boldsymbol{X})^2\right]\cdot\mathbb{E}\left[\|u_{K_1}(T)\|^2\right] \\
=& \frac{4}{N}\sum_{j=1}^{K_2}\left\{\mathbb{E}\left[\frac{\left|\rho'\left(u_{K_1}^\top(T)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X})\right)\right|^2}{\pi_0(T,\boldsymbol{X})}\cdot\pi_0(T,\boldsymbol{X})\cdot\|u_{K_1}(T)v_{K_2,j}(\boldsymbol{X})\|^2\right] + \mathbb{E}[v_{K_2,j}(\boldsymbol{X})^2]\mathbb{E}\left[\|u_{K_1}(T)\|^2\right]\right\} \\
&+ \frac{2}{N}\sum_{j=1}^{K_2}\mathbb{E}\left[v_{K_2,j}(\boldsymbol{X})^2\right]\cdot\mathbb{E}\left[\|u_{K_1}(T)\|^2\right] \\
\leq& \frac{4}{N}\sum_{j=1}^{K_2}\left\{\frac{\left(\sup_{\gamma\in\Gamma_1}\rho'(\gamma)\right)^2}{\eta_1}\cdot\mathbb{E}\left[\pi_0(T,\boldsymbol{X})\cdot\|u_{K_1}(T)v_{K_2,j}(\boldsymbol{X})\|^2\right] + \mathbb{E}[v_{K_2,j}(\boldsymbol{X})^2]\mathbb{E}\left[\|u_{K_1}(T)\|^2\right]\right\} \\
&+ \frac{2}{N}\sum_{j=1}^{K_2}\mathbb{E}\left[v_{K_2,j}(\boldsymbol{X})^2\right]\cdot\mathbb{E}\left[\|u_{K_1}(T)\|^2\right] \\
=& \frac{4}{N}\sum_{j=1}^{K_2}\left\{\frac{\left(\sup_{\gamma\in\Gamma_1}\rho'(\gamma)\right)^2}{\eta_1}\cdot\mathbb{E}\left[v_{K_2,j}(\boldsymbol{X})^2\right]\mathbb{E}\left[\|u_{K_1}(T)\|^2\right] + \mathbb{E}[v_{K_2,j}(\boldsymbol{X})^2]\mathbb{E}\left[\|u_{K_1}(T)\|^2\right]\right\} \\
&+ \frac{2}{N}\sum_{j=1}^{K_2}\mathbb{E}\left[v_{K_2,j}(\boldsymbol{X})^2\right]\cdot\mathbb{E}\left[\|u_{K_1}(T)\|^2\right] \\
\leq& \frac{1}{N}\left\{\frac{4}{\eta_1}\left(\sup_{\gamma\in\Gamma_1}\rho'(\gamma)\right)^2 + 4 + 2\right\}\cdot\mathbb{E}\left[\|u_{K_1}(T)\|^2\right]\sum_{j=1}^{K_2}\mathbb{E}\left[v_{K_2,j}(\boldsymbol{X})^2\right] \\
=& \frac{1}{N}\left\{\frac{4}{\eta_1}\left(\sup_{\gamma\in\Gamma_1}\rho'(\gamma)\right)^2 + 6\right\}K_1K_2 \leq C_4^2\frac{K}{N}\,,
\end{aligned}$$

where the last inequality follows by Assumption 1.8 and C_4 is a finite universal constant.

Let $\epsilon > 0$, fix $C_5(\epsilon) > 0$ (to be chosen later) and define

$$\hat{\Upsilon}_{K_1\times K_2}(\epsilon) := \left\{ \Lambda \in \mathbb{R}^{K_1\times K_2} : \|\Lambda - \Lambda^*_{K_1\times K_2}\| \le C_5(\epsilon)C_4\sqrt{\frac{K}{N}} \right\}.$$

For $\forall \Lambda \in \hat{\Upsilon}_{K_1\times K_2}(\epsilon), \forall (t,\boldsymbol{x}) \in \mathcal{T}\times\mathcal{X}$, we can have

$$\begin{aligned}
&\left|u_{K_1}(t)^\top \Lambda v_{K_2}(\boldsymbol{x}) - u_{K_1}(t)^\top \Lambda^*_{K_1\times K_2} v_{K_2}(\boldsymbol{x})\right| \\
\le &\|\Lambda - \Lambda^*_{K_1\times K_2}\| \sup_{t\in\mathcal{T}} \|u_{K_1}(t)\| \sup_{\boldsymbol{x}\in\mathcal{X}} \|v_{K_2}(\boldsymbol{x})\| \le C_5(\epsilon)C_4\sqrt{\frac{K}{N}}\zeta_1(K_1)\zeta_2(K_2),
\end{aligned}$$

thus for large enough N, in accordance with Assumption 1.4 and (14), we have

$$\begin{aligned}
u_{K_1}(t)^\top \Lambda v_{K_2}(\boldsymbol{x}) \in &\left[u_{K_1}(t)^\top \Lambda^*_{K_1\times K_2} v_{K_2}(\boldsymbol{x}) - C_5(\epsilon)C_4\zeta_1(K_1)\zeta_2(K_2)\sqrt{\frac{K}{N}}, \right. \\
&\left. u_{K_1}(t)^\top \Lambda^*_{K_1\times K_2} v_{K_2}(\boldsymbol{x}) + C_5(\epsilon)C_4\zeta_1(K_1)\zeta_2(K_2)\sqrt{\frac{K}{N}} \right] \\
\subset &\left[\underline{\gamma} - CK^{-\alpha} - C_5(\epsilon)C_4\zeta_1(K_1)\zeta_2(K_2)\sqrt{\frac{K}{N}}, \right. \\
&\left. \overline{\gamma} + CK^{-\alpha} + C_5(\epsilon)C_4\zeta_1(K_1)\zeta_2(K_2)\sqrt{\frac{K}{N}} \right] \subset \Gamma_2(\epsilon)\ ,
\end{aligned} \tag{28}$$

where $\Gamma_2(\epsilon) := \left[\underline{\gamma} - 1 - C_5(\epsilon), \overline{\gamma} + 1 + C_5(\epsilon)\right]$ is a compact set and independent of $(t,\boldsymbol{x})$.

For any $\Lambda \in \partial\hat{\Upsilon}_{K_1\times K_2}(\epsilon)$, there exists $\bar{\Lambda}$ on the line joining Λ and $\Lambda^*_{K_1\times K_2}$ such that

$$\begin{aligned}
\hat{G}_{K_1\times K_2}(\Lambda) =&\hat{G}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2}) + \sum_{j=1}^{K_2}(\lambda_j - \lambda^*_j)^\top \frac{\partial}{\partial\lambda_i}\hat{G}_{K_1\times K_2}(\lambda^*_1,\ldots,\lambda^*_{K_2}) \\
&+ \frac{1}{2}\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j - \lambda^*_j)^\top \frac{\partial^2}{\partial\lambda_i\partial\lambda_l}\hat{G}_{K_1\times K_2}(\bar{\lambda}_1,\ldots,\bar{\lambda}_{K_2})(\lambda_l - \lambda^*_l)\ ,
\end{aligned}$$

where $\bar{\lambda}_j$ denotes the j-th column of $\bar{\Lambda}$. For the second order term in above equality, note that $u^\top_{K_1}(t)\bar{\Lambda}v_{K_2}(\boldsymbol{x}) \in \Gamma_2(\epsilon)$ for all $(t,\boldsymbol{x}) \in \mathcal{T}\times\mathcal{X}$, we can further deduce that

$$\sum_{l=1}^{K_2}\sum_{j=1}^{K_2}(\lambda_j - \lambda^*_j)^\top \frac{\partial^2}{\partial\lambda_i\partial\lambda_l}\hat{G}_{K_1\times K_2}(\bar{\lambda}_1,\ldots,\bar{\lambda}_{K_2})(\lambda_l - \lambda^*_l) \tag{29}$$

$$
\begin{aligned}
=&\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_2}\sum_{l=1}^{K_2}(\lambda_j-\lambda_j^*)^\top u_{K_1}(T_i)\rho''\left(u_{K_1}^\top(T_i)\bar{\Lambda}v_{K_2}(\boldsymbol{X}_i)\right)(\lambda_l-\lambda_l^*)^\top u_{K_1}(T_i)v_{K_2,j}(\boldsymbol{X}_i)v_{K_2,l}(\boldsymbol{X}_i)\\
\leq&-\frac{\bar{b}(\epsilon)}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_2}\sum_{l=1}^{K_2}(\lambda_j-\lambda_j^*)^\top u_{K_1}(T_i)u_{K_1}(T_i)^\top(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X}_i)v_{K_2,l}(\boldsymbol{X}_i)\\
=&-\frac{\bar{b}(\epsilon)}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_2}\sum_{l=1}^{K_2}\frac{1}{\pi_0(T_i,\boldsymbol{X}_i)}(\lambda_j-\lambda_j^*)^\top\pi_0(T_i,\boldsymbol{X}_i)u_{K_1}(T_i)u_{K_1}(T_i)^\top(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X}_i)v_{K_2,l}(\boldsymbol{X}_i)\\
\leq&-\frac{\bar{b}(\epsilon)}{N\eta_2}\sum_{i=1}^{N}\sum_{j=1}^{K_2}\sum_{l=1}^{K_2}(\lambda_j-\lambda_j^*)^\top\pi_0(T_i,\boldsymbol{X}_i)u_{K_1}(T_i)u_{K_1}(T_i)^\top(\lambda_l-\lambda_l^*)v_{K_2,j}(\boldsymbol{X}_i)v_{K_2,l}(\boldsymbol{X}_i)\\
=&-\frac{\bar{b}(\epsilon)}{\eta_2}\hat{S}_N\ ,
\end{aligned}
$$

where $-\bar{b}(\epsilon):=\sup_{\gamma\in\Gamma_2(\epsilon)}\rho''(\gamma)<\infty$.

Define the event set

$$
\begin{aligned}
E_N:=\Bigg\{&\hat{S}_N>\frac{1}{2}\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^2,\ \Lambda\neq\Lambda^*_{K_1\times K_2};\\
&\text{and }\left\|\left\{\frac{1}{N}\sum_{i=1}^{N}u_{K_1}(T_i)-\mathbb{E}[u_{K_1}(T)]\right\}\left\{\frac{1}{N}\sum_{l=1}^{N}v_{K_2}^\top(\boldsymbol{X}_l)-\mathbb{E}[v_{K_2}^\top(\boldsymbol{X})]\right\}\right\|\leq\frac{1}{N^{1/4}}\cdot\sqrt{\frac{K}{N}}\Bigg\}.
\end{aligned}
$$

Note that

$$
\left\|\left\{\frac{1}{N}\sum_{i=1}^{N}u_{K_1}(T_i)-\mathbb{E}[u_{K_1}(T)]\right\}\left\{\frac{1}{N}\sum_{l=1}^{N}v_{K_2}^\top(\boldsymbol{X}_l)-\mathbb{E}[v_{K_2}^\top(\boldsymbol{X})]\right\}\right\|=O_P\left(\frac{\sqrt{K}}{N}\right).
$$

By (25), we can deduece that for any $\epsilon>0$, there exists $N_0(\epsilon)\in\mathbb{N}$ such that $N>N_0(\epsilon)$ large enough

$$
\begin{aligned}
&\mathbb{P}\left((E_N)^c\right)<\mathbb{P}\left(\left|\hat{S}_N-\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^2\right|\geq\frac{1}{2}\left\|\Lambda-\Lambda^*_{K_1\times K_2}\right\|^2,\ \Lambda\neq\Lambda^*_{K_1\times K_2}\right)\\
&\quad+\mathbb{P}\left(\left\|\left\{\frac{1}{N}\sum_{i=1}^{N}u_{K_1}(T_i)-\mathbb{E}[u_{K_1}(T)]\right\}\left\{\frac{1}{N}\sum_{l=1}^{N}v_{K_2}^\top(\boldsymbol{X}_l)-\mathbb{E}[v_{K_2}^\top(\boldsymbol{X})]\right\}\right\|>\frac{1}{N^{1/4}}\cdot\sqrt{\frac{K}{N}}\right)\\
<&\frac{\epsilon}{4}+\frac{\epsilon}{4}=\frac{\epsilon}{2}\ .
\end{aligned}
\tag{30}
$$

Therefore, **on the event** E_N, for large enough N, we can deduce that for any $\Lambda \in \partial\hat{\Upsilon}_{K_1\times K_2}(\epsilon)$,

$$
\begin{aligned}
&\hat{G}_{K_1\times K_2}(\Lambda) - \hat{G}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2}) \qquad (31)\\
&= \sum_{j=1}^{K_2} (\lambda_j - \lambda_j^*)^\top \frac{\partial}{\partial \lambda_j} \hat{G}_{K_1\times K_2}(\lambda_1^*, \ldots, \lambda_{K_2}^*)\\
&\quad + \sum_{l=1}^{K_2}\sum_{j=1}^{K_2} \frac{1}{2}(\lambda_j - \lambda_j^*)^\top \frac{\partial^2}{\partial \lambda_j \partial \lambda_l} \hat{G}_{K_1\times K_2}(\bar{\lambda}_1, \ldots, \bar{\lambda}_{K_2})(\lambda_l - \lambda_l^*)\\
&\le \left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\| \left\|\nabla \hat{G}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2})\right\| - \frac{\bar{b}(\epsilon)}{2\eta_2}\hat{S}_N \text{ (by (29))}\\
&\le \left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\| \left\|\nabla \hat{G}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2})\right\| - \frac{\bar{b}(\epsilon)}{4\eta_2}\left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\|^2\\
&\le \left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\| \left(\left\|\nabla \hat{H}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2})\right\| + \frac{1}{N^{1/4}}\cdot\sqrt{\frac{K}{N}} - \frac{\bar{b}(\epsilon)}{4\eta_2}\left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\| \right) \quad \text{(by (26))}\\
&\le \left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\| \left(\left\|\nabla \hat{H}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2})\right\| - \frac{1}{2}\cdot\frac{\bar{b}(\epsilon)}{4\eta_2}\left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\| \right)
\end{aligned}
$$

where the second and the last inequality follow from definition of the event E_N.

Note that for sufficiently large N, by Chebyshev's inequality and (27) we have

$$
\begin{aligned}
&\mathbb{P}\left\{ \left\|\nabla \hat{H}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2})\right\| \ge \frac{\bar{b}(\epsilon)}{8\eta_2}\left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\| \right\} \qquad (32)\\
\le &\frac{64\cdot\eta_2^2}{\bar{b}(\epsilon)^2}\cdot\frac{\mathbb{E}\left[\left\|\nabla \hat{H}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2})\right\|^2\right]}{\left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\|^2} \le \frac{64\eta_2^2}{\bar{b}(\epsilon)^2 C_5^2(\epsilon)} \le \frac{\epsilon}{2}
\end{aligned}
$$

where the last inequality holds by choosing

$$
C_5(\epsilon) \ge \sqrt{\frac{128\cdot\eta_2^2}{\bar{b}(\epsilon)^2\epsilon}}\ .
$$

Therefore, for sufficiently large N, by (30) and (32) we can derive

$$
\begin{aligned}
&\mathbb{P}\left((E_N)^c \text{ or } \left\|\nabla \hat{H}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2})\right\| \ge \frac{\bar{b}(\epsilon)}{8\eta_2}\left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\| \right) \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\\
\Rightarrow &\mathbb{P}\left(E_N \text{ and } \left\|\nabla \hat{H}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2})\right\| < \frac{\bar{b}(\epsilon)}{8\eta_2}\left\|\Lambda - \Lambda^*_{K_1\times K_2}\right\| \right) > 1-\epsilon\ . \qquad (33)
\end{aligned}
$$

With (31) and (33), we can obtain that

$$\mathbb{P}\left\{\hat{G}_{K_1\times K_2}(\Lambda) - \hat{G}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2}) < 0, \quad \forall \Lambda \in \partial\hat{\Upsilon}_{K_1\times K_2}(\epsilon)\right\} \geq 1-\epsilon\ .$$

Note that the event $\left\{\hat{G}_{K_1\times K_2}(\Lambda^*_{K_1\times K_2}) > \hat{G}_{K_1\times K_2}(\Lambda),\ \forall \Lambda \in \partial\hat{\Upsilon}_{K_1\times K_2}(\epsilon)\right\}$ implies that there exists a local maximizer in the interior of $\hat{\Upsilon}_{K_1\times K_2}(\epsilon)$. Since $\hat{G}_{K_1\times K_2}(\cdot)$ is strictly concave and $\hat{\Lambda}_{K_1\times K_2}$ is the unique global maximizer of $\hat{G}_{K_1\times K_2}$, then we get

$$\mathbb{P}\left(\hat{\Lambda}_{K_1\times K_2} \in \hat{\Upsilon}_{K_1\times K_2}(\epsilon)\right) > 1-\epsilon\ , \tag{34}$$

i.e. $\left\|\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\| = O_p\left(\sqrt{\frac{K}{N}}\right)$.

■

3.3 Corollary 3.3

The next corollary states that $\hat{\pi}_K(t,\boldsymbol{x})$ is arbitrarily close to $\pi^*_K(t,\boldsymbol{x})$.

Corollary 3.3 *Under Assumption 1.2-1.4, we have*

$$\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} |\hat{\pi}_K(t,\boldsymbol{x}) - \pi^*_K(t,\boldsymbol{x})|^2 = O_p\left(\zeta(K)\sqrt{\frac{K}{N}}\right),$$

and

$$\int_{\mathcal{T}\times\mathcal{X}} |\hat{\pi}_K(t,\boldsymbol{x}) - \pi^*_K(t,\boldsymbol{x})|^2 dF_{T,X}(t,\boldsymbol{x}) = O_p\left(\frac{K}{N}\right),$$

and

$$\frac{1}{N}\sum_{i=1}^{N} |\hat{\pi}_K(T_i,\boldsymbol{X}_i) - \pi^*_K(T_i,\boldsymbol{X}_i)|^2 = O_p\left(\frac{K}{N}\right).$$

Proof. From the proof of Lemma 3.2, we know the facts $\mathbb{P}\left(\hat{\Lambda}_{K_1\times K_2} \in \hat{\Upsilon}_{K_1\times K_2}(\epsilon)\right) > 1-\epsilon$ and (28). Then for any element $\tilde{\Lambda}_{K_1\times K_2}$ lying on the line joining $\hat{\Lambda}_{K_1\times K_2}$ and $\Lambda^*_{K_1\times K_2}$, we can have that $\mathbb{P}(u_{K_1}(t)^\top\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x}) \in \Gamma_2(\epsilon)$ for all $(t,\boldsymbol{x}) \in \mathcal{T}\times\mathcal{X}) \geq 1-\epsilon$, which implies

$$\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} |\rho''(u_{K_1}(t)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x}))| = O_p(1). \tag{35}$$

Using Mean Value Theorem, Lemma 3.1, and (35), we can obtain that

$$\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} |\hat{\pi}_K(t,\boldsymbol{x}) - \pi^*_K(t,\boldsymbol{x})|$$

$$
\begin{aligned}
&= \sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} |\rho'\left(u_{K_1}(t)\hat{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right) - \rho'\left(u_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)| \\
&\leq \sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} |\rho''(u_{K_1}(t)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x}))| \sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} \left|u_{K_1}(t)\hat{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x}) - u_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right| \\
&\leq O_p(1)\cdot\|\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\|\cdot\sup_{t\in\mathcal{T}}\|u_{K_1}(t)\|\cdot\sup_{\boldsymbol{x}\in\mathcal{X}}\|v_{K_2}(\boldsymbol{x})\| \\
&\leq O_p(1)\cdot O_p\left(\sqrt{\frac{K}{N}}\right)\zeta_1(K_1)\cdot\zeta_2(K_2) = O_p\left(\zeta(K)\sqrt{\frac{K}{N}}\right).
\end{aligned}
$$

Note that by Mean Value Theorem and (35), we can deduce that

$$
\begin{aligned}
&\int_{\mathcal{T}\times\mathcal{X}} |\hat{\pi}_K(t,\boldsymbol{x}) - \pi^*_K(t,\boldsymbol{x})|^2 dF_{T,X}(t,\boldsymbol{x}) \\
\leq &\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} |\rho''(u_{K_1}(t)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x}))|^2 \int_{\mathcal{T}\times\mathcal{X}} \left|u_{K_1}(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})\right|^2 dF_{T,X}(t,\boldsymbol{x}) \\
\leq &O_p(1)\cdot\int_{\mathcal{T}\times\mathcal{X}} \left|u_{K_1}(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})\right|^2 dF_{T,X}(t,\boldsymbol{x}).
\end{aligned}
$$

We estimate $\int_{\mathcal{T}\times\mathcal{X}} \left|u_{K_1}(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})\right|^2 dF_{T,X}(t,\boldsymbol{x})$. Note that $\mathbb{E}[u_{K_1}(T)[u_{K_1}(T)^\top] = I_{K_1\times K_1}$, $\mathbb{E}[v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top] = I_{K_2\times K_2}$, (13) and Assumption 1.2, we can deduce that

$$
\begin{aligned}
&\int_{\mathcal{T}\times\mathcal{X}} \left|u_{K_1}(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})\right|^2 dF_{T,X}(t,\boldsymbol{x}) \\
\leq &\int_{\mathcal{T}\times\mathcal{X}} u_{K_1}^\top(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})v_{K_2}(\boldsymbol{x})^\top\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}^\top u_{K_1}(t)dF_{T,X}(t,\boldsymbol{x}) \\
= &\int_{\mathcal{T}\times\mathcal{X}} \frac{1}{\pi_0(t,\boldsymbol{x})}\pi_0(t,\boldsymbol{x})u_{K_1}^\top(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})v_{K_2}(\boldsymbol{x})^\top\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}^\top u_{K_1}(t)dF_{T,X}(t,\boldsymbol{x}) \\
\leq &\frac{1}{\eta_1}\int_{\mathcal{T}\times\mathcal{X}} \pi_0(t,\boldsymbol{x})\cdot u_{K_1}^\top(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})v_{K_2}(\boldsymbol{x})^\top\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}^\top u_{K_1}(t)dF_{T,X}(t,\boldsymbol{x}) \\
= &\frac{1}{\eta_1}\int_{\mathcal{T}} u_{K_1}^\top(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}\left(\int_{\mathcal{X}} v_{K_2}(\boldsymbol{x})v_{K_2}(\boldsymbol{x})^\top dF_X(x)\right)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}^\top u_{K_1}(t)dF_T(t) \\
= &\frac{1}{\eta_1}\int_{\mathcal{T}} u_{K_1}^\top(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}^\top u_{K_1}(t)dF_T(t) \\
= &\frac{1}{\eta_1}\mathrm{tr}\left(\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}^\top\int_{\mathcal{T}} u_{K_1}(t)u_{K_1}^\top(t)dF_T(t)\right) \\
= &\frac{1}{\eta_1}\mathrm{tr}\left(\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\}^\top\right) \\
= &\frac{1}{\eta_1}\cdot\left\|\hat{\Lambda}_{K_2\times K_2} - \Lambda^*_{K_1\times K_2}\right\|^2 = O_p\left(\frac{K}{N}\right).
\end{aligned} \tag{36}
$$

Then we obtain

$$\int_{\mathcal{T}\times\mathcal{X}} |\hat{\pi}_K(t,\boldsymbol{x}) - \pi_K^*(t,\boldsymbol{x})|^2 dF_{T,X}(t,\boldsymbol{x}) = O_p\left(\frac{K}{N}\right).$$

Similar to (23), we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left|u_{K_1}^\top(T_i)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{X}_i)\right|^2 - \int_{\mathcal{T}\times\mathcal{X}}\left|u_{K_1}(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{x})\right|^2 dF_{T,X}(t,\boldsymbol{x}) \\
=&O_p\left(\frac{\zeta(K)}{\sqrt{N}}\cdot\|\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\|^2\right) = O_p\left(\frac{\zeta(K)}{\sqrt{N}}\cdot\frac{K}{N}\right) = o_p\left(\frac{K}{N}\right). \qquad (37)
\end{aligned}$$

where the last equality holds in light of Assumption 1.4. Hence, with (36) and (37), we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}|\hat{\pi}_K(T_i,\boldsymbol{X}_i) - \pi_K^*(T_i,\boldsymbol{X}_i)|^2 \\
\leq&\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}|\rho''(u_{K_1}(t)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x}))|^2\cdot\frac{1}{N}\sum_{i=1}^{N}\left|u_{K_1}(T_i)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{X}_i)\right|^2 \\
\leq&O_p(1)\cdot\int_{\mathcal{T}\times\mathcal{X}}\left|u_{K_1}(t)\left\{\hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2}\right\} v_{K_2}(\boldsymbol{x})\right|^2 dF_{T,X}(t,\boldsymbol{x}) + o_p\left(\frac{K}{N}\right) \\
\leq&O_p\left(\frac{K}{N}\right) + o_p\left(\frac{K}{N}\right) = O_p\left(\frac{K}{N}\right).
\end{aligned}$$

■

4 Efficient Estimation

4.1 Proof of Theorem 4

Because $\widehat{\boldsymbol{\beta}}$ (resp. $\boldsymbol{\beta}^*$) is a unique minimizer of $N^{-1}\sum_{i=1}^{N}\widehat{\pi}_K(T_i,\boldsymbol{X}_i)L\left(Y_i - g(T_i;\boldsymbol{\beta})\right)$ (resp. $\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L\left(Y - g(T;\boldsymbol{\beta})\right)\right]$), from the theory of M-estimation (van der Vaart, 1998, Theorem 5.7), if the following condition holds:

$$\sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\widehat{\pi}_K(T_i,\boldsymbol{X}_i)L\left(Y_i - g(T_i;\boldsymbol{\beta})\right) - \mathbb{E}\left[\pi_0(T,\boldsymbol{X})L\left(Y - g(T;\boldsymbol{\beta})\right)\right]\right| \xrightarrow{p} 0.$$

then $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$. Note that

$$\sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\widehat{\pi}_K(T_i,\boldsymbol{X}_i)L\left(Y_i - g(T_i;\boldsymbol{\beta})\right) - \mathbb{E}\left[\pi_0(T,\boldsymbol{X})L\left(Y - g(T;\boldsymbol{\beta})\right)\right]\right|$$

$$\leq \sup_{\boldsymbol{\beta}\in\Theta} \left| \frac{1}{N}\sum_{i=1}^{N} \left\{\widehat{\pi}_K(T_i, \boldsymbol{X}_i) - \pi_0(T_i, \boldsymbol{X}_i)\right\} L\left(Y_i - g(T_i;\boldsymbol{\beta})\right) \right| \tag{38}$$

$$+ \sup_{\boldsymbol{\beta}\in\Theta} \left| \frac{1}{N}\sum_{i=1}^{N} \pi_0(T_i, \boldsymbol{X}_i) L\left(Y_i - g(T_i;\boldsymbol{\beta})\right) - \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) L\left(Y - g(T;\boldsymbol{\beta})\right)\right] \right|. \tag{39}$$

We first show (38) is of $o_P(1)$. By Theorem 3, $\widehat{\pi}_K(\cdot) \xrightarrow{L^2(F_N)} \pi_0(\cdot)$, using Cauchy-Scharwz' inequality and Assumption 1.5, we have

$$\begin{aligned}
|(38)| \leq & \left\{ \frac{1}{N}\sum_{i=1}^{N} \left\{\widehat{\pi}_K(T_i, \boldsymbol{X}_i) - \pi_0(T_i, \boldsymbol{X}_i)\right\}^2 \right\}^{1/2} \cdot \sup_{\boldsymbol{\beta}\in\Theta} \left\{ \frac{1}{N}\sum_{i=1}^{N} L\left(Y_i - g(T_i;\boldsymbol{\beta})\right)^2 \right\}^{1/2} \\
\leq & o_P(1) \cdot \left\{ \sup_{\boldsymbol{\beta}\in\Theta} \mathbb{E}\left[L\left(Y - g(T;\boldsymbol{\beta})\right)^2\right] + o_P(1) \right\}^{1/2} = o_P(1) \quad \text{(by Assumption 5)}
\end{aligned}$$

To show (39) is of $o_P(1)$, by (Newey and McFadden, 1994, Lemma 2.4), it is sufficient to require the following conditions holds:

1. Θ is compact;
2. $L(Y - g(T;\boldsymbol{\beta}))$ is continuous in $\boldsymbol{\beta}$;
3. $\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in\Theta} |L(Y - g(T;\boldsymbol{\beta}))|\right] < \infty$.

which are the imposed Assumption 1.5.

4.2 Proof of Theorem 5

The proposed estimator $\hat{\boldsymbol{\beta}}$ is a special case of Chen, Linton, and Van Keilegom (2003), where the authors establish the consistency and asymptotic normality of a class of semiparametric optimization estimators under that the criterion function does not satisfy standard smoothness conditions. The asymptotic distribution of the proposed estimator can be derived by applying Theorem 2 of Chen, Linton, and Van Keilegom (2003).

Using their notation, we denote

$$\begin{aligned}
M_N(\boldsymbol{\beta}, \pi(\cdot)) &:= \frac{1}{N}\sum_{i=1}^{N} \pi(T_i, \boldsymbol{X}_i) L'\left(Y_i - g(T_i;\boldsymbol{\beta})\right) m(T_i;\boldsymbol{\beta}), \\
M(\boldsymbol{\beta}, \pi(\cdot)) &:= \mathbb{E}[M_N(\boldsymbol{\beta}, \pi(\cdot))] = \mathbb{E}\left[\pi(T, \boldsymbol{X}) L'\left(Y - g(T;\boldsymbol{\beta})\right) m(T;\boldsymbol{\beta})\right].
\end{aligned}$$

The ordinary derivative $\Gamma_1(\boldsymbol{\beta}, \pi(\cdot))$ in $\boldsymbol{\beta}$ of $M(\boldsymbol{\beta}, \pi(\cdot))$ is

$$\begin{aligned}\Gamma_1(\boldsymbol{\beta}, \pi(\cdot))(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}) =& : \lim_{\tau \to 0} \frac{M(\boldsymbol{\beta} + \tau(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}), \pi(\cdot)) - M(\boldsymbol{\beta}, \pi(\cdot))}{\tau} \\ =& \nabla_{\boldsymbol{\beta}} \mathbb{E}\left[\pi(T, \boldsymbol{X}) L'\left(Y - g(T; \boldsymbol{\beta})\right) m(T; \boldsymbol{\beta})\right],\end{aligned}$$

and the functional derivative $\Gamma_2(\boldsymbol{\beta}, \pi_0(\cdot))[\pi(\cdot) - \pi_0(\cdot)]$ of $M(\boldsymbol{\beta}, \pi_0(\cdot))$ along the direction $\pi(\cdot) - \pi_0(\cdot)$ is

$$\begin{aligned}\Gamma_2(\boldsymbol{\beta}, \pi_0(\cdot))[\pi(\cdot) - \pi_0(\cdot)] :=& \lim_{\tau \to 0} \frac{M(\boldsymbol{\beta}, \pi_0(\cdot) + \tau(\pi(\cdot) - \pi_0(\cdot))) - M(\boldsymbol{\beta}, \pi_0(\cdot))}{\tau} \\ =& \mathbb{E}\left[(\pi(T, \boldsymbol{X}) - \pi_0(T, \boldsymbol{X})) L'\left(Y - g(T; \boldsymbol{\beta})\right) m(T; \boldsymbol{\beta})\right].\end{aligned}$$

In order to apply Theorem 2 of Chen, Linton, and Van Keilegom (2003), we need to verify their Conditions (2.1)-(2.6) hold. Conditions (2.1)-(2.5) of Chen, Linton, and Van Keilegom (2003) can be easily verified by using following facts:

- Theorem 4 ensures $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \xrightarrow{p} 0$;
- Assumption 1.6 (iv) implies $\|M_N(\hat{\boldsymbol{\beta}}, \hat{\pi}_K(\cdot))\| = \left\|N^{-1} \sum_{i=1}^N \hat{\pi}_K(T_i, \boldsymbol{X}_i) m(T_i; \hat{\boldsymbol{\beta}}) L'\{Y_i - g(T_i; \hat{\boldsymbol{\beta}})\}\right\| = o_P(1/\sqrt{N})$;
- Assumption 1.8 implies $K = o_p(N^{1/2})$ and $K^{-\alpha} = o_p(N^{-1/2})$, then by Theorem 2 we have $\int_{\mathcal{T} \times \mathcal{X}} |\hat{\pi}_K(t, \boldsymbol{x}) - \pi_0(t, \boldsymbol{x})|^2 dF_{T,X}(t, \boldsymbol{x}) = O_p\left(K^{-\alpha}\right) + O_p\left(\sqrt{K/N}\right) = o_P(N^{-1/2}) + o_P(N^{-1/4}) \leq o_p(N^{-1/4})$.

The most important step toward the application of Theorem 2 of Chen, Linton, and Van Keilegom (2003) is to check their Condition (2.6) holds, which states that there exits some finite matrix V_1 such that

$$\sqrt{N}\left\{M_N(\boldsymbol{\beta}^*, \pi_0(\cdot)) + \Gamma_2(\boldsymbol{\beta}^*, \pi_0(\cdot))[\hat{\pi}_K(\cdot) - \pi_0(\cdot)]\right\} \xrightarrow{d} N(0, V_1). \tag{40}$$

If Conditions (2.1)-(2.6) hold, Theorem 2 of Chen, Linton, and Van Keilegom (2003) ensures that

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \xrightarrow{d} \mathcal{N}(0, \Omega),$$

where $\Omega := \Gamma_1(\boldsymbol{\beta}^*, \pi_0(\cdot))^{-1} V_1 (\Gamma_1(\boldsymbol{\beta}^*, \pi_0(\cdot))^{-1})^\top = H_0^{-1} V_1 (H_0^{-1})^\top$. However, Chen, Linton, and Van Keilegom (2003) do not give the expression of V_1 and the verification of (40) is difficult which

is also admitted by the authors themselves (see the first paragraph in Section 3.3 of Chen, Linton, and Van Keilegom (2003)). In Section 4.3, we prove (40) holds and give

$$V_1 = \mathbb{E}[\psi(Y, T, \boldsymbol{X}; \boldsymbol{\beta}^*)\psi(Y, T, \boldsymbol{X}; \boldsymbol{\beta}^*)^\top].$$

Therefore, we can have $\Omega = V_{eff}$ which justifies Theorem 5.

4.3 Proof of (40)

Before proving (40), we prepare some preliminary notation and results that will be used later. Since $\hat{\Lambda}_{K_1\times K_2}$ is a unique maximizer of the concave function $\hat{G}_{K_1\times K_2}$, then

$$\frac{1}{N}\sum_{i=1}^{N}\rho'\left(u_{K_1}(T_i)^\top\hat{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)u_{K_1}(T_i)v_{K_2}(\boldsymbol{X}_i)^\top - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{l=1}^{N}u_{K_1}(T_l)v_{K_2}(\boldsymbol{X}_i)^\top = 0.$$

Using Mean Value Theorem, we can have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\rho'\left(u_{K_1}(T_i)^\top\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)u_{K_1}(T_i)v_{K_2}(\boldsymbol{X}_i)^\top\\
+&\frac{1}{N}\sum_{i=1}^{N}\rho''\left(u_{K_1}(T_i)^\top\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)u_{K_1}(T_i)u_{K_1}(T_i)^\top\left\{\hat{\Lambda}_{K_1\times K_2}-\Lambda^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X}_i)v_{K_2}(\boldsymbol{X}_i)^\top\\
=&\frac{1}{N^2}\sum_{i=1}^{N}\sum_{l=1}^{N}u_{K_1}(T_l)v_{K_2}(\boldsymbol{X}_i)^\top\ ,
\end{aligned}\tag{41}$$

where $\tilde{\Lambda}_{K_1\times K_2}$ lies on the line joining from $\hat{\Lambda}_{K_1\times K_2}$ to $\Lambda^*_{K_1\times K_2}$. We define the following notation:

$$\hat{A}_{K_1\times K_2} := \hat{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2},\tag{42}$$

$$\tilde{A}_{K_1\times K_2} := \tilde{\Lambda}_{K_1\times K_2} - \Lambda^*_{K_1\times K_2},\tag{43}$$

and

$$\begin{aligned}
&A^*_{K_1\times K_2} := \nabla\hat{G}_{K_1\times K_2}\left(\Lambda^*_{K_1\times K_2}\right)\\
=&\frac{1}{N}\sum_{i=1}^{N}\rho'\left(u_{K_1}(T_i)^\top\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)u_{K_1}(T_i)v_{K_2}(\boldsymbol{X}_i)^\top - \left(\frac{1}{N}\sum_{l=1}^{N}u_{K_1}(T_l)\right)\left(\frac{1}{N}\sum_{i=1}^{N}v_{K_2}(\boldsymbol{X}_i)^\top\right).
\end{aligned}\tag{44}$$

In light of (27) we have

$$\left\|A^*_{K_1\times K_2}\right\| = O_p\left(\sqrt{\frac{K}{N}}\right).$$

From (41), $A^*_{K_1\times K_2}$ can also be written as

$$A^*_{K_1\times K_2} = -\frac{1}{N}\sum_{i=1}^{N}\rho''\left(u_{K_1}(T_i)^\top\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)u_{K_1}(T_i)u_{K_1}(T_i)^\top\left\{\hat{\Lambda}_{K_1\times K_2}-\Lambda^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{X}_i)v_{K_2}(\boldsymbol{X}_i)^\top. \tag{45}$$

We now start to (40). We decompose $\sqrt{N}\left\{M_N(\boldsymbol{\beta}^*,\pi_0(\cdot))+\Gamma_2(\boldsymbol{\beta}^*,\pi_0(\cdot))[\hat{\pi}_K(\cdot)-\pi_0(\cdot)]\right\}$ as follows:

$$\begin{aligned}
&\sqrt{N}\left\{M_N(\boldsymbol{\beta}^*,\pi_0(\cdot))+\Gamma_2(\boldsymbol{\beta}^*,\pi_0(\cdot))[\hat{\pi}_K(\cdot)-\pi_0(\cdot)]\right\}\\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\pi_0(T_i,\boldsymbol{X}_i)L'\left\{Y_i-g\left(T_i;\boldsymbol{\beta}^*\right)\right\}m(T_i;\boldsymbol{\beta}^*)+\int_{\mathcal{T}}\int_{\mathcal{X}}\left(\hat{\pi}_K(t,\boldsymbol{x})-\pi_0(t,\boldsymbol{x})\right)\varepsilon(\boldsymbol{x},t;\boldsymbol{\beta}^*)m(t;\boldsymbol{\beta}^*)dF_{X,T}(\boldsymbol{x},t)\right\}\\
=&\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\left(\pi^*_K(t,\boldsymbol{x})-\pi_0(t,\boldsymbol{x})\right)dF_{X,T}(\boldsymbol{x},t)\\
&+\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}\left(\hat{\pi}_K(t,\boldsymbol{x})-\pi^*_K(t,\boldsymbol{x})\right)\varepsilon(\boldsymbol{x},t;\boldsymbol{\beta}^*)m(t;\boldsymbol{\beta}^*)dF_{X,T}(\boldsymbol{x},t)\\
&+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\pi_0(T_i,\boldsymbol{X}_i)L'\left\{Y_i-g\left(T_i;\boldsymbol{\beta}^*\right)\right\}m(T_i;\boldsymbol{\beta}^*)\\
=&\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\left(\pi^*_K(t,\boldsymbol{x})-\pi_0(t,\boldsymbol{x})\right)dF_{X,T}(\boldsymbol{x},t) &(46)\\
&+\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}\left(\hat{\pi}_K(t,\boldsymbol{x})-\pi^*_K(t,\boldsymbol{x})\right)\varepsilon(\boldsymbol{x},t;\boldsymbol{\beta}^*)m(t;\boldsymbol{\beta}^*)dF_{X,T}(\boldsymbol{x},t) &(47)\\
&\quad-\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u^\top_{K_1}(t)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u^\top_{K_1}(t)\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})m(t;\boldsymbol{\beta}^*)dF_{X,T}(\boldsymbol{x},t)\\
&+\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u^\top_{K_1}(t)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u^\top_{K_1}(t)\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})m(t;\boldsymbol{\beta}^*)dF_{X,T}(\boldsymbol{x},t) &(48)\\
&\quad-\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u^\top_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u^\top_{K_1}(t)A^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})m(t;\boldsymbol{\beta}^*)dF_{X,T}(\boldsymbol{x},t)\\
&+\sqrt{N}\int_{\mathcal{X}}\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u^\top_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u^\top_{K_1}(t)A^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})m(t;\boldsymbol{\beta}^*)dF_{X,T}(\boldsymbol{x},t) &(49)\\
&\quad+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big\{\pi_0(T_i,\boldsymbol{X}_i)m(T_i;\boldsymbol{\beta}^*)\varepsilon(T_i,\boldsymbol{X}_i;\boldsymbol{\beta}^*)-\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{X}_i\right]\\
&\qquad\qquad-\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|T=T_i\right]\Big\}
\end{aligned}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big\{\pi_0(T_i,\boldsymbol{X}_i)L'\left\{Y_i-g\left(T_i;\boldsymbol{\beta}^*\right)\right\}m(T_i;\boldsymbol{\beta}^*)-\pi_0(T_i,\boldsymbol{X}_i)m(T_i;\boldsymbol{\beta}^*)\varepsilon(T_i,\boldsymbol{X}_i;\boldsymbol{\beta}^*) \tag{50}$$
$$+\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{X}_i\right]+\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|T=T_i\right]\Big\},$$

where $\hat{A}_{K_1\times K_2}$ and $A^*_{K_1\times K_2}$ are defined in (42) and (45). We show that the terms (46)-(49) are all of $o_p(1)$, while the term (50) is asymptotically normal.

For term (46)**:** By Lemma 3.1 and Assumption 1.4, we can deduce that

$$\begin{aligned}&\left\|\sqrt{N}\cdot\mathbb{E}\left[m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\left(\pi^*_K(T,\boldsymbol{X})-\pi_0(T,\boldsymbol{X})\right)\right]\right\|\\ \leq&\sqrt{N}\sup_{t\in\mathcal{T}}\|m(t;\boldsymbol{\beta}^*)\|\cdot\mathbb{E}[|\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|^2]^{\frac{1}{2}}\cdot\mathbb{E}\left[|\pi^*_K(T,\boldsymbol{X})-\pi_0(T,\boldsymbol{X})|^2\right]^{\frac{1}{2}}=O\left(\sqrt{N}K^{-\alpha}\right)\ .\end{aligned}$$

For term (47)**:** By Mean Value Theorem and the definition of $\hat{A}_{K_1\times K_2}$ in (42), the term (47) is exactly equal to zero.

For term (48)**:** We can telescope (48) as follows:

$$\begin{aligned}&\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u^\top_{K_1}(t)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u^\top_{K_1}(t)\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t)\\ &-\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u^\top_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u^\top_{K_1}(t)A^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t)\\ =&\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\left\{\rho''\left(u^\top_{K_1}(t)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)-\rho''\left(u^\top_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)\right\}\\ &\qquad\times u^\top_{K_1}(t)\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t) \qquad (51)\\ &+\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u^\top_{K_1}(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u^\top_{K_1}(t)\\ &\qquad\times\left\{\hat{A}_{K_1\times K_2}-A^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t). \qquad (52)\end{aligned}$$

For the term (51), by Mean Value Theorem,

$$(51)=\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho'''\left(\xi_3(t,\boldsymbol{x})\right)\left\{u^\top_{K_1}(t)\tilde{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right\}\left\{u^\top_{K_1}(t)\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right\}dF_{X,T}(\boldsymbol{x},t).$$

Since $\xi_3(t,\boldsymbol{x})$ lies between $u_{K_1}(t)^\top\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})$ and $u_{K_1}(t)^\top\tilde{\Lambda}^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})$, which implies $\xi_3(t,\boldsymbol{x})$ lies between $u_{K_1}(t)^\top\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})$ and $u_{K_1}(t)^\top\hat{\Lambda}^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})$. Then in light of (28) and (34), we

have $\mathbb{P}\left(\xi_3(t,\boldsymbol{x}) \in \Gamma_2(\epsilon),\ \forall (t,\boldsymbol{x}) \in \mathcal{T}\times\mathcal{X}\right) > 1-\epsilon$, therefore,

$$\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} \left|\rho'''\left(\xi_3(t,\boldsymbol{x})\right)\right| = O_p(1)\ . \tag{53}$$

With (36), (53), the fact $\|\tilde{A}_{K_1\times K_2}\| \le \|\hat{A}_{K_1\times K_2}\|$, Lemma 3.2, and Assumption 1.4, we can derive that

$$\begin{aligned}
\|(51)\| \le & \sqrt{N} \sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} \left|\rho'''\left(\xi_3(t,\boldsymbol{x})\right)\right| \sup_{t\in\mathcal{T}} \|m(t;\boldsymbol{\beta}^*)\| \cdot \sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} |\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)| \\
& \cdot \int_{\mathcal{T}}\int_{\mathcal{X}} \left|u_{K_1}(t)^\top \tilde{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{x})\right| \cdot \left|u_{K_1}^\top(t) \hat{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{x})\right| dF_{X,T}(\boldsymbol{x},t) \\
\le & \sqrt{N}\cdot O_p(1)\cdot O(1)\cdot O(1)\cdot \left\{\int_{\mathcal{T}}\int_{\mathcal{X}} \left|u_{K_1}(t)^\top \tilde{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{x})\right|^2 dF_{X,T}(\boldsymbol{x},t)\right\}^{\frac{1}{2}} \\
& \cdot \left\{\int_{\mathcal{T}}\int_{\mathcal{X}} \left|u_{K_1}^\top(t) \hat{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{x})\right|^2 dF_{X,T}(\boldsymbol{x},t)\right\}^{\frac{1}{2}} \\
= & \sqrt{N}\cdot O_p(1)\cdot O(1)\cdot O(1)\cdot O_p\left(\sqrt{\frac{K}{N}}\right)\cdot O_p\left(\sqrt{\frac{K}{N}}\right) = O_p\left(\sqrt{\frac{K^2}{N}}\right) \qquad \text{(by ((36)))}.
\end{aligned} \tag{54}$$

For the term (52), we first compute the probability order of $\|A^*_{K_1\times K_2} - \hat{A}_{K_1\times K_2}\|$. Using (45), the fact $\rho''(v) = -\rho'(v)$ and Mean Value Theorem, we have

$$\begin{aligned}
& A^*_{K_1\times K_2} - \hat{A}_{K_1\times K_2} \\
= & -\frac{1}{N}\sum_{i=1}^N \rho''\left(u_{K_1}^\top(T_i)\Lambda^*_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i)\right) u_{K_1}(T_i) u_{K_1}(T_i)^\top \hat{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i) v_{K_2}^\top(\boldsymbol{X}_i) \\
& -\frac{1}{N}\sum_{i=1}^N \rho'''\left(\xi_3(T_i,\boldsymbol{X}_i)\right)\left\{u_{K_1}(T)^\top \tilde{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i)\right\} u_{K_1}(T_i) u_{K_1}(T_i)^\top \hat{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i) v_{K_2}^\top(\boldsymbol{X}_i) \\
& -\hat{A}_{K_1\times K_2} \\
= & \frac{1}{N}\sum_{i=1}^N \left\{\rho'\left(u_{K_1}^\top(T_i)\Lambda^*_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i)\right) u_{K_1}(T_i) u_{K_1}(T_i)^\top \hat{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i) v_{K_2}^\top(\boldsymbol{X}_i) - \hat{A}_{K_1\times K_2}\right\} \qquad (55) \\
& -\frac{1}{N}\sum_{i=1}^N \rho'''\left(\xi_3(T_i,\boldsymbol{X}_i)\right)\left\{u_{K_1}(T_i)^\top \tilde{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i)\right\} u_{K_1}(T_i) u_{K_1}(T_i)^\top \hat{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i) v_{K_2}^\top(\boldsymbol{X}_i). \qquad (56)
\end{aligned}$$

For the term (55), by (13) we can write $\hat{A}_{K_1\times K_2}$ as

$$\hat{A}_{K_1\times K_2} = \mathbb{E}_{T,\boldsymbol{X}}\left[\pi_0(T,\boldsymbol{X}) u_{K_1}(T) u_{K_1}(T)^\top \hat{A}_{K_1\times K_2} v_{K_2}(\boldsymbol{X}) v_{K_2}^\top(\boldsymbol{X})\right],$$

where $\mathbb{E}_{T,\boldsymbol{X}}[\cdot]$ denotes taking expectation with respect to $(T, \boldsymbol{X})$. We telescope (55) as follows:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left\{\rho'\left(u_{K_1}^\top(T_i)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)u_{K_1}(T_i)u_{K_1}(T_i)^\top\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)v_{K_2}^\top(\boldsymbol{X}_i)-\hat{A}_{K_1\times K_2}\right\}\\
=&\frac{1}{N}\sum_{i=1}^{N}\left\{\left\{\rho'\left(u_{K_1}^\top(T_i)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)-\pi_0(T_i,\boldsymbol{X}_i)\right\}u_{K_1}(T_i)u_{K_1}(T_i)^\top\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)v_{K_2}^\top(\boldsymbol{X}_i)\right\} \qquad (57)\\
&-\frac{1}{N}\sum_{i=1}^{N}\Big\{\pi_0(T_i,\boldsymbol{X}_i)u_{K_1}(T_i)u_{K_1}(T_i)^\top\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)v_{K_2}^\top(\boldsymbol{X}_i)\\
&\qquad\qquad-\mathbb{E}_{T,\boldsymbol{X}}\left[\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X})v_{K_2}^\top(\boldsymbol{X})\right]\Big\}. \qquad (58)
\end{aligned}$$

For the term (57), by Lemmas 3.1 and 3.2 and (36), we have that

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{i=1}^{N}\left\{\rho'\left(u_{K_1}^\top(T_i)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)-\pi_0(T_i,\boldsymbol{X}_i)\right\}u_{K_1}(T_i)u_{K_1}(T_i)^\top\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)v_{K_2}^\top(\boldsymbol{X}_i)\right\|\\
\le&\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\rho'\left(u_{K_1}^\top(T_i)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)-\pi_0(T_i,\boldsymbol{X}_i)\right|^2\|u_{K_1}(T_i)\|^2\|v_{K_2}(\boldsymbol{X}_i)\|^2}\\
&\cdot\sqrt{\frac{1}{N}\sum_{i=1}^{N}|u_{K_1}(T_i)^\top\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)|^2}\\
\le&O(\zeta(K)K^{-\alpha})\cdot\left[\int_{\mathcal{T}\times\mathcal{X}}|u_{K_1}(t)\widehat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})|^2dF_{T,X}(t,\boldsymbol{x})+\|\widehat{A}_{K_1\times K_2}\|^2\cdot O_p\left(\zeta(K)\sqrt{\frac{K}{N}}\right)\right]^{1/2}\\
\le&O(\zeta(K)K^{-\alpha})\cdot O_p(\|\widehat{A}_{K_1\times K_2}\|)\\
=&O(\zeta(K)K^{-\alpha})\cdot O_p\left(\sqrt{\frac{K}{N}}\right)=O_p\left(N^{-\frac{1}{2}}\zeta(K)\cdot K^{\frac{1}{2}-\alpha}\right).
\end{aligned}$$

For the term (58), define the linear map $\mathcal{J}(\cdot):\mathbb{R}^{K_1\times K_2}\to\mathbb{R}$ by

$$\mathcal{J}(M):=\frac{1}{N}\sum_{i=1}^{N}\left\{\pi_0(T_i,\boldsymbol{X}_i)u_{K_1}(T_i)u_{K_1}(T_i)^\top Mv_{K_2}(\boldsymbol{X}_i)v_{K_2}^\top(\boldsymbol{X}_i)-\mathbb{E}_{T,\boldsymbol{X}}\left[\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top Mv_{K_2}(\boldsymbol{X})v_{K_2}^\top(\boldsymbol{X})\right]\right\},$$

then (58) $=\mathcal{J}(\hat{A}_{K_1\times K_2})$. For any fixed $M\in\mathbb{R}^{K_1\times K_2}$, by (13) and $M=\mathbb{E}[\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top M$ $\cdot v_{K_2}(\boldsymbol{X})v_{K_2}^\top(\boldsymbol{X})]$, then we have

$$\mathbb{E}\left[\mathcal{J}(M)^2\right]$$

$$
\begin{aligned}
=&\frac{1}{N}\cdot\mathbb{E}\left[\left\|\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top Mv_{K_2}(\boldsymbol{X})v_{K_2}^\top(\boldsymbol{X})-\mathbb{E}\left[\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top Mv_{K_2}(\boldsymbol{X})v_{K_2}^\top(\boldsymbol{X})\right]\right\|^2\right]\\
\leq&\frac{1}{N}\cdot\mathbb{E}\left[\left\|\pi_0(T,\boldsymbol{X})u_{K_1}(T)u_{K_1}(T)^\top Mv_{K_2}(\boldsymbol{X})v_{K_2}^\top(\boldsymbol{X})\right\|^2\right]\\
\leq&\frac{1}{N}\cdot\eta_2\cdot\mathbb{E}\left[\pi_0(T,\boldsymbol{X})\cdot\|u_{K_1}(T)\|^4\|v_{K_2}(\boldsymbol{X})\|^4\right]\cdot\|M\|^2\\
=&\frac{1}{N}\cdot\eta_2\cdot\mathbb{E}[\|u_{K_1}(T)\|^4]\cdot\mathbb{E}[\|v_{K_2}(\boldsymbol{X})\|^4]\cdot\|M\|^2\\
\leq&\frac{1}{N}\cdot\eta_2\cdot\zeta_1(K)^2\cdot\zeta_2(K)^2\cdot\mathbb{E}[\|u_{K_1}(T)\|^2]\cdot\mathbb{E}[\|v_{K_2}(\boldsymbol{X})\|^2]\cdot\|M\|^2\\
=&\|M\|^2\cdot O\left(\zeta(K)^2\frac{K}{N}\right).
\end{aligned}
$$

Using Chebyshev's inequality we have

$$
|\mathcal{J}(M)|=\|M\|O_p\left(\zeta(K)\sqrt{\frac{K}{N}}\right),
$$

then in light of Lemma 3.2,

$$
(58)=\mathcal{J}(\hat{A}_{K_1\times K_2})=\|\hat{A}_{K_1\times K_2}\|O_p\left(\zeta(K)\sqrt{\frac{K}{N}}\right)=O_p\left(\zeta(K)\frac{K}{N}\right).
$$

Therefore,

$$
(55)=(57)+(58)=O_p\left(N^{-\frac{1}{2}}\zeta(K)\cdot K^{\frac{1}{2}-\alpha}\right)+O_p\left(\zeta(K)\frac{K}{N}\right).
$$

For the term (56), in light of (53) and Lemma 3.2, we can deduce that

$$
\begin{aligned}
&\left\|\frac{1}{N}\sum_{i=1}^N\rho'''\left(\xi_3(T_i,\boldsymbol{X}_i)\right)\left\{u_{K_1}(T_i)^\top\tilde{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right\}\left\{u_{K_1}(T_i)u_{K_1}(T_i)^\top\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)v_{K_2}^\top(\boldsymbol{X}_i)\right\}\right\|\\
\leq&\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}\left|\rho'''\left(\xi_3(t,\boldsymbol{x})\right)\right|\cdot\zeta(K)\cdot\frac{1}{N}\sum_{i=1}^N\left|u_{K_1}(T_i)^\top\tilde{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right|\cdot\left|u_{K_1}(T_i)^\top\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right|\\
\leq&\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}\left|\rho'''\left(\xi_3(t,\boldsymbol{x})\right)\right|\cdot\zeta(K)\cdot\sqrt{\frac{1}{N}\sum_{i=1}^N\left|u_{K_1}(T_i)^\top\tilde{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right|^2}\cdot\sqrt{\frac{1}{N}\sum_{i=1}^N\left|u_{K_1}(T_i)^\top\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right|^2}\\
\leq&O(1)\cdot\zeta(K)\cdot O_p(\|\tilde{A}_{K_1\times K_2}\|)\cdot O_p(\|\hat{A}_{K_1\times K_2}\|)\leq O_p(1)\cdot\zeta(K)\cdot O_p\left(\sqrt{\frac{K}{N}}\right)\cdot O_p\left(\sqrt{\frac{K}{N}}\right)\leq O_p\left(\zeta(K)\frac{K}{N}\right).
\end{aligned}
$$

Now, we can obtain

$$\|\hat{A}_{K_1\times K_2}-A^*_{K_1\times K_2}\| = (55)+(56) = O_p\left(N^{-\frac{1}{2}}\zeta(K)K^{\frac{1}{2}-\alpha}\right)+O_p\left(\zeta(K)\frac{K}{N}\right)+O_p\left(\zeta(K)\frac{K}{N}\right)$$
$$= O_p\left(N^{-\frac{1}{2}}\zeta(K)\cdot K^{\frac{1}{2}-\alpha}\right)+O_p\left(\zeta(K)\frac{K}{N}\right)\ . \tag{59}$$

Using (59), Assumptions 1.7 and 1.4, for large enough N, we have

$$(52) = \left\|\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u_{K_1}^\top(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u_{K_1}^\top(t)\left\{\hat{A}_{K_1\times K_2}-A^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t)\right\|$$
$$\leq\sqrt{N}\sup_{t\in\mathcal{T}}\|m(t;\boldsymbol{\beta}^*)\|\sup_{\gamma\in\Gamma_1}|\rho''(\gamma)|\cdot\mathbb{E}\left[|\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|^2\right]^{\frac{1}{2}}\cdot\left[\int_{\mathcal{T}\times\mathcal{X}}\left(u_{K_1}(t)\left\{\hat{A}_{K_1\times K_2}-A^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})\right)^2 dF_{T,X}(t,\boldsymbol{x})\right]^{\frac{1}{2}}$$
$$\leq\sqrt{N}\cdot O(1)\cdot O(1)\cdot O(1)\cdot O(1)\cdot O(\|\hat{A}_{K_1\times K_2}-A^*_{K_1\times K_2}\|)$$
$$\leq O_p\left(\zeta(K)\cdot K^{\frac{1}{2}-\alpha}\right)+O_p\left(\zeta(K)\frac{K}{\sqrt{N}}\right), \tag{60}$$

where the second inequality holds since by using the same argument of establishing (36), we have

$$\int_{\mathcal{T}\times\mathcal{X}}\left(u_{K_1}(t)\left\{\hat{A}_{K_1\times K_2}-A^*_{K_1\times K_2}\right\}v_{K_2}(\boldsymbol{x})\right)^2 dF_{T,X}(t,\boldsymbol{x}) = O(\|\hat{A}_{K_1\times K_2}-A^*_{K_1\times K_2}\|).$$

Therefore, by combining (54) and (60), we can obtain that

$$(48) = (51)+(52) = O_p\left(\sqrt{\frac{K^2}{N}}\right)+O_p\left(\zeta(K)\cdot K^{\frac{1}{2}-\alpha}\right)+O_p\left(\zeta(K)\frac{K}{\sqrt{N}}\right)$$
$$= O_p\left(\zeta(K)\cdot K^{\frac{1}{2}-\alpha}\right)+O_p\left(\zeta(K)\frac{K}{\sqrt{N}}\right).$$

For term (49)**:** By the definition of $A^*_{K_1\times K_2}$ in (44), we have

$$(49) = \frac{1}{\sqrt{N}}\sum_{i=1}^N\Bigg\{\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u_{K_1}^\top(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u_{K_1}^\top(t) \tag{61}$$
$$\times\left\{u_{K_1}(T_i)\rho'\left(u_{K_1}(T_i)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)v_{K_2}^\top(\boldsymbol{X}_i)\right\}v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t)+m(T_i;\boldsymbol{\beta}^*)\varepsilon(T_i,\boldsymbol{X}_i;\boldsymbol{\beta}^*)\pi_0(T_i,\boldsymbol{X}_i)\Bigg\}$$
$$-\frac{1}{\sqrt{N}}\sum_{i=1}^N\Bigg\{\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho''\left(u_{K_1}^\top(t)\Lambda^*_{K_1\times K_2}v_{K_2}^\top(\boldsymbol{x})\right)u_{K_1}^\top(t)\left(\frac{1}{N}\sum_{l=1}^N u_{K_1}(T_l)\right) \tag{62}$$
$$\times\left(\frac{1}{N}\sum_{j=1}^N v_{K_2}^\top(\boldsymbol{X}_j)\right)v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t)+\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{X}_i\right]$$
$$+\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|T=T_i\right]\Bigg\}.$$

We shall show that both (61) and (62) are of $o_p(1)$. Noting $\rho'' = -\rho'$, we can telescope (61) as follows:

$$
\begin{aligned}
(61) =& \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Bigg\{ \int_{\mathcal{T}} \int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*) \varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*) \rho' \left(u_{K_1}^\top(t) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right) u_{K_1}^\top(t) && (63) \\
& \qquad \times \left\{ u_{K_1}(T_i) \Big[-\rho' \left(u_{K_1}(T_i) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{X}_i) \right) + \pi_0(T_i, \boldsymbol{X}_i) \Big] v_{K_2}^\top(\boldsymbol{X}_i) \right\} v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x},t) \Bigg\} \\
& - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Bigg\{ \int_{\mathcal{T}} \int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*) \varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*) \left\{ \rho' \left(u_{K_1}^\top(t) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right) - \pi_0(t,\boldsymbol{x}) \right\} u_{K_1}^\top(t) && (64) \\
& \qquad \times \left\{ u_{K_1}(T_i) \pi_0(T_i, \boldsymbol{X}_i) v_{K_2}^\top(\boldsymbol{X}_i) \right\} v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x},t) \Bigg\} \\
& - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Bigg\{ \int_{\mathcal{T}} \int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*) \varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*) \pi_0(t,\boldsymbol{x}) u_{K_1}^\top(t) \left\{ u_{K_1}(T_i) \pi_0(T_i, \boldsymbol{X}_i) v_{K_2}^\top(\boldsymbol{X}_i) \right\} v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x},t) \\
& \qquad + m(T_i;\boldsymbol{\beta}^*) \varepsilon(T_i, \boldsymbol{X}_i;\boldsymbol{\beta}^*) \pi_0(T_i, \boldsymbol{X}_i) \Bigg\} . && (65)
\end{aligned}
$$

We shall show that (63), (64) and (65) are all of $o_p(1)$. Note that second moment of (63) is

$$
\begin{aligned}
\mathbb{E}[|(63)|^2] =& \mathbb{E} \Bigg[\Bigg| \int_{\mathcal{T}} \int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*) \varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*) \rho' \left(u_{K_1}^\top(t) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right) u_{K_1}^\top(t) \\
& \qquad \times \left\{ u_{K_1}(T_i) \Big[-\rho' \left(u_{K_1}(T_i) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{X}_i) \right) + \pi_0(T_i, \boldsymbol{X}_i) \Big] v_{K_2}^\top(\boldsymbol{X}_i) \right\} v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x},t) \Bigg|^2 \Bigg] \\
=& \mathbb{E} \Bigg[\Bigg| \int_{\mathcal{T}} \int_{\mathcal{X}} \pi_0(t,\boldsymbol{x}) \cdot m(t;\boldsymbol{\beta}^*) \varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*) \left[\frac{\rho' \left(u_{K_1}^\top(t) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right)}{\pi_0(t,\boldsymbol{x})} \right] u_{K_1}^\top(t) \\
& \qquad \times \left\{ u_{K_1}(T_i) \Big[-\rho' \left(u_{K_1}(T_i) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{X}_i) \right) + \pi_0(T_i, \boldsymbol{X}_i) \Big] v_{K_2}^\top(\boldsymbol{X}_i) \right\} v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x},t) \Bigg|^2 \Bigg] \\
\le& \mathbb{E} \Bigg[\Bigg| \int_{\mathcal{T}} \int_{\mathcal{X}} \pi_0(t,\boldsymbol{x}) \cdot m(t;\boldsymbol{\beta}^*) \varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*) \left[\frac{\rho' \left(u_{K_1}^\top(t) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right)}{\pi_0(t,\boldsymbol{x})} \right] u_{K_1}^\top(t) \left\{ u_{K_1}(T_i) v_{K_2}^\top(\boldsymbol{X}_i) \right\} v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x},t) \Bigg|^2 \Bigg] \\
& \qquad \times \sup_{(t,\boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}} \left\{ -\rho' \left(u_{K_1}^\top(t) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right) + \pi_0(t,\boldsymbol{x}) \right\}^2 \\
=& \left\{ \mathbb{E} \left[\left| m(T_i;\boldsymbol{\beta}^*) \varepsilon(T_i, \boldsymbol{X}_i;\boldsymbol{\beta}^*) \left[\frac{\rho' \left(u_{K_1}^\top(T_i) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{X}_i) \right)}{\pi_0(T_i, \boldsymbol{X}_i)} \right] \right|^2 \right] + o(1) \right\} \times \sup_{(t,\boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}} \left\{ -\pi_K^*(t,\boldsymbol{x}) + \pi_0(t,\boldsymbol{x}) \right\}^2 \\
=& O(1) \cdot O(K^{-2\alpha} \zeta(K)^2) = O(K^{-2\alpha} \zeta(K)^2) \to 0, \text{ (by Assumption 1.4)}
\end{aligned}
$$

where the third equality holds because

$$
\int_{\mathcal{T}} \int_{\mathcal{X}} \pi_0(t,\boldsymbol{x}) \cdot m(t;\boldsymbol{\beta}^*) \varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*) \left[\frac{\rho' \left(u_{K_1}^\top(t) \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right)}{\pi_0(t,\boldsymbol{x})} \right] u_{K_1}^\top(t) \left\{ u_{K_1}(T) v_{K_2}^\top(\boldsymbol{X}) \right\} v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x},t)
$$

is the weighted L^2-projection of $m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\left[\frac{\rho'\left(u_{K_1}^\top(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)}{\pi_0(t,\boldsymbol{x})}\right]$ on the space linearly spanned by $\{u_{K_1}(t), v_{K_2}(\boldsymbol{x})\}$ with the weighted measure $\pi_0(t,\boldsymbol{x})dF_{T,X}(t,\boldsymbol{x})$. Similarly, we can also show (64) and (65) are of $o_p(1)$. Therefore, (61) is of $o_p(1)$.

For the term (62), since $\rho''(v) = -\rho'(v)$ and the fact $\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\right] = 0$, we telescope it as follows:

$$
\begin{aligned}
(62) =& \sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho'\left(u_{K_1}^\top(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u_{K_1}^\top(t)\left(\frac{1}{N}\sum_{l=1}^N u_{K_1}(T_l) - \mathbb{E}\left[u_{K_1}(T)\right]\right) \\
&\times\left(\frac{1}{N}\sum_{j=1}^N v_{K_2}^\top(\boldsymbol{X}_j) - \mathbb{E}[v_{K_2}^\top(\boldsymbol{X})]\right)v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t) \qquad (66)\\
&+\frac{1}{\sqrt{N}}\sum_{i=1}^N\Big\{\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho'\left(u_{K_1}^\top(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u_{K_1}^\top(t)\mathbb{E}\left[u_{K_1}(T)\right]v_{K_2}^\top(\boldsymbol{X}_i)v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t) \\
&\qquad - \mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|\boldsymbol{X}=\boldsymbol{X}_i\right]\Big\} \qquad (67)\\
&+\frac{1}{\sqrt{N}}\sum_{l=1}^N\Big\{\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho'\left(u_{K_1}^\top(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u_{K_1}^\top(t)u_{K_1}(T_l)\mathbb{E}[v_{K_2}^\top(\boldsymbol{X})]v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t) \\
&\qquad - \mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|T=T_i\right]\Big\} \qquad (68)\\
&-\frac{1}{\sqrt{N}}\sum_{i=1}^N\Big\{\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\rho'\left(u_{K_1}^\top(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u_{K_1}^\top(t)\mathbb{E}\left[u_{K_1}(T)\right]\mathbb{E}[v_{K_2}^\top(\boldsymbol{X})]v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t) \\
&\qquad - \mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\right]\Big\}. \qquad (69)
\end{aligned}
$$

For the term (66), since

$$
\begin{aligned}
\left\|\frac{1}{N}\sum_{l=1}^N u_{K_1}(T_l) - \mathbb{E}\left[u_{K_1}(T)\right]\right\| &= O_p\left(\sqrt{\frac{K_1}{N}}\right)\ ,\\
\left\|\frac{1}{N}\sum_{j=1}^N v_{K_2}(\boldsymbol{X}_j) - \mathbb{E}\left[v_{K_2}(\boldsymbol{X})\right]\right\| &= O_p\left(\sqrt{\frac{K_2}{N}}\right)\ ,\\
\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}\left|\rho'\left(u_{K_1}^\top(t)\Lambda^*_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)\right| &= O(1)\ ,
\end{aligned}
$$

and by Assumptions 1.4, 1.6 and 1.7, we can deduce that

$$
(66) = \sqrt{N}\cdot O(\zeta(K))O_p\left(\sqrt{\frac{K_1}{N}}\right)O_p\left(\sqrt{\frac{K_2}{N}}\right) = O_p\left(\zeta(K)\sqrt{\frac{K}{N}}\right) = o_p(1)\ .
$$

For the term (67), noting the fact that $\mathbb{E}\left[\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)|\boldsymbol{X}\right] = \int_{\mathcal{T}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{X};\boldsymbol{\beta}^*)$ $dF_T(t)$, we can rewrite (67) as follows:

$$(67) = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left\{\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\frac{\pi_K^*(t,\boldsymbol{x})}{\pi_0(t,\boldsymbol{x})}u_{K_1}^\top(t)\mathbb{E}\left[u_{K_1}(T)\right]\, v_{K_2}^\top(\boldsymbol{X}_j)v_{K_2}(\boldsymbol{x})dF_X(\boldsymbol{x})dF_T(t)\right.$$
$$\left. - \int_{\mathcal{T}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{X}_j;\boldsymbol{\beta}^*)dF_T(t)\right\}.$$

By computing the second moment of (67), we can obtain that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\frac{\pi_K^*(t,\boldsymbol{x})}{\pi_0(t,\boldsymbol{x})}u_{K_1}^\top(t)\mathbb{E}\left[u_{K_1}(T)\right]v_{K_2}^\top(\boldsymbol{X})v_{K_2}(\boldsymbol{x})dF_X(\boldsymbol{x})dF_T(t) - \int_{\mathcal{T}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{X};\boldsymbol{\beta}^*)dF_T(t)\right\|^2\right]\\
\le&\mathbb{E}\left[\left\|\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\frac{\pi_K^*(t,\boldsymbol{x})}{\pi_0(t,\boldsymbol{x})}u_{K_1}^\top(t)u_{K_1}(T^*)v_{K_2}^\top(\boldsymbol{X}^*)v_{K_2}(\boldsymbol{x})dF_X(\boldsymbol{x})dF_T(t) - m(T^*;\boldsymbol{\beta}^*)\varepsilon(T^*,\boldsymbol{X}^*;\boldsymbol{\beta}^*)\right\|^2\right]\\
\le&2\cdot\mathbb{E}\left[\left\|\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)u_{K_1}^\top(t)u_{K_1}(T^*)v_{K_2}^\top(\boldsymbol{X}^*)v_{K_2}(\boldsymbol{x})dF_X(\boldsymbol{x})dF_T(t) - m(T^*;\boldsymbol{\beta}^*)\varepsilon(T^*,\boldsymbol{X}^*;\boldsymbol{\beta}^*)\right\|^2\right]\\
&+2\cdot\mathbb{E}\left[\left\|\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)\frac{\pi_K^*(t,\boldsymbol{x})-\pi_0(t,\boldsymbol{x})}{\pi_0(t,\boldsymbol{x})}u_{K_1}^\top(t)u_{K_1}(T^*)v_{K_2}^\top(\boldsymbol{X}^*)v_{K_2}(\boldsymbol{x})dF_X(\boldsymbol{x})dF_T(t)\right\|^2\right]\to 0,
\end{aligned}$$

where $T^*\sim F_T$, $\boldsymbol{X}^*\sim F_X$, and T^* is independent of $\boldsymbol{X}^*$; the first inequality holds by Jensen's inequality; the last convergence result follows from Lemma 3.1 and the fact that

$$\int_{\mathcal{T}}\int_{\mathcal{X}} m(t;\boldsymbol{\beta}^*)\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*)u_{K_1}^\top(t)u_{K_1}(T^*)v_{K_2}^\top(\boldsymbol{X}^*)v_{K_2}(\boldsymbol{x})dF_X(\boldsymbol{x})dF_T(t)$$

is the L^2-projection of $m(T^*;\boldsymbol{\beta}^*)\varepsilon(T^*,\boldsymbol{X}^*;\boldsymbol{\beta}^*)$ on the space spanned by $\{u_{K_1}(T^*), v_{K_2}(\boldsymbol{X}^*)\}$. Thus (67) is of $o_p(1)$ by Chebyshev's inequality. Similar argument can be applied to show that both (68) and (69) are of $o_p(1)$. Therefore, we can have that

$$|(62)| \le |(66)| + |(67)| + |(68)| = o_p(1)\ .$$

Then, we can obtain that

$$|(49)| \le |(61)| + |(62)| = o_p(1)\ .$$

Summing up all orders (46)-(49) and using Assumption 1.8, we have

$$\begin{aligned}
&(46)+(47)+(48)+(49)\\
=&O(\sqrt{N}K^{-\alpha}) + 0 + \left\{O_p\left(\zeta(K)\cdot K^{\frac{1}{2}-\alpha}\right) + O_p\left(\zeta(K)\frac{K}{\sqrt{N}}\right)\right\} + o_p(1) = o_p(1).
\end{aligned}$$

5 Some Extensions

5.1 Proof of Theorem 7

(Consistency). Let

$$\widehat{\gamma} = \left[\sum_{i=1}^{N} u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\sum_{i=1}^{N} u_{K_1}(T_i)\hat{\pi}_K(T_i, \boldsymbol{X}_i)Y_i\right]$$

then $\widehat{\theta}_t = \hat{\gamma}^\top u_{K_1}(t)$. By assumption, there exists $\gamma^* \in \mathbb{R}^{K_1}$ such that

$$\sup_{t\in\mathcal{T}} \left|\theta_t - (\gamma^*)^\top u_{K_1}(t)\right| = O(K_1^{-\tilde{\alpha}}). \tag{70}$$

We first claim that

$$\|\hat{\gamma} - \gamma^*\| = O_p\left(\zeta(K)\left\{\sqrt{\frac{K}{N}} + K^{-\alpha}\right\} + K_1^{-\tilde{\alpha}}\right), \tag{71}$$

whose proof will be established later. Using (70) and (71), we can deduce that

$$\begin{aligned}
&\int_{\mathcal{T}}\left[\widehat{\theta}_t - \theta_t\right]^2 dF_T(t)\\
=&\int_{\mathcal{T}}[\hat{\gamma}^\top u_{K_1}(t) - (\gamma^*)^\top u_{K_1}(t) + (\gamma^*)^\top u_{K_1}(t) - \theta_t]^2 dF_T(t)\\
\leq& 2(\hat{\gamma}-\gamma^*)^\top\left[\int_{\mathcal{T}} u_{K_1}(t)u_{K_1}(t)^\top dF_T(t)\right](\hat{\gamma}-\gamma^*) + 2\int_{\mathcal{T}}[(\gamma^*)^\top u_{K_1}(t) - \theta_t]^2 dF_T(t)\\
\leq& 2\|\hat{\gamma}-\gamma^*\|^2\cdot\lambda_{\max}\left(\mathbb{E}[u_{K_1}(T)u_{K_1}(T)^\top]\right) + 2\sup_{t\in\mathcal{T}}|(\gamma^*)^\top u_{K_1}(t) - \theta_t|^2\\
=& O_p\left(\zeta(K)^2\left\{\frac{K}{N} + K^{-2\alpha}\right\} + K_1^{-2\tilde{\alpha}}\right),
\end{aligned}$$

and

$$\begin{aligned}
\sup_{t\in\mathcal{T}}|\hat{\theta}_t - \theta_t| =& \sup_{t\in\mathcal{T}}\left|\hat{\gamma}^\top u_{K_1}(t) - (\gamma^*)^\top u_{K_1}(t) + (\gamma^*)^\top u_{K_1}(t) - \theta_t\right|\\
\leq& \sup_{t\in\mathcal{T}}\|u_{K_1}(t)\|\cdot\|\hat{\gamma}-\gamma^*\| + \sup_{t\in\mathcal{T}}|(\gamma^*)^\top u_{K_1}(t) - \theta_t|\\
\leq& O_p\left[\zeta_1(K_1)\left(\zeta(K)\left\{\sqrt{\frac{K}{N}} + K^{-\alpha}\right\} + K_1^{-\tilde{\alpha}}\right)\right] + O(K_1^{-\tilde{\alpha}})
\end{aligned}$$

$$=O_p\left[\zeta_1(K_1)\left(\zeta(K)\left\{\sqrt{\frac{K}{N}}+K^{-\alpha}\right\}+K_1^{-\tilde{\alpha}}\right)\right].$$

Finally, we come back to prove (71). Note that

$$\begin{aligned}
\hat{\gamma}-\gamma^* &= \left[\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\sum_{i=1}^N \hat{\pi}_K(T_i,\boldsymbol{X}_i)u_{K_1}(T_i)Y_i\right]-\gamma^* \\
&= \left[\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\sum_{i=1}^N u_{K_1}(T_i)\left\{\hat{\pi}_K(T_i,\boldsymbol{X}_i)-\pi_0(T_i,\boldsymbol{X}_i)\right\}Y_i\right] \\
&\quad+\left[\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\sum_{i=1}^N u_{K_1}(T_i)\{\pi_0(T_i,\boldsymbol{X}_i)Y_i-\mathbb{E}[\pi_0(T_i,\boldsymbol{X}_i)Y_i|T_i]\}\right] \\
&\quad+\left[\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\sum_{i=1}^N u_{K_1}(T_i)\left\{\mathbb{E}[\pi_0(T_i,\boldsymbol{X}_i)Y_i|T_i]-(\gamma^*)^\top u_{K_1}(T_i)\right\}\right] \\
&\equiv A_{1N}+A_{2N}+A_{3N}.
\end{aligned}$$

We first compute the probability order of A_{1N}. We use the following notation:

$$\begin{aligned}
\hat{H}_N &:= \left(\left\{\hat{\pi}_K(T_1,X_1)-\pi_0(T_1,X_1)\right\}Y_1,...,\left\{\hat{\pi}_K(T_N,X_N)-\pi_0(T_N,X_N)\right\}Y_N\right)^\top, \\
U_{N\times K_1} &:= \left(u_{K_1}(T_1),...,u_{K_1}(T_N)\right)^\top, \\
\hat{\Phi}_{K_1\times K_1} &:= \frac{1}{N}\sum_{i=1}^N u_{K_1}(T)u_{K_1}^\top(T).
\end{aligned}$$

Then we can obtain that

$$\begin{aligned}
\|A_{1N}\|^2 &= \left\|\left[\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\sum_{i=1}^N u_{K_1}(T_i)\left\{\hat{\pi}_K(T_i,\boldsymbol{X}_i)-\pi_0(T_i,\boldsymbol{X}_i)\right\}Y_i\right]\right\|^2 \\
&= N^{-2}\mathrm{tr}\left(\hat{\Phi}_{K_1\times K_1}^{-1}U_{N\times K_1}^\top\hat{H}_N\hat{H}_N^\top U_{N\times K_1}\hat{\Phi}_{K_1\times K_1}^{-1}\right) \\
&= N^{-2}\mathrm{tr}\left(U_{N\times K_1}^\top\hat{H}_N\hat{H}_N^\top U_{N\times K_1}\hat{\Phi}_{K_1\times K_1}^{-1}\hat{\Phi}_{K_1\times K_1}^{-1}\right) \\
&= N^{-2}\mathrm{tr}\left(\hat{\Phi}_{K_1\times K_1}^{-1/2}U_{N\times K_1}^\top\hat{H}_N\hat{H}_N^\top U_{N\times K_1}\hat{\Phi}_{K_1\times K_1}^{-1/2}\hat{\Phi}_{K_1\times K_1}^{-1}\right) \\
&\le \lambda_{\max}(\hat{\Phi}_{K_1\times K_1}^{-1})N^{-2}\mathrm{tr}\left(\hat{\Phi}_{K_1\times K_1}^{-1/2}U_{N\times K_1}^\top\hat{H}_N\hat{H}_N^\top U_{N\times K_1}\hat{\Phi}_{K_1\times K_1}^{-1/2}\right) \\
&= \lambda_{\max}(\hat{\Phi}_{K_1\times K_1}^{-1})N^{-1}\mathrm{tr}\left(\hat{H}_N\hat{H}_N^\top U_{N\times K_1}(U_{N\times K_1}^\top U_{N\times K_1})^{-1}U_{N\times K_1}^\top\right)
\end{aligned}$$

$$
\begin{aligned}
\leq & [\lambda_{\min}(\hat{\Phi}_{K_1\times K_1})]^{-1} N^{-1} \|\hat{H}_N\|^2 \\
= & [\lambda_{\min}(\hat{\Phi}_{K_1\times K_1})]^{-1} \cdot \frac{1}{N} \sum_{i=1}^{N} \left\{\hat{\pi}_K(T_i, \boldsymbol{X}_i) - \pi_0(T_i, \boldsymbol{X}_i)\right\}^2 Y_i^2 \\
\leq & [\lambda_{\min}(\hat{\Phi}_{K_1\times K_1})]^{-1} \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |\hat{\pi}_K(t,x) - \pi_0(t,x)|^2 \cdot \frac{1}{N} \sum_{i=1}^{N} Y_i^2 \\
\leq & O_p(1) \cdot O_p\left(\zeta(K)^2 K^{-2\alpha} + \frac{\zeta(K)^2 K}{N}\right) \cdot O_p(1) \\
= & O_p\left(\zeta(K)^2 K^{-2\alpha} + \frac{\zeta(K)^2 K}{N}\right), \qquad (72)
\end{aligned}
$$

where the first inequality follows from the fact that $tr(AB) \leq \lambda_{\max}(B) tr(A)$ for any symmetric matrix B and positive semidefinite matrix A, the second inequality follows from the same fact and the fact that $U_{N\times K_1}(U_{N\times K_1}^\top U_{N\times K_1})^{-1} U_{N\times K_1}^\top$ is a projection matrix with maximum eigenvalue 1, and the fourth inequality follows from the facts that $|\lambda_{\min}(\hat{\Phi}_{K_1\times K_1})|^{-1} = O_p(1)$, Lemma 3.1 and Corollary 3.3, and $N^{-1}\sum_{i=1}^N Y_i^2 = O_p(1)$.

Next, we compute the probability order of A_{2N}. Let

$$
\varepsilon_i := \pi_0(T_i, \boldsymbol{X}_i) Y_i - \mathbb{E}[\pi_0(T_i, \boldsymbol{X}_i) Y_i | T_i] \text{ and } \mathcal{E}_N := (\varepsilon_1, \ldots, \varepsilon_N)^\top.
$$

We can deduce that

$$
\begin{aligned}
\|A_{2N}\|^2 = & \left\| \left[\sum_{i=1}^{N} u_{K_1}(T_i) u_{K_1}(T_i)^\top\right]^{-1} \left[\sum_{i=1}^{N} u_{K_1}(T_i)\varepsilon_i\right] \right\|^2 \\
= & N^{-2} \mathrm{tr}\left(\hat{\Phi}_{K_1\times K_1}^{-1} U_{N\times K_1}^\top \mathcal{E}_N \mathcal{E}_N^\top U_{N\times K_1} \hat{\Phi}_{K_1\times K_1}^{-1}\right) \\
= & N^{-2} \mathrm{tr}\left(U_{N\times K_1}^\top \mathcal{E}_N \mathcal{E}_N^\top U_{N\times K_1} \hat{\Phi}_{K_1\times K_1}^{-1} \hat{\Phi}_{K_1\times K_1}^{-1}\right) \\
\leq & [\lambda_{\min}(\hat{\Phi}_{K_1\times K_1})]^{-2} N^{-2} \|U_{N\times K_1}^\top \mathcal{E}_N\|^2 = O_p\left(\frac{K_1}{N}\right),
\end{aligned}
$$

where the last equality follows that $|\lambda_{\min}(\hat{\Phi}_{K_1\times K_1})|^{-1} = O_p(1)$ and $N^{-2}\|U_{N\times K_1}^\top \mathcal{E}_N\|^2 = O_p(K_1/N)$ by Markov's inequality.

We finally compute the probability order of A_{3N}. Let

$$R_N(\gamma^*) = \left(\left\{ \mathbb{E}[\pi_0(T_1, X_1)Y_1|T_1] - (\gamma^*)^\top u_{K_1}(T_1) \right\}, \ldots, \left\{ \mathbb{E}[\pi_0(T_N, X_N)Y_N|T_N] - (\gamma^*)^\top u_{K_1}(T_N) \right\} \right)^\top,$$

then

$$\begin{aligned}
\|A_{3N}\|^2 =& \left\| \left[\sum_{i=1}^N u_{K_1}(T_i) u_{K_1}(T_i)^\top \right]^{-1} \left[\sum_{i=1}^N u_{K_1}(T_i) \left\{ \mathbb{E}[\pi_0(T_i, \boldsymbol{X}_i)Y_i|T_i] - (\gamma^*)^\top u_{K_1}(T_i) \right\} \right] \right\|^2 \\
=& N^{-2} \left\| \hat{\Phi}_{K_1\times K_1}^{-1} U_{N\times K_1}^\top R_N(\gamma^*) \right\|^2 \\
=& N^{-2} \mathrm{tr} \left(\hat{\Phi}_{K_1\times K_1}^{-1} U_{N\times K_1}^\top R_N(\gamma^*) R_N(\gamma^*)^\top U_{N\times K_1} \hat{\Phi}_{K_1\times K_1}^{-1} \right) \\
=& N^{-2} \mathrm{tr} \left(U_{N\times K_1}^\top R_N(\gamma^*) R_N(\gamma^*)^\top U_{N\times K_1} \hat{\Phi}_{K_1\times K_1}^{-1} \hat{\Phi}_{K_1\times K_1}^{-1} \right) \\
=& N^{-2} \mathrm{tr} \left(\hat{\Phi}_{K_1\times K_1}^{-1/2} U_{N\times K_1}^\top R_N(\gamma^*) R_N(\gamma^*)^\top U_{N\times K_1} \hat{\Phi}_{K_1\times K_1}^{-1/2} \hat{\Phi}_{K_1\times K_1}^{-1} \right) \\
\leq& \lambda_{\max}(\hat{\Phi}_{K_1\times K_1}^{-1}) N^{-2} \mathrm{tr} \left(\hat{\Phi}_{K_1\times K_1}^{-1/2} U_{N\times K_1}^\top R_N(\gamma^*) R_N(\gamma^*)^\top U_{N\times K_1} \hat{\Phi}_{K_1\times K_1}^{-1/2} \right) \\
=& \lambda_{\max}(\hat{\Phi}_{K_1\times K_1}^{-1}) N^{-1} \mathrm{tr} \left(R_N(\gamma^*) R_N(\gamma^*)^\top U_{N\times K_1} (U_{N\times K_1}^\top U_{N\times K_1})^{-1} U_{N\times K_1}^\top \right) \\
\leq& [\lambda_{\min}(\hat{\Phi}_{K_1\times K_1})]^{-1} N^{-1} \|R_N(\gamma^*)\|^2 \\
=& [\lambda_{\min}(\hat{\Phi}_{K_1\times K_1})]^{-1} \cdot \frac{1}{N} \sum_{i=1}^N \left\{ \mathbb{E}[\pi_0(T_i, \boldsymbol{X}_i)Y_i|T_i] - (\gamma^*)^\top u_{K_1}(T_i) \right\}^2 = O_p(K_1^{-2\tilde{\alpha}}),
\end{aligned}$$

where the first inequality follows from the fact that $\mathrm{tr}(AB) \leq \lambda_{\max}(B)\mathrm{tr}(A)$ for any symmetric matrix B and positive semidefinite matrix A, the second inequality follows from the same fact and the fact that $U_{N\times K_1}(U_{N\times K_1}^\top U_{N\times K_1})^{-1}U_{N\times K_1}^\top$ is a projection matrix with maximum eigenvalue 1, and the last equality follows from the fact that $|\lambda_{\min}(\hat{\Phi}_{K_1\times K_1})|^{-1} = O_p(1)$ and the fact that $\frac{1}{N}\sum_{i=1}^N \left\{ \mathbb{E}[\pi_0(T_i, \boldsymbol{X}_i)Y_i|T_i] - (\gamma^*)^\top u_{K_1}(T_i) \right\}^2 \leq \sup_{t\in\mathcal{T}} |\mathbb{E}[\pi_0(T, X)Y|T = t] - (\gamma^*)^\top u_{K_1}(t)|^2 = O(K_1^{-2\tilde{\alpha}})$. Hence, we complete the proof of (71).

(Asymptotic Normality). We have the following decomposition for $\widehat{\theta}_t - \theta(t)$:

$$\begin{aligned}
&\hat{\theta}_t - \theta_t = u_{K_1}(t)^\top(\hat{\gamma} - \gamma^*) + [(\gamma^*)^\top u_{K_1}(t) - \theta_t] \\
=& u_{K_1}(t)^\top \left[\frac{1}{N} \sum_{i=1}^N u_{K_1}(T_i) u_{K_1}(T_i)^\top \right]^{-1} \left[\frac{1}{N} \sum_{i=1}^N u_{K_1}(T_i) \left\{ \hat{\pi}_K(T_i, \boldsymbol{X}_i)Y_i - \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i)Y_i|T_i\right] \right\} \right]
\end{aligned}$$

$$
\begin{aligned}
&+u_{K_1}(t)^\top \left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\cdot\left\{\mathbb{E}\left[\pi_0(T_i,\boldsymbol{X}_i)Y_i|T_i\right]-(\gamma^*)^\top u_{K_1}(T_i)\right\}\right]\\
&\quad+\left[(\gamma^*)^\top u_{K_1}(t)-\theta_t\right]\\
\equiv& b_{1N}(t)+b_{2N}(t)+b_{3N}(t).
\end{aligned}
$$

We shall show that $b_{1N}(t)$ contributes to the asymptotic variance; and $b_{2N}(t)+b_{3N}(t)$ contributes to the asymptotic bias which is asymptotically negligible. Thus to complete the proof of asymptotic normality, it is sufficient to prove the following results:

(i) $V_t \geq c\|u_{K_1}(t)\|^2$ for some $c>0$;

(ii) $\sqrt{N}V_t^{-1/2}b_{1N}(t)\xrightarrow{d}N(0,1)$;

(iii) $\sqrt{N}V_t^{-1/2}b_{2N}(t)=o_p(1)$;

(iv) $\sqrt{N}V_t^{-1/2}b_{3N}(t)=o_p(1)$.

We first prove Result (i). By assumption, $\lambda_{\min}\left(\mathbb{E}\left[b_{K_1}(T,\boldsymbol{X},Y)b_{K_1}^\top(T,\boldsymbol{X},Y)\right]\right)\geq \underline{c}$, we have

$$
\begin{aligned}
V_t =&u_{K_1}^\top(t)\Phi_{K_1\times K_1}^{-1}\mathbb{E}\left[b_{K_1}(T,\boldsymbol{X},Y)b_{K_1}^\top(T,\boldsymbol{X},Y)\right]\Phi_{K_1\times K_1}^{-1}u_{K_1}(t)\\
\geq&\underline{c}\cdot u_{K_1}^\top(t)\Phi_{K_1\times K_1}^{-1}\Phi_{K_1\times K_1}^{-1}u_{K_1}(t)\\
\geq&\underline{c}\cdot\lambda_{\min}^2\left(\Phi_{K_1\times K_1}^{-1}\right)\|u_{K_1}(t)\|^2.
\end{aligned}
$$

For the claim (ii). Let

$$
\tilde{b}_{1N}(t)=u_{K_1}(t)^\top\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^N b_{K_1}(T_i,\boldsymbol{X}_i,Y_i)\right].
$$

Similar to the proof of (40), we can show

$$
\sqrt{N}V_t^{-1/2}\cdot\left(b_{1N}(t)-\tilde{b}_{1N}(t)\right)=o_p(1)\ .
$$

Then

$$
\begin{aligned}
&\sqrt{N}V_t^{-1/2}b_{1N}(t)=\sqrt{N}V_t^{-1/2}\tilde{b}_{1N}(t)+o_p(1)\\
=&\sqrt{N}V_t^{-1/2}u_{K_1}(t)^\top\hat{\Phi}_{K_1\times K_1}^{-1}N^{-1}\sum_{i=1}^N b_{K_1}(T_i,\boldsymbol{X}_i,Y_i)
\end{aligned}
$$

$$
\begin{aligned}
=&\sqrt{N}V_t^{-1/2}u_{K_1}(t)^\top \Phi_{K_1\times K_1}^{-1}\cdot N^{-1}\sum_{i=1}^N b_{K_1}(T_i,\boldsymbol{X}_i,Y_i)\\
&+\sqrt{N}V_t^{-1/2}u_{K_1}(t)^\top\left[\hat{\Phi}_{K_1\times K_1}^{-1}-\Phi_{K_1\times K_1}^{-1}\right]\cdot N^{-1}\sum_{i=1}^N b_{K_1}(T_i,\boldsymbol{X}_i,Y_i)\\
\equiv& b_{1N}^{(1)}(t)+b_{1N}^{(2)}(t). \qquad (73)
\end{aligned}
$$

For $b_{1N}^{(1)}(t)$, we can simply apply the Liapounov CLT and show that $b_{1N}^{(1)}(t)\xrightarrow{d}N(0,1)$. For $b_{1N}^{(2)}(t)$, we can deduce that

$$
\begin{aligned}
|b_{1N,2}^{(2)}(t)|^2\leq&\left\{V_K^{-1}\|u_{K_1}(t)\|^2\right\}\cdot\left\|\hat{\Phi}_{K_1\times K_1}^{-1}-\Phi_{K_1\times K_1}^{-1}\right\|^2\cdot\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N b_{K_1}(T_i,\boldsymbol{X}_i,Y_i)\right\|^2\\
\leq&O_p(1)\cdot O_P\left(\zeta_1(K_1)^2\cdot\frac{K_1}{N}\right)\cdot O_P(K_1)=O_P\left(\zeta_1(K_1)^2\cdot\frac{K_1^2}{N}\right)=o_P(1),
\end{aligned}
$$

where the second inequality by noting the following facts

$$
\begin{aligned}
&\left\|\hat{\Phi}_{K_1\times K_1}^{-1}-\Phi_{K_1\times K_1}^{-1}\right\|^2\\
=&\mathrm{tr}\left(\left\{\hat{\Phi}_{K_1\times K_1}^{-1}-\Phi_{K_1\times K_1}^{-1}\right\}\left\{\hat{\Phi}_{K_1\times K_1}^{-1}-\Phi_{K_1\times K_1}^{-1}\right\}\right)\\
=&\mathrm{tr}\left(\hat{\Phi}_{K_1\times K_1}^{-1}\left\{\hat{\Phi}_{K_1\times K_1}-\Phi_{K_1\times K_1}\right\}\Phi_{K_1\times K_1}^{-1}\Phi_{K_1\times K_1}^{-1}\left\{\hat{\Phi}_{K_1\times K_1}-\Phi_{K_1\times K_1}\right\}\hat{\Phi}_{K_1\times K_1}^{-1}\right)\\
=&\mathrm{tr}\left(\left\{\hat{\Phi}_{K_1\times K_1}-\Phi_{K_1\times K_1}\right\}\Phi_{K_1\times K_1}^{-1}\Phi_{K_1\times K_1}^{-1}\left\{\hat{\Phi}_{K_1\times K_1}-\Phi_{K_1\times K_1}\right\}\hat{\Phi}_{K_1\times K_1}^{-1}\hat{\Phi}_{K_1\times K_1}^{-1}\right)\\
\leq&\lambda_{\min}\left(\hat{\Phi}_{K_1\times K_1}\right)^{-2}\lambda_{\min}\left(\Phi_{K_1\times K_1}\right)^{-2}\cdot\mathrm{tr}\left(\left\{\hat{\Phi}_{K_1\times K_1}-\Phi_{K_1\times K_1}\right\}\left\{\hat{\Phi}_{K_1\times K_1}-\Phi_{K_1\times K_1}\right\}\right)\\
\leq&O_p(1)\cdot O_p(1)\cdot O_P\left(\zeta_1(K_1)^2\cdot\frac{K_1}{N}\right)=O_P\left(\zeta_1(K_1)^2\cdot\frac{K_1}{N}\right),
\end{aligned}
$$

and

$$
\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N b_{K_1}(T_i,\boldsymbol{X}_i,Y_i)\right\|^2\right]=\mathbb{E}\left[\|b_{K_1}(T,\boldsymbol{X},Y)\|^2\right]=O(K_1).
$$

Thus (ii) holds.

For (iii), by Cauchy-Schwarz's inequality, we can obtain that

$$
\sqrt{N}V_t^{-1/2}|b_{2N}(t)|
$$

$$
\begin{aligned}
=&N^{-1/2}V_t^{-1/2}\left|u_{K_1}(t)^\top \hat{\Phi}_{K_1\times K_1}^{-1} U_{N\times K_1}^\top R_N(\gamma^*)\right| \\
\leq&V_t^{-1/2}\left\{u_{K_1}(t)^\top \hat{\Phi}_{K_1\times K_1}^{-1}\left(N^{-1}U_{N\times K_1}^\top U_{N\times K_1}\right)\hat{\Phi}_{K_1\times K_1}^{-1}u_{K_1}(t)\right\}^{\frac{1}{2}}\left\{R_N(\gamma^*)^\top R_N(\gamma^*)\right\}^{\frac{1}{2}} \\
\leq&V_t^{-1/2}\left\{u_{K_1}(t)^\top \hat{\Phi}_{K_1\times K_1}^{-1}u_{K_1}(t)\right\}^{\frac{1}{2}}\left\{R_N(\gamma^*)^\top R_N(\gamma^*)\right\}^{\frac{1}{2}} \\
\leq&\{V_t^{-1/2}\|u_{K_1}(t)\|\}\cdot|\lambda_{\max}(\Phi_{K_1\times K_1}^{-1})|^{\frac{1}{2}}\cdot O(\sqrt{N}\cdot K_1^{-\tilde{\alpha}}) \\
=&O(1)\cdot O_p(1)\cdot o_p(1)=o_p(1)\ .
\end{aligned}
$$

Similarly, we can show show that $\sqrt{N}V_t^{-1/2}|b_{3N}(t)| = o_p(1)$. This completes the proof of the Theorem.

5.2 Proof of Theorem 9

Note that

$$
\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot\widehat{\theta}_{t_0,t_1|t_0}=\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot\widehat{\theta}_{t_1|t_0}-\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot\widehat{\theta}_{t_0|t_0}.
$$

Consider the term $\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot\widehat{\theta}_{t_0|t_0}$. Since $\widehat{\theta}_{t_0|t_0}$ is a nonparametric series estimator of $\theta_{t_0|t_0}$, by using a similar argument of proving Theorem 6 (see also Newey (1997)), we have

$$
\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot\widehat{\theta}_{t_0|t_0}=V_{t_1,t_0|t_0}^{-1/2}\cdot u_{K_1}(t_0)^\top\Phi_{K_1\times K_1}^{-1}\cdot\frac{1}{\sqrt{N}}\sum_{i=1}^N b_{3,K_1}(T_i,Y_i)+o_P(1), \tag{74}
$$

where

$$
b_{3,K_1}(T_i,Y_i)=u_{K_1}(T_i)\{Y_i-\mathbb{E}[Y_i|T_i]\}.
$$

Consider the term $\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot\widehat{\theta}_{t_1|t_0}$. Let $\delta:=t_1-t_0$, and

$$
\widehat{\gamma}:=\left[\sum_{i=1}^N\frac{\widehat{\pi}_K(T_i,\boldsymbol{X}_i)}{\widehat{\pi}_K(T_i-\delta,\boldsymbol{X}_i)}\cdot Y_i\cdot u_{K_1}^\top(T_i)\right]\left[\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}^\top(T_i)\right]^{-1}.
$$

Then $\widehat{\theta}_{t_1|t_0}=\widehat{\gamma}^\top u_{K_1}(t_1)$. We have the following decomposition for $\widehat{\theta}_{t_1|t_0}-\theta_{t_1|t_0}$:

$$
\begin{aligned}
&\widehat{\theta}_{t_1|t_0}-\theta_{t_1|t_0}=u_{K_1}(t_1)^\top(\hat{\gamma}-\gamma^*)+[(\gamma^*)^\top u_{K_1}(t_1)-\theta_{t_1|t_0}] \\
=&u_{K_1}(t_1)^\top\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\left\{\frac{\hat{\pi}_K(T_i,\boldsymbol{X}_i)}{\hat{\pi}_K(T_i-\delta,\boldsymbol{X}_i)}Y_i-\frac{\hat{\pi}_K(T_i,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}Y_i\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+u_{K_1}(t_1)^\top \left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\left\{\frac{\hat\pi_K(T_i,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}Y_i-\mathbb{E}\left[\frac{\pi_0(T_i,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}Y_i\middle|T_i\right]\right\}\right]\\
&+u_{K_1}(t_1)^\top \left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)u_{K_1}(T_i)^\top\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\cdot\left\{\mathbb{E}\left[\frac{\pi_0(T_i,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}Y_i\middle|T_i\right]-(\gamma^*)^\top u_{K_1}(T_i)\right\}\right]\\
&\quad+\left[(\gamma^*)^\top u_{K_1}(t_1)-\theta_{t_1|t_0}\right]\\
\equiv& b_{1N}(t_1)+b_{2N}(t_1)+b_{3N}(t_1)+b_{4N}(t_1).
\end{aligned}
$$

Similar to the proof of Theorem 7 (pp 50, (iii) and (iv)), we can show

$$
\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}|b_{3N}(t_1)|\to 0 \text{ and } \sqrt{N}V_{t_1,t_0|t_0}^{-1/2}|b_{4N}(t_1)|\to 0. \tag{75}
$$

Consider $b_{1N}(t_1)$. Similar to (73), we can show that

$$
\begin{aligned}
&\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}(t_1)\\
=&-\sqrt{N}u_{K_1}(t_1)^\top\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\left\{\frac{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)}\right\}\hat\pi_K(T_i,\boldsymbol{X}_i)Y_i\right]+o_P(1).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}(t_1)\\
=&-\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot u_{K_1}(t_1)^\top\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\left\{\frac{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)^2}\right\}\pi_0(T_i,\boldsymbol{X}_i)Y_i\right]\\
&-\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot u_{K_1}(t_1)^\top\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\left\{\frac{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}\right\}\right.\\
&\qquad\left.\times\left\{\frac{\hat\pi_K(T_i,\boldsymbol{X}_i)}{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)}-\frac{\pi_0(T_i,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}\right\}Y_i\right]\\
=&-\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot u_{K_1}(t_1)^\top\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\left\{\frac{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)^2}\right\}\pi_0(T_i,\boldsymbol{X}_i)Y_i\right]\\
&-\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot u_{K_1}(t_1)^\top\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\left\{\frac{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}\right\}\right.\\
&\qquad\left.\times\left\{\frac{\hat\pi_K(T_i,\boldsymbol{X}_i)-\pi_0(T_i,\boldsymbol{X}_i)}{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)}\right\}Y_i\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot u_{K_1}(t_1)^\top\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}u_{K_1}(T_i)\left\{\frac{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)}{\pi_0(t_0,\boldsymbol{X}_i)}\right\}\right.\\
&\qquad\left.\cdot\left\{\frac{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,X_i)\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)}\right\}\pi_0(T_i,\boldsymbol{X}_i)Y_i\right]\\
\equiv&\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}^{(1)}(t_1)+\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}^{(2)}(t_1)+\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}^{(3)}(t_1).
\end{aligned}
$$

Consider $\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}^{(2)}(t_1)$. The conditions $\lambda_{\min}(\Sigma_{2K_1\times 2K_1})>\underline{c}>0$ and $\lambda_{\min}(\Phi_{K_1\times K_1})>\underline{c}>0$ imply $V_{t_1,t_0|t_0}^{-1}\geq c\cdot\|u_{K_1}(t_1)\|^2$ for some $c>0$. Similar to (72), we can show that

$$
\begin{aligned}
&\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\left|b_{1N}^{(2)}(t_1)\right|\leq\sqrt{N}\left\{V_{t_1,t_0|t_0}^{-1/2}\cdot\|u_{K_1}(t_1)\|\right\}\\
&\cdot\left\|\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}u_{K_1}(T_i)\cdot\left\{\frac{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}\right\}\left\{\frac{\hat\pi_K(T_i,\boldsymbol{X}_i)-\pi_0(T_i,\boldsymbol{X}_i)}{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)}\right\}Y_i\right]\right\|\\
\leq&\sqrt{N}\cdot O_P(1)\cdot\left\{\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)}{\pi_0(T_i-\delta,\boldsymbol{X}_i)}\right\}^2\cdot\left\{\frac{\hat\pi_K(T_i,\boldsymbol{X}_i)-\pi_0(T_i,\boldsymbol{X}_i)}{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)}\right\}^2Y_i^2\right\}^{1/2}\\
\leq&\sqrt{N}\cdot O_P(1)\cdot\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}}|\hat\pi_K(t,\boldsymbol{x})-\pi_0(t,\boldsymbol{x})|^2\cdot\left\{\frac{1}{N}\sum_{i=1}^{N}Y_i^2\right\}^{1/2}\\
\leq&O_P\left(\sqrt{N}\cdot\zeta^2(K)\cdot\left\{K^{-2\alpha}+\frac{K}{N}\right\}\right)=o_P(1).
\end{aligned}
\tag{76}
$$

Similarly, we can also show

$$
\sqrt{N}\cdot V_{t_1,t_0|t_0}^{-1/2}\cdot\left|b_{1N}^{(3)}(t_1)\right|=o_P(1).
\tag{77}
$$

We next consider $b_{1N}^{(1)}(t_1)$. We shall find the influence representation for $N^{-1/2}u_{K_1}(t_1)\Phi_{K_1\times K_1}^{-1}\cdot\sum_{i=1}^{N}u_{K_1}(T_i)\{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_0(T_i-\delta,\boldsymbol{X}_i)\}\pi_0(T_i,\boldsymbol{X}_i)Y_i/\pi_0(t_0,\boldsymbol{X}_i)^2$. To achieve this goal, we consider the asymptotic behavior of $N^{-1/2}\sum_{i=1}^{N}\{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)\phi(T_i,\boldsymbol{X}_i,Y_i)-\mathbb{E}[\pi_0(T-\delta,\boldsymbol{X})\phi(T,\boldsymbol{X},Y)]\}$, where $\phi(T,\boldsymbol{X},Y)$ denotes a general L^2 random variable. Define $\mu(t,\boldsymbol{x}):=\mathbb{E}\left[\phi(T,\boldsymbol{X},Y)|T=t,\boldsymbol{X}=\boldsymbol{x}\right]$. Similar to the proof of (40) in Section 4.3, we have the following decomposition:

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)\phi(T_i,\boldsymbol{X}_i,Y_i)-\mathbb{E}\left[\pi_0(T-\delta,\boldsymbol{X})\phi(T,\boldsymbol{X},Y)\right]\right\}\\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\left(\hat\pi_K(T_i-\delta,\boldsymbol{X}_i)-\pi_K^*(T_i-\delta,\boldsymbol{X}_i)\right)\phi(T_i,\boldsymbol{X}_i,Y_i)\right.
\end{aligned}
\tag{78}
$$

$$
\begin{aligned}
&\quad - \int_{\mathcal{T}} \int_{\mathcal{X}} \left(\hat{\pi}_K(t-\delta, \boldsymbol{x}) - \pi_K^*(t-\delta, \boldsymbol{x})\right) \mu(\boldsymbol{x}, t) dF_{X,T}(\boldsymbol{x}, t) \Bigg\} \\
&+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Bigg\{ \left(\pi_K^*(T_i - \delta, \boldsymbol{X}_i) - \pi_0(T_i - \delta, \boldsymbol{X}_i)\right) \phi(T_i, \boldsymbol{X}_i, Y_i) \qquad (79) \\
&\qquad - \int_{\mathcal{T}} \int_{\mathcal{X}} \mu(t, \boldsymbol{x}) \left(\pi_K^*(t-\delta, \boldsymbol{x}) - \pi_0(t-\delta, \boldsymbol{x})\right) dF_{X,T}(\boldsymbol{x}, t) \Bigg\} \\
&+ \sqrt{N} \int_{\mathcal{T}} \int_{\mathcal{X}} \mu(t, \boldsymbol{x}) \left(\pi_K^*(t-\delta, \boldsymbol{x}) - \pi_0(t-\delta, \boldsymbol{x})\right) dF_{X,T}(\boldsymbol{x}, t) \qquad (80) \\
&+ \sqrt{N} \int_{\mathcal{T}} \int_{\mathcal{X}} \left(\hat{\pi}_K(t-\delta, \boldsymbol{x}) - \pi_K^*(t-\delta, \boldsymbol{x})\right) \mu(\boldsymbol{x}, t) dF_{X,T}(\boldsymbol{x}, t) \qquad (81) \\
&\quad - \sqrt{N} \int_{\mathcal{T}} \int_{\mathcal{X}} \mu(t, \boldsymbol{x}) \rho'' \left(u_{K_1}^\top (t-\delta) \tilde{\Lambda}_{K_1 \times K_2} v_{K_2}(\boldsymbol{x})\right) u_{K_1}^\top (t-\delta) \hat{A}_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x}, t) \\
&+ \sqrt{N} \int_{\mathcal{T}} \int_{\mathcal{X}} \mu(t, \boldsymbol{x}) \rho'' \left(u_{K_1}^\top (t-\delta) \tilde{\Lambda}_{K_1 \times K_2} v_{K_2}(\boldsymbol{x})\right) u_{K_1}^\top (t-\delta) \hat{A}_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x}, t) \qquad (82) \\
&\quad - \sqrt{N} \int_{\mathcal{T}} \int_{\mathcal{X}} \mu(t, \boldsymbol{x}) \rho'' \left(u_{K_1}^\top (t-\delta) \Lambda_{K_1 \times K_2}^* v_{K_2}(\boldsymbol{x})\right) u_{K_1}^\top (t-\delta) A_{K_1 \times K_2}^* v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x}, t) \\
&+ \sqrt{N} \int_{\mathcal{X}} \mu(t, \boldsymbol{x}) \rho'' \left(u_{K_1}^\top (t-\delta) \Lambda_{K_1 \times K_2}^* v_{K_2}(\boldsymbol{x})\right) u_{K_1}^\top (t-\delta) A_{K_1 \times K_2}^* v_{K_2}(\boldsymbol{x}) dF_{X,T}(\boldsymbol{x}, t) \qquad (83) \\
&\quad + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Bigg\{ \pi_0(T_i, \boldsymbol{X}_i) \frac{f_{T|X}(T_i + \delta | \boldsymbol{X}_i)}{f_{T|X}(T_i | \boldsymbol{X}_i)} \mu(T_i + \delta, \boldsymbol{X}_i) \\
&\qquad - \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) \frac{f_{T|X}(T_i + \delta | \boldsymbol{X}_i)}{f_{T|X}(T_i | \boldsymbol{X}_i)} \mu(T_i + \delta, \boldsymbol{X}_i) \middle| \boldsymbol{X}_i \right] \\
&\qquad - \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) \frac{f_{T|X}(T_i + \delta | \boldsymbol{X}_i)}{f_{T|X}(T_i | \boldsymbol{X}_i)} \mu(T_i + \delta, \boldsymbol{X}_i) \middle| T_i \right] \\
&\qquad + \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) \frac{f_{T|X}(T_i + \delta | \boldsymbol{X}_i)}{f_{T|X}(T_i | \boldsymbol{X}_i)} \mu(T_i + \delta, \boldsymbol{X}_i)\right] \\
&\quad + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Bigg\{ \pi_0(T_i - \delta, \boldsymbol{X}_i) \phi(T_i, \boldsymbol{X}_i, Y_i) - \pi_0(T_i, \boldsymbol{X}_i) \frac{f_{T|X}(T_i + \delta | \boldsymbol{X}_i)}{f_{T|X}(T_i | \boldsymbol{X}_i)} \mu(T_i + \delta, \boldsymbol{X}_i) \qquad (84) \\
&\qquad + \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) \frac{f_{T|X}(T_i + \delta | \boldsymbol{X}_i)}{f_{T|X}(T_i | \boldsymbol{X}_i)} \mu(T_i + \delta, \boldsymbol{X}_i) \middle| \boldsymbol{X}_i \right] - \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) \frac{f_{T|X}(T_i + \delta | \boldsymbol{X}_i)}{f_{T|X}(T_i | \boldsymbol{X}_i)} \mu(T_i + \delta, \boldsymbol{X}_i)\right] \\
&\qquad + \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) \frac{f_{T|X}(T_i + \delta | \boldsymbol{X}_i)}{f_{T|X}(T_i | \boldsymbol{X}_i)} \mu(T_i + \delta, \boldsymbol{X}_i) \middle| T_i \right] - \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) \frac{f_{T|X}(T_i + \delta | \boldsymbol{X})}{f_{T|X}(T_i | \boldsymbol{X}_i)} \mu(T_i + \delta, \boldsymbol{X}_i)\right] \Bigg\}.
\end{aligned}
$$

Using changing of variables, the first term of (83) can be written as follows:

$$\begin{aligned}
&\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}\mu(t,\boldsymbol{x})\rho''\left(u_{K_1}^{\top}(t-\delta)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u_{K_1}^{\top}(t-\delta)\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})f_{X,T}(\boldsymbol{x},t)dxdt\\
=&\sqrt{N}\int_{\mathcal{T}}\int_{\mathcal{X}}\left\{\frac{f_{T|X}(t+\delta|\boldsymbol{x})}{f_{T|X}(t|\boldsymbol{x})}\right\}\mu(t+\delta,\boldsymbol{x})\rho''\left(u_{K_1}^{\top}(t)\tilde{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})\right)u_{K_1}^{\top}(t)\hat{A}_{K_1\times K_2}v_{K_2}(\boldsymbol{x})dF_{X,T}(\boldsymbol{x},t).
\end{aligned}$$

Using a similar argument of showing that (46)-(49) are all $o_p(1)$, we can show that (80)-(83) are all $o_p(1)$. By substituting $\phi(T,\boldsymbol{X},Y)=u_{K_1}(t_1)^{\top}\Phi_{K_1\times K_1}u_{K_1}(T)\pi_0(T,\boldsymbol{X})Y/\pi_0(T-\delta,\boldsymbol{X})^2$, we can obtain

$$\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}^{(1)}(t_1)=V_{t_1,t_0|t_0}^{-1/2}\cdot u_{K_1}^{\top}(t_1)\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}b_{1,K_1}(T_i,\boldsymbol{X}_i,Y_i)\right]+o_P(1), \tag{85}$$

where

$$\begin{aligned}
b_{1,K_1}(T_i,\boldsymbol{X}_i,Y_i)=&\frac{f_{T|X}(T_i+\delta|\boldsymbol{X})}{f_{T|X}(T_i|\boldsymbol{X}_i)}\frac{\pi_0(T_i,\boldsymbol{X}_i)^2}{\pi_0(T_i-\delta,\boldsymbol{X}_i)^2}\cdot\mathbb{E}[Y_i|T_i,\boldsymbol{X}_i]\cdot u_{K_1}(T_i)\\
&-\mathbb{E}\left[\frac{f_{T|X}(T_i+\delta|\boldsymbol{X})}{f_{T|X}(T_i|\boldsymbol{X}_i)}\frac{\pi_0(T_i,\boldsymbol{X}_i)^2}{\pi_0(T_i-\delta,\boldsymbol{X}_i)^2}\cdot Y_i\cdot u_{K_1}(T_i)\middle|\boldsymbol{X}_i\right]\\
&-\mathbb{E}\left[\frac{f_{T|X}(T_i+\delta|\boldsymbol{X})}{f_{T|X}(T_i|\boldsymbol{X}_i)}\frac{\pi_0(T_i,\boldsymbol{X}_i)^2}{\pi_0(T_i-\delta,\boldsymbol{X}_i)^2}\cdot Y_i\cdot u_{K_1}(T_i)\middle|T_i\right]\\
&+\mathbb{E}\left[\frac{f_{T|X}(T_i+\delta|\boldsymbol{X})}{f_{T|X}(T_i|\boldsymbol{X}_i)}\frac{\pi_0(T_i,\boldsymbol{X}_i)^2}{\pi_0(T_i-\delta,\boldsymbol{X}_i)^2}\cdot Y_i\cdot u_{K_1}(T_i)\right].
\end{aligned}$$

By combining (85), (76), and (77), we have

$$\begin{aligned}
&\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}(t_1)\\
=&\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}^{(1)}(t_1)+\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}^{(2)}(t_1)+\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{1N}^{(3)}(t_1)\\
=&V_{t_1,t_0|t_0}^{-1/2}\cdot u_{K_1}(t_1)^{\top}\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}b_{1,K_1}(T_i,\boldsymbol{X}_i,Y_i)\right]+o_P(1).
\end{aligned} \tag{86}$$

Similar to the proof of (40), we can show

$$\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\cdot b_{2N}(t_1)=u_{K_1}(t_1)^{\top}\Phi_{K_1\times K_1}^{-1}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}b_{2,K_1}(T_i,\boldsymbol{X}_i,Y_i)\right]+o_P(1), \tag{87}$$

where

$$b_{2,K_1}(T_i, \boldsymbol{X}_i, Y_i) = \frac{\pi_0(T_i, \boldsymbol{X}_i)}{\pi_0(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}(T_i) - \mathbb{E}\left[\frac{\pi_0(T_i, \boldsymbol{X}_i)}{\pi_0(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}(T_i)\middle| T_i, \boldsymbol{X}_i\right] + \mathbb{E}\left[\frac{\pi_0(T_i, \boldsymbol{X}_i)}{\pi_0(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}(T_i)\middle| \boldsymbol{X}_i\right] - \mathbb{E}\left[\frac{\pi_0(T_i, \boldsymbol{X}_i)}{\pi_0(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}(T_i)\right].$$

Therefore, by combining (75), (86) and (87), we can have that

$$\sqrt{N} V_{t_1,t_0|t_0}^{-1/2} \cdot \widehat{\theta}_{t_1|t_0} = V_{t_1,t_0|t_0}^{-1/2} \cdot u_{K_1}(t_1)^\top \Phi_{K_1 \times K_1}^{-1} \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \{b_{1,K_1}(T_i, \boldsymbol{X}_i, Y_i) + b_{2,K_1}(T_i, \boldsymbol{X}_i, Y_i)\} + o_P(1). \tag{88}$$

By combining (74) and (88), we can obtain

$$\begin{aligned} &\sqrt{N} V_{t_1,t_0|t_0}^{-1/2} \cdot \widehat{\theta}_{t_1,t_0|t_0} = \sqrt{N} V_{t_1,t_0|t_0}^{-1/2} \cdot \left\{\widehat{\theta}_{t_1|t_0} - \widehat{\theta}_{t_0|t_0}\right\} \\ =& V_{t_1,t_0|t_0}^{-1/2} \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big\{ u_{K_1}(t_1)^\top \Phi_{K_1 \times K_1}^{-1} \cdot \{b_{1,K_1}(T_i, \boldsymbol{X}_i, Y_i) + b_{2,K_1}(T_i, \boldsymbol{X}_i, Y_i)\} \\ &\qquad - u_{K_1}(t_0)^\top \Phi_{K_1 \times K_1}^{-1} \cdot b_{3,K_1}(T_i, Y_i) \Big\} + o_P(1), \end{aligned}$$

which implies $\sqrt{N} V_{t_1,t_0|t_0}^{-1/2} \cdot \widehat{\theta}_{t_1,t_0|t_0} \xrightarrow{d} N(0,1)$ by Liapounov CLT.

6 Variance Estimation in Monte Carlo Simulations

6.1 Proposed Variance Estimator

In Monte Carlo simulations, the estimated parameter is the average treatment effects, which corresponds to a differentiable loss function $L(v) = v^2$. The variance estimator can be simply defined as follows:

$$\begin{aligned} \hat{V}_{eff} = &\left[\frac{1}{N} \sum_{i=1}^{N} m(T_i; \hat{\boldsymbol{\beta}}) m(T_i; \hat{\boldsymbol{\beta}})^\top\right]^{-1} \\ &\times \left\{\frac{1}{N} \sum_{j=1}^{N} \left[\hat{\psi}(Y_j, T_j, \boldsymbol{X}_j; \hat{\boldsymbol{\beta}}) - \text{mean}(\hat{\psi})\right] \cdot \left[\hat{\psi}(Y_j, T_j, \boldsymbol{X}_j; \hat{\boldsymbol{\beta}}) - \text{mean}(\hat{\psi})\right]^\top\right\} \\ &\times \left[\frac{1}{N} \sum_{i=1}^{N} m(T_i; \hat{\boldsymbol{\beta}}) m(T_i; \hat{\boldsymbol{\beta}})^\top\right]^{-1}, \end{aligned} \tag{89}$$

where

$$\hat{\psi}(Y,T,\boldsymbol{X};\hat{\boldsymbol{\beta}}) = \hat{\pi}_{K'}(T,\boldsymbol{X})m(T;\hat{\boldsymbol{\beta}})\{Y-\hat{\mathbb{E}}[Y|T,\boldsymbol{X}]\}+\widehat{\mathbb{E}}\left[\{Y-g(T;\widehat{\boldsymbol{\beta}})\}\pi_0(T,\boldsymbol{X})m(T;\widehat{\boldsymbol{\beta}})|\boldsymbol{X}\right],$$

and

$$\widehat{\mathbb{E}}\left[Y|T,\boldsymbol{X}\right] = \left[\sum_{i=1}^{N} Y_i w_{K_0}(T_i,\boldsymbol{X}_i)^\top\right]\left[\sum_{i=1}^{N} w_{K_0}(T_i,\boldsymbol{X}_i)w_{K_0}(T_i,\boldsymbol{X}_i)^\top\right]^{-1} w_{K_0}(T,\boldsymbol{X})$$

and

$$\begin{aligned}\widehat{\mathbb{E}}\left[\{Y-g(T;\widehat{\boldsymbol{\beta}})\}\pi_0(T,\boldsymbol{X})m(T;\widehat{\boldsymbol{\beta}})|\boldsymbol{X}\right] = &\left[\sum_{i=1}^{N}\hat{\pi}_{K'}(T_i,\boldsymbol{X}_i)(Y_i-g(T_i;\hat{\boldsymbol{\beta}}))m(T_i;\hat{\boldsymbol{\beta}})v_{M_0}(\boldsymbol{X}_i)^\top\right]\\ &\times\left[\sum_{i=1}^{N} v_{M_0}(\boldsymbol{X}_i)v_{M_0}(\boldsymbol{X}_i)^\top\right]^{-1} v_{M_0}(\boldsymbol{X}),\end{aligned}$$

and

$$\text{mean}(\hat{\psi}) := \frac{1}{N}\sum_{j=1}^{N}\hat{\psi}(Y_j,T_j,\boldsymbol{X}_j;\hat{\boldsymbol{\beta}}).$$

6.2 True Values of V_{eff} in Monte Carlo Simulations

In Section 9 of the main paper, Monte Carlo simulations on variance estimation are performed. Computing the bias, standard deviation, and RMSE of the variance estimator $\hat{V}_{eff}$ requires to compute the true variance V_{eff}. We describe how to compute V_{eff} under DGP-L1 in Section 6.2.1 and DGP-NL1 in Section 6.2.2. DGP-L2 and DGP-NL2 are omitted since they can be handled in the same way as DGP-L1 and DGP-NL1.

To reduce notation, the single covariate X_1 is redefined as X. Note that the influence function is written as

$$\begin{aligned}\psi(Y,T,X;\boldsymbol{\beta}^*) =&\pi_0(T,X)m(T;\boldsymbol{\beta}_0)\{Y-g(T;\boldsymbol{\beta}^*)\}-\mathbb{E}\left[\{Y-g(T;\boldsymbol{\beta}^*)\}\pi_0(T,X)m(T;\boldsymbol{\beta}^*)|T,X\right]\\ &+\mathbb{E}\left[\{Y-g(T;\boldsymbol{\beta}^*)\}\pi_0(T,X)m(T;\boldsymbol{\beta}^*)|T\right]+\mathbb{E}\left[\{Y-g(T;\boldsymbol{\beta}^*)\}\pi_0(T,X)m(T;\boldsymbol{\beta}^*)|X\right]\\ &-\mathbb{E}\left[\{Y-g(T;\boldsymbol{\beta}^*)\}\pi_0(T,X)m(T;\boldsymbol{\beta}^*)\right]\\ =&\pi_0(T,X)m(T;\boldsymbol{\beta}^*)\{Y-g(T;\boldsymbol{\beta}^*)\}-\mathbb{E}\left[\{Y-g(T;\boldsymbol{\beta}^*)\}\pi_0(T,X)m(T;\boldsymbol{\beta}^*)|T,X\right]\\ &+\mathbb{E}\left[\{Y-g(T;\boldsymbol{\beta}^*)\}\pi_0(T,X)m(T;\boldsymbol{\beta}^*)|X\right]. \qquad (90)\end{aligned}$$

6.2.1 DGP-L1

Recall that DGP-L1 is

$$T = 1 + \rho_{T,X} \cdot X + \xi, \quad Y = 1 + X + T + \varepsilon. \quad (\rho_{T,X} = 0.2)$$

We have

$$\mathbb{E}[Y|T, X] = 1 + X + T, \quad \mathbb{E}[Y(t)] = g(t; \boldsymbol{\beta}^*) = 1 + t.$$

We directly compute

$$\begin{aligned}
\mathbb{E}\left[\{Y - g(T;\boldsymbol{\beta}^*)\}\pi_0(T,X)m(T;\boldsymbol{\beta}^*)|T=t, X=x\right] &= \int \{y - g(t;\boldsymbol{\beta}^*)\} \cdot \frac{f_T(t)}{f_{T|X}(t|x)} \cdot m(t) \cdot f_{Y|T,X}(y|t,x)dy \\
&= m(t) \cdot \pi_0(t,x) \cdot \{\mathbb{E}[Y|X=x, T=t] - g(t;\boldsymbol{\beta}^*)\} = m(t) \cdot \pi_0(t,x) \cdot \{1 + x + t - (1+t)\} \\
&= m(t) \cdot \pi_0(t,x) \cdot x \qquad (91)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\{Y - g(T;\boldsymbol{\beta}^*)\}\pi_0(T,X)m(T;\boldsymbol{\beta}^*)|X=x\right] &= \int m(t) \cdot \pi_0(t,x) \cdot x \cdot f_{T|X}(t|x)dt \\
&= x \cdot \int m(t)f(t)dt = x \cdot \mathbb{E}[m(T)]. \qquad (92)
\end{aligned}$$

Substitute (91) and (92) into (90) to get

$$\begin{aligned}
\psi(Y,T,X;\boldsymbol{\beta}^*) =& \pi_0(T,X)m(T)\{Y - 1 - T\} - m(T) \cdot \pi_0(T,X) \cdot X + X \cdot \mathbb{E}[m(T)] \\
=& \pi_0(T,X) \cdot m(T) \cdot \{Y - 1 - X - T\} + X \cdot \mathbb{E}[m(T)] \\
=& \pi_0(T,X) \cdot \{Y - 1 - X - T\} \cdot \begin{bmatrix} 1 \\ T \end{bmatrix} + \begin{bmatrix} X \\ X \end{bmatrix}. \qquad (93)
\end{aligned}$$

To compute $\pi_0(t,x)$, note that

$$\begin{aligned}
f_{T|X}(t|x) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t - 1 - \rho_{T,X} \cdot x)^2}{2}\right), \\
f_T(t) &= \frac{1}{\sqrt{2\pi \cdot (1 + \rho_{T,X}^2)}} \exp\left(-\frac{(t-1)^2}{2 \cdot (1 + \rho_{T,X}^2)}\right), \quad (T \sim N(1, 1 + \rho_{T,X}^2)).
\end{aligned}$$

Hence,

$$\pi_0(t,x) = \frac{f_T(t)}{f_{T|X}(t|x)} = \frac{1}{\sqrt{1+\rho_{T,X}^2}} \exp\left\{ \frac{\rho_{T,X}^2 \cdot (t-1)^2 + (1+\rho_{T,X}^2) \cdot \rho_{T,X}^2 \cdot x^2 - 2 \cdot (1+\rho_{T,X}^2) \cdot \rho_{T,X} \cdot x(t-1)}{2(1+\rho_{T,X}^2)} \right\}. \tag{94}$$

Using (93) and (94), the true variance can be computed as

$$V_{eff} = \left[\frac{1}{N} \sum_{i=1}^{N} m(T_i) m(T_i)^\top \right]^{-1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{\psi}(Y_i, T_i, X_i; \boldsymbol{\beta}^*) \psi(Y_i, T_i, X_i; \boldsymbol{\beta}^*)^\top \right\} \left[\frac{1}{N} \sum_{i=1}^{N} m(T_i) m(T_i)^\top \right]^{-1}. \tag{95}$$

Based on a simulated sample with large enough size $N = 10^8$, it follows that $V_{11} = 3.142$, $V_{12} = -1.097$, and $V_{22} = 1.097$.

6.2.2 DGP-NL1

Recall that DGP-NL1 is

$$T = \rho_{T,X} \cdot X^2 + \xi, \quad Y = X^2 + T + \epsilon. \quad (\rho_{T,X} = 0.1)$$

We have

$$\mathbb{E}[Y|T,X] = X^2 + T, \quad \mathbb{E}[Y(t)] = g(t; \boldsymbol{\beta}^*) = 1 + t.$$

Eq. (91) is now rewritten as

$$\mathbb{E}\left[\{Y - g(T; \boldsymbol{\beta}^*)\} \pi_0(T,X) m(T; \boldsymbol{\beta}^*) | T = t, X = x \right] = m(t) \cdot \pi_0(t,x) \cdot \{x^2 - 1\}.$$

Eq. (92) is now rewritten as

$$\mathbb{E}\left[\{Y - g(T; \boldsymbol{\beta}^*)\} \pi_0(T,X) m(T; \boldsymbol{\beta}^*) | X = x \right] = \{x^2 - 1\} \cdot \mathbb{E}[m(T)].$$

Substitute those equations into (90) to get

$$\psi(Y,T,X; \boldsymbol{\beta}^*) = \pi_0(T,X) \cdot \{Y - X^2 - T\} \cdot \begin{bmatrix} 1 \\ T \end{bmatrix} + \begin{bmatrix} X^2 - 1 \\ \{X^2 - 1\} \cdot \rho_{T,X} \end{bmatrix}. \tag{96}$$

To compute $\pi_0(t,x)$, note that

$$f_{T|X}(t|x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t-\rho_{T,X}\cdot x^2)^2}{2}\right),$$
$$\hat{f}_T(t) = \frac{1}{Nh}\sum_{i=1}^{N} k\left(\frac{T_i - t}{h}\right) = \frac{1}{Nh}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(T_i-t)^2}{2h^2}\right),$$

where $k(\cdot)$ is the kernel function which can be taken as $k(z) = (2\pi)^{-1/2}\exp(-z^2/2)$, and the bandwith can be taken as, say, $h = 0.1$. Then

$$\pi_0(t,x) = \frac{\hat{f}_T(t)}{f_{T|X}(t|x)} = \frac{1}{Nh}\sum_{i=1}^{N} \exp\left(-\frac{(T_i-t)^2}{2h^2} + \frac{(t-\rho_{T,X}\cdot x^2)^2}{2}\right). \tag{97}$$

Using (96) and (97), the true variance can be computed from (95). Based on a simulated sample with large enough size $N = 50000$, it follows that $V_{11} = 3.043$, $V_{12} = -0.118$, and $V_{22} = 1.074$.

References

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press.

Cattaneo, M. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138–154.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models When the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259–276.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79(1), 147–168.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4, chap. 36, pp. 2111–2245. Citeseer.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89(427), 846–866.

Tchetgen Tchetgen, E. J., and I. Shpitser (2012): "Semiparametric theory for causal me-

diation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis," *The Annals of Statistics*, 40(3), 1816–1845.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press.