ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

PARTIAL MEAN PROCESSES WITH GENERATED REGRESSORS: CONTINUOUS TREATMENT EFFECTS AND NONSEPARABLE MODELS

Ying-Ying Lee

Number 706
May 2014

Manor Road Building, Manor Road, Oxford OX1 3UQ

Partial Mean Processes with Generated Regressors: Continuous Treatment Effects and Nonseparable Models

Ying-Ying Lee *
University of Oxford †

May 2014

Abstract

Partial mean processes with generated regressors arise in several important econometric problems, such as the distribution of potential outcomes with continuous treatments and the quantile structural function in a nonseparable triangular model. This paper proposes a fully nonparametric estimator for the partial mean process, where the second step consists of a kernel regression on regressors that are estimated in the first step. The main contribution is a uniform expansion that characterizes in detail how the estimation error associated with the generated regressor affects the limiting distribution of the marginal integration estimator. The general results are illustrated with three examples: control variables in triangular models (Newey, Powell, and Vella, 1999; Imbens and Newey, 2009), the generalized propensity score for a continuous treatment (Hirano and Imbens, 2004), and the propensity score for sample selection (Das, Newey, and Vella, 2003).

Keywords: Continuous treatment, partial means, nonseparable models, generated regressors, control function
JEL Classification: C13, C14, C31

1 Introduction

This paper studies objects of interest that are identified through conditional expectations with some of the conditioning variables estimated at a preliminary step, known as *generated regressors*. A *generated regressor* can be viewed as an infinite-dimensional nuisance parameter in many economics examples: the control variables in triangular models (Newey, Powell, and Vella, 1999; Imbens and Newey, 2009), the generalized propensity score for continuous treatment (Hirano and Imbens, 2004), or the propensity score for sample selection (Das, Newey, and Vella, 2003). The *partial mean* is a marginal integration of such conditional expectations where the generated regressor is averaged out and the value of certain other explanatory variables of interest is fixed. We propose a nonparametric multi-step estimator for

*The author is deeply grateful to Jack Porter and Bruce Hansen for invaluable guidance and encouragement. The author thanks Xiaoxia Shi, Chris Taber, and the seminar participants in UW-Madison for helpful comments and discussion. I also thank Debopam Bhattacharya, Juan Carlo Escanciano, Yu-Chin Hsu, David Jacho-Chávez, Arthur Lebwel, Enno Mammen, Christoph Rothe, Anne Vanhems, Ingrid van Keilegom, Melanie Schienle, Kyungchul Song, as well as the seminar participants in Academia Sinica, UCL, University of Bristol, University of Mannheim, University of Oxford, TSE, 2012/2013 Midwest Econometrics Group Meeting, 2012 Info-Metrics Nonparametric Conference, 2013 SETA, and 2014 Cambridge Nonparametric and Semiparametric Methods Conference. All remaining errors are mine.

†Department of Economics, University of Oxford. Manor Road Building, Manor Road, OX1 3UQ, United Kingdom. E-mail: ying-ying.lee@economics.ox.ac.uk Website: https://sites.google.com/site/yyleelilian/

this partial mean and elaborate how the estimation error associated with the generated regressor affects the limiting properties of the estimator.

Our results provide new insights on nonparametric estimation of continuous treatment effect models. Two key features of these models are non-separability in the unobservables and heterogeneity in treatment intensity effects. The proposed methods capture heterogeneous treatment intensity effects by estimating an array of distributional structural features that can be applied to a variety of economic questions. For example, when evaluating a social program, researchers might be interested in how the length of exposure to the program affects the wage distribution. The proposed method includes inference on smooth functionals of the outcome distribution process. In this example, a researcher could consider how inequality responds to the length of exposure to an anti-poverty program by tracing out the Gini coefficient of the wage distribution by time in the program. In demand analysis, we can estimate the Engel curve that describes how the distribution of household food consumption responds to an exogenous change in total expenditure. In particular, we provide a uniform inference method for the quantile structural function in nonseparable triangular simultaneous equations models (Imbens and Newey, 2009). Analyzing these questions often involves generated regressors to account for the endogeneity of the continuous explanatory variable. Understanding the estimation error of the generated regressors is fundamental to perform correct inference.

We define the weighted partial mean process by

$$\mathbb{E}\big[F_{Y|T\Lambda}(y|t,\Lambda)\cdot W\big] \equiv \int \mathbb{E}\big[\mathbf{1}_{\{Y\leq y\}}\big|T=t,\Lambda=\lambda\big]\cdot w \;\; dF_{\Lambda W}(\lambda,w). \tag{1}$$

The conditional expectation $F_{Y|T\Lambda}(y|t,\Lambda)$ is the conditional cumulative distribution function (cdf) of the outcome Y given the continuous explanatory variables T and nuisance variables Λ. The regressor $\Lambda = \Lambda(\cdot)$ is a function of observables and can be estimated in the first step. The conditional cdf $F_{Y|T\Lambda}(y|t,\Lambda)$ is estimated nonparametrically by a kernel regression in a second step. The third step is sample analogue of a marginal integration that averages out the generated regressor Λ and the weight W, but fixes the value of the continuous treatment T at t. Because the regression function $F_{Y|T\Lambda}(y|t,\Lambda)$ contains more arguments than being averaged over in the third step, this is known as a *partial mean* in the terminology of Newey (1994b). The partial mean is a nonparametric function of the continuous variables t and hence is estimated at a convergence rate slower than regular root-n.

This paper builds on and extends the partial mean literature in three ways: First, the dependent variable is $\{\mathbf{1}_{\{Y\leq y\}} : y \in \mathcal{Y}\}$ a process indexed by the threshold y, making it possible to estimate the whole distribution and also various distributional features, such as the mean, quantiles, and inequality measures. This setup intrinsically allows the dependent variable to be Y or discrete. Second, the regressors Λ can be unobserved and estimated parametrically or nonparametrically in the first step. Third, our stochastic expansion of the estimator is uniform over the treatment level t and the threshold value y, which could be used as an intermediate step in a more complicated estimation procedure.

There are two infinite-dimensional parameters in the partial mean: the generated regressor $\Lambda(\cdot)$ and the regression function $F_{Y|T\Lambda}(y|t,\cdot)$. The generated regressor enters the partial mean through two channels: first, it is an *argument* of the regression function that directly enters the partial mean; second, it is a *regressor* that determines the functional form of the regression $F_{Y|T\Lambda}(y|t,\cdot)$. The *argument* role of the generated regressor and the regression function can be analyzed as unknown functions that have been studied in the extensive literature; for example, Pakes and Pollard (1989),

Andrews (1994), Sherman (1994), Chen, Linton, and Van Keilegom (2003), Ichimura and Lee (2010), among many others. The challenge lies in the *regressor* role of the generated regressor. There is a growing literature on nonparametric regression with generated regressors, including Sperlich (2009), Mammen, Rothe, and Schienle (2012), Mammen, Rothe, and Schienle (2013), Hahn and Ridder (2013), Escanciano, Jacho-Chávez, and Lewbel (2014), Song (2008), Song (2014), among others. Most of the previous work focuses on the impact of generated regressors on the nonparametric regression and its application in semiparametric models. Our main contribution is a concrete representation of the impact of generated regressors on the nonparametric partial mean. The new uniform expansion distinguishes the *argument* and *regressor* roles of the generated regressor and contains three important *elements* — (i) partial mean/full mean structure, [1] (ii) generated regressor as an index function of observables, and (iii) projection error of the weight $W - \mathbb{E}[W|\Lambda]$. This finding appears to be of significant practical importance by providing conditions under which the estimation error is negligible. The new message is that these elements are generic for marginal integration estimators that involve a kernel regression with generated regressors, in both nonparametric or semiparametric models.

Another set of our results is weak convergence of the partial mean of the weighted conditional cdf in Eq. (1) allowing for generated regressors. A multiplier bootstrap method is shown to be valid for uniform inference, which enables functional hypotheses tests for the whole distribution, such as tests for no effect or stochastic dominance. By extending the results to the Hadamard-differentiable functionals of the partial mean process, we are able to provide the limiting distribution and uniform inference method for estimating common inequality measures and various distributional structural features; for example, the Lorenz curves and the Gini coefficients (Bhattacharya, 2007; Rothe, 2010; Firpo and Pinto, 2011; Donald, Hsu, and Barrett, 2012; Chernozhukov, Fernández-Val, and Melly, 2013). Because our partial mean estimator converges at a nonparametric rate, the standard Donsker property in empirical process theory is not directly applied. Deriving the weak convergence and multiplier method demands more involved technical arguments.

The partial mean is a statistical object with broad applications, such as additive nonparametric models and differential equation solution introduced in Newey (1994b). Nevertheless, we focus on the potential outcome framework that is convenient for interpretation and relatively simplifies notation. The potential outcome corresponding to the treatment level t, denoted by $Y(t)$, is the counterfactual outcome that would have occurred given t. For each observation, we only observe the outcome from the chosen or received treatment level; the potential outcomes given other treatment levels are latent. The key causal object of interest is the unconditional distribution of $Y(t)$ by averaging out other covariates and unobservable heterogeneity. We consider identification of the distributional causal effects via functionals of the weighted partial means in Eq. (1). One motivating example is the control function approach that assumes the continuous treatment T to be exogenous conditional on generated regressors Λ. [2]

The Hadamard-differentiable functionals cover a wide class of regression functions with generated regressors that have been of interest in the econometrics and statistics literature. Some notable

[1] The nonparametric convergence rate of a partial mean depends on the number of regressors that are not averaged over. On the other hand, a *full mean* averages over all the arguments and its nonparametric estimation can be root-n consistent. This insight on the convergence rate is a well-known feature of partial means and full means.

[2] According to Matzkin (2007), "A *control function* is a function of observable variables such that conditioning on its value purges any statistical dependence that may exist between the observable and unobservable explanatory variables in an original model."

examples are as follows. The unconditional mean $\mathbb{E}[Y(t)]$ is the *average structural function* or the *dose-response function* (Blundell and Powell, 2003; Flores, 2007). The equivalent variation in welfare analysis studied by Bhattacharya (2013) can also be expressed as the average structural function. The unconditional quantile function $F^{-1}_{Y(t)}(\tau)$ defines the *quantile structural function* (Imbens and Newey, 2009). The difference between two treatment levels is the unconditional quantile treatment effect. The local average response or the marginal mean treatment effect on the treated is based on $\mathbb{E}\big[Y(t)\big|T=\bar{t}\big]$, the expected outcome given a hypothetical value t for those currently choosing treatment level $\bar{t}$, holding their other observables and unobservables fixed at baseline values $\bar{t}$ (Altonji and Matzkin, 2005; Florens, Heckman, Meghir, and Vytlacil, 2008).

We derive a uniform expansion of the three-step estimator characterizing the impact of generated regressors on the final estimator. The generated regressors can be general functions, such as regression, density, or quantile regression. We can estimate the generated regressors by a (semi)parametric or nonparametric method that admits a uniform asymptotic linear representation. [3] We illustrate the usefulness of our general results by two examples. The first is the control variable as in the triangular simultaneous equations models in Newey, Powell, and Vella (1999) and Imbens and Newey (2009). A novel finding is that the estimation error of the control variable diminishes at root-n rate. So the average and quantile structural functions can be estimated as if the true control variable was observed. The key insight is that the marginal integration estimator averages out the generated regressor, so its estimation error converges at a faster rate. Similarly for the nonparametric sample selection model with endogeneity in Das, Newey, and Vella (2003), the estimation errors from the propensity score and control function are ignorable to our partial mean estimator. So we provide a more precise alternative to their two-step series estimator. Moreover, our stochastic expansion, which is uniform over t and y, is useful when a partial mean is an intermediate step in a semiparametric setting or multi-step estimation. In such cases, the estimation error from the control variables might not be ignorable and is characterized by our expansion.

The second example is the generalized propensity score (GPS), defined as the conditional density function of treatment given observable characteristics. Under the unconfoundedness assumption, the GPS is known to reduce dimensionality in the second-step regression (Hirano and Imbens, 2004) and allow a weaker smoothness condition. Regressing on the GPS is common practice in program evaluation where the length of participation is often taken as the continuous treatment (Agüero, Carter, and Woolard, 2010; Flores, Flores-Lagunes, Gonzalez, and Neumann, 2012; Kluve, Schneider, Uhlendorff, and Zhao, 2012). Nonetheless, to the best of our knowledge, this paper is the first presentation of a complete limit theory of nonparametric regression on the estimated GPS. We show that for estimating the overall distribution $F_{Y(t)}(y)$, the GPS does not result in an efficiency gain, over controlling directly for the whole set of the observables. This is mainly because the sampling variation from estimating the GPS $f_{T|X}(t|X)$ is a partial mean and hence converges at the same rate as the final estimator. This finding parallels the binary treatment case by Hahn and Ridder (2013) and Mammen, Rothe, and Schienle (2013) for nonparametric regression on the propensity score in Heckman, Ichimura, and Todd (1998).

To extract the influence from the *regressor* role of the generated regressor, we implement stochas-

[3] Song (2014) considers the generated regressor to be single-index that might not have a linear representation and provides conditions under which the estimation error is ignorable.

tic equicontinuity arguments in empirical process theory on the second-step kernel regression and the third-step partial mean. An intuition for our analysis is that the partial mean averages over the arguments of the regression function and the kernel regression averages over the generated regressors. By the symmetric structure of the kernel function, we deal with the *regressor* role of the generated regressor by viewing it as an "argument" of the partial mean. This is the key to generally analyze our marginal integration estimator for nonparametric partial mean and semiparametric full mean models. To implement this approach, we modify the stochastic equicontinuity argument for the kernel estimator with respect to the regressors developed in Mammen, Rothe, and Schienle (2012) (MRS12, hereafter). While MRS12 contribute a detailed characterization of how the estimation error of the generated regressor affects the conditional mean regression estimator, we focus on the partial mean of such regression and do not control the approximation error from linearizing the second-step regression estimator. Our third step is in the spirit of Escanciano, Jacho-Chávez, and Lewbel (2014)(EJL14, hereafter) who use a stochastic equicontinuity argument on a full mean based on a general class of regression functions; in contrast, we make use of the kernel regression estimator in the nonparametric partial mean. Hahn and Ridder (2013) are among the first to characterize the asymptotic variance of a semiparametric marginal integration estimator contributed by generated regressors by using Newey (1994a) path-derivative method. It is nontrivial to extend their theoretical approach to the nonparametric context of partial mean for fixing continuous regressors, although the same intuition carries over. To our best knowledge, this is the first paper that analyzes a marginal integration estimator for general partial means and full means, accounting for generated regressors. Our stochastic expansion recovers the theoretical results of Hahn and Ridder (2013) for the semiparametric case.

The rest of the paper is organized as follows: Section 2 introduces the setup and causal parameters of interest. Readers who are only interested in the estimation and statistical properties might skip this section. Section 3 outlines three-step nonparametric kernel estimation for the weighted partial mean process with generated regressors in Eq. (1). Section 4 presents the main asymptotic theorems. Section 5 illustrates the usefulness of our results by economic examples. Our stochastic equicontinuity arguments for the partial mean apply to semiparametric models with generated regressors, which complements the important findings in Hahn and Ridder (2013), Mammen, Rothe, and Schienle (2013), and EJL14. Although semiparametric models are not the focus of this paper, the insights of *the elements* — index and projection of the weight — in our stochastic expansion carry over and derives new results in two examples: (i) the policy effects in Rothe (2010) and Imbens and Newey (2009) with control variables; (ii) the average treatment effect on the treated for a multi-valued discrete treatment by regressing on the propensity score. Section 6 presents the results for semiparametric models. Section 7 presents the limit theories for treatment/policy effects using the functional delta method for the Hadamard-differentiable policy functionals. We explicitly carry out the limit theory for estimating the quantile processes. A multiplier bootstrap method enables a uniform inference. Section 8 concludes this paper. The proofs are in the Appendix where we start with a heuristic summary.

2 Potential Outcome and Nonseparable Models

This section introduces the potential outcome framework and the causal objects of interest, which are identified by functionals of the partial means in Eq. (1). The treatment effect model is known to be

equivalent to a nonseparable outcome with a general disturbance, where the outcome equation is $Y = \phi(T, \epsilon)$, e.g., Imbens and Newey (2009), White and Chalak (2013). The error ϵ represents unobservable individual heterogeneity. No functional form assumption is imposed on the general disturbances ϵ, like monotonicity, dimensionality, or separability. Let $Y(t) \equiv \phi(t, \epsilon)$ denote the potential outcome corresponding to the treatment level t, where the randomness comes from the unobserved disturbances ϵ. If the treatment value $\bar{t}$ is chosen by an economic agent i, then $T_i = \bar{t}$ and the observed outcome $Y_i = Y_i(\bar{t})$ is one of the potential outcomes $\{Y_i(t) = \phi(t, \epsilon_i)\}_{t \in \mathcal{T}}$. For example, we observe the quantity demanded at the observed price, but cannot observe what the demand would have been given other prices. With regard to program evaluation, we observe the wage of a participant after one year in a job training program, but we wish to learn what the wage would have been if the participant had stayed in the program for two years.

2.1 Treatment effects and structural functions

The cumulative distribution function (cdf) of the potential outcome $F_{Y(t)}(y) = \mathbb{E}[\mathbf{1}_{\{\phi(t,\epsilon) \leq y\}}]$ is the outcome distribution when the value of the treatment T is fixed at t and the expectation is taken with respect to the marginal density of ϵ, $f_\epsilon(\epsilon)$. In other words, it is the unconditional outcome distribution if, hypothetically, the economic agent had been assigned to the treatment level t. The cdf of $Y(t)$ for those who have chosen their treatment level $\bar{t}$ is defined by $F_{Y(t)|T}(y|\bar{t}) = \mathbb{E}\big[\mathbf{1}_{\{\phi(t,\epsilon) \leq y\}} \big| T = \bar{t}\big]$, where the expectation is taken with respect to the conditional density $f_{\epsilon|T}(\epsilon|\bar{t})$.

An array of estimands are often of interest based on these causal outcome distributions. We consider a general class of functionals Γ on $F_{Y(t)}(\cdot)$ and $F_{Y(t)|T}(\cdot|\bar{t})$. For example, if interest centers on the quantile treatment effect (QTE), we let Γ be the quantile operator $\Gamma(F_{Y(t)}) = Q_\tau(Y(t)) \equiv \inf\{y : F_{Y(t)}(y) \geq \tau\}$. The QTE corresponding to a change from t to $\bar{t}$ is $\Gamma(F_{Y(\bar{t})}) - \Gamma(F_{Y(t)})$. Similarly, the QTE *on the treated* $\bar{t}$ is $\Gamma(F_{Y(\bar{t})|T}(y|t)) - \Gamma(F_{Y(t)|T}(y|t))$. When the outcome structural function ϕ is monotone in the error ϵ, the *quantile structural function* $Q_\tau(Y(t))$ equals the structural function evaluated at (t, τ), $\phi(t, \tau)$, by a normalization. If interest is on the mean treatment effect, then let Γ be the mean operator. When the outcome is separable $Y = \phi(T) + \epsilon$, the *average structural function* $\mathbb{E}[Y(t)]$ equals $\phi(t)$ up to an additive constant. Other inequality measures are also applicable, such as the coefficient of variation, the interquantile range, the Theil index, the Gini coefficient, the Lorenz curve.

2.2 Identification

We use conditional independence and common support assumptions to show that the partial mean process in Eq. (1) identifies the causal outcome distributions $F_{Y(t)}(\cdot)$ and $F_{Y(t)|T}(\cdot|\bar{t})$. The results here could be straightforwardly generalized to consider the conditional potential outcome distribution where the conditioning set consists of exogenous observables or discrete covariates.

Assumption 1 (Conditional Independence Assumption)
T and ϵ are independent conditional on Λ. Or for any $t \in \mathcal{T}$, the potential outcome $Y(t)$ is independent of the treatment T, given Λ.

In words, given the value of Λ, the treatment T is independent of the unobservable heterogeneity ϵ. The leading examples are the following types of conditioning variables Λ. Under unconfoundedness

or selection on observables X, Assumption 1 is satisfied by $\Lambda = X$ or $\Lambda = f_{T|X}(t|X)$ the generalized propensity score in Hirano and Imbens (2004). When unconfoundedness is violated, one approach to satisfying Assumption 1 is through control variables as in the triangular simultaneous equations model. For example, Imbens and Newey (2009) show the conditional distribution function of the endogenous variable given the instrumental variables is a control variable $\Lambda = F_{T|Z}(T|Z)$, where Z is an exogenous instrumental variable. The conditional distribution of the potential outcome $Y(t)$ given Λ is identified by Assumption 1,

$$F_{Y(t)|\Lambda}(y|\Lambda) \equiv \mathbb{E}[\mathbf{1}_{\{Y(t)\leq y\}}|\Lambda] = \mathbb{E}[\mathbf{1}_{\{Y(t)\leq y\}}|T=\bar{t},\Lambda] = \mathbb{E}[\mathbf{1}_{\{Y\leq y\}}|T=t,\Lambda] \equiv F_{Y|T\Lambda}(y|t,\Lambda) \quad (2)$$

$\forall \bar{t} \in \mathcal{T}$. That is, conditional on the control function Λ, the distribution of the potential outcome for choosing treatment intensity t is invariant to the current treatment intensity $\bar{t}$.

The following common support Assumption 2 is also known as the overlapping assumption for discrete treatments, i.e., the propensity score $Pr(T = t|\Lambda)$ cannot be exactly zero or one.

Assumption 2 (Common Support)
For $t \in \mathcal{T}$, the support of Λ conditional on $T = t$ equals the support of Λ.

By Assumptions 1 and 2, and following Eq. (2), the partial mean process with generated regressor in Eq. (1) identifies $F_{Y(t)}(y)$ with $W = 1$ and identifies $F_{Y(t)|T}(y|\bar{t})$ with $W(\lambda) = f_{T|\Lambda}(\bar{t}|\lambda)/f_T(\bar{t})$. The identification for the overall cdf $F_{Y(t)}$ has been shown in Theorem 3 in Imbens and Newey (2009).

In addition to the counterfactual of changing the treatment intensity t through the potential outcome $Y(t)$, a general usage of this weight function W would allow us to consider a wide variety of counterfactual objects. Following the idea of Oaxaca (1973), Blinder (1973), DiNardo, Fortin, and Lemieux (1996), Rothe (2010), and Chernozhukov, Fernández-Val, and Melly (2013), we could define the weight to be a ratio of a *counterfactual density* and the status-quo density of the observable characteristics, $W(X) = f_{X^*}(X)/f_X(X)$ under the unconfoundedness assumption $\Lambda = X$. An important choice of the counterfactual density $f_{X^*}(X)$ is $f_{X|T}(X|\bar{t})$ for those currently being treated or choosing $\bar{t}$. [4] [5]

Bseides the above regression-type identification, we present results for propensity-score weighting identification for completeness.

Remark (Propensity-score weighting)
Propensity-score weighting identification for continuous treatment variables relies on the introduction of a kernel function $K_h(T-t) \equiv \frac{1}{h^{d_t}}\Pi_{l=1}^{d_t} k(\frac{T_l - t_l}{h})$, where k is any conventional kernel and d_t is the dimension of a vector t. By a calculation involving a Taylor expansion of the kernel,[6] $F_{Y|T\Lambda}(y|t,\Lambda) = \lim_{h\to 0} \mathbb{E}[\mathbf{1}_{\{Y\leq y\}}K_h(T-t)|\Lambda]/f_{T|\Lambda}(t|\Lambda)$. Together with Eq. (2) and the law of iterated expectations,

[4] The sampling variation resulting from estimating this weight $W = f_{T|X}(\bar{t}|X)/f_T(\bar{t})$ on the inference of the weighted partial mean is in a previous version of this paper. We separate the results to another paper for brevity and focus on the issue of generated regressors.

[5] The weight will include a fixed trimming function, where the density of the conditioning variables are bounded away from zero, as in Newey (1994b). In fairness, the choice of fixed trimming function can affect the interpretation of the estimands considered. The subpopulation is selected such that the observables do not take extreme values.

[6] The identification argument will depend on the smoothness in $F_{Y|T\Lambda}(y|t,\Lambda)$ and $f_{T|\Lambda}(t|\Lambda)$ matching the order of the kernel k. Let r denote the order of the kernel k, and assume $F_{Y|T\Lambda}(y|t,\Lambda)$ and $f_{T|\Lambda}(t|\Lambda)$ are r-order continuously differentiable in t with uniformly bounded derivatives.

the propensity-score weighting identification is

$$\mathbb{E}[F_{Y(t)|\Lambda}(y|\Lambda)W] = \lim_{h\to 0}\mathbb{E}\Big[\frac{\mathbf{1}_{\{Y\le y\}}K_h(T-t)}{\mathbb{E}[K_h(T-t)|\Lambda]}\mathbb{E}[W|\Lambda]\Big] = \lim_{h\to 0}\mathbb{E}\Big[\frac{\mathbf{1}_{\{Y\le y\}}K_h(T-t)}{f_{T|\Lambda}(t|\Lambda)}\mathbb{E}[W|\Lambda]\Big]. \quad (3)$$

Flores, Flores-Lagunes, Gonzalez, and Neumann (2012) estimate the continuous treatment effect based on equations (1) and (3) nonparametrically with a parametric generalized propensity score without providing a limit theory. We do not exploit estimation based on this propensity-score weighting identification. [7]

3 Estimation

This section introduces a general procedure to estimate the process in Eq. (1)

$$\Big\{\ t \to F_{Y(t)}\big(y;\Lambda,W\big) \equiv \int \mathbb{E}\big[\mathbf{1}_{\{Y\le y\}}\big|T=t,\Lambda=\lambda\big]\cdot w\ dF_{\Lambda W}(\lambda,w) : y\in\mathcal{Y}\ \Big\}.$$

We denote the partial mean process by $F_{Y(t)}\big(y;\Lambda,W\big)$, where Λ collects the nuisance conditioning variables being averaged out in the partial mean and might contain the generated regressors. The notation of the potential outcome $Y(t)$ is convenient to represent the value of the outcome at which the continuous treatment variable T is fixed. The weight W is the additional variable that is not involved in the regression.

We will discuss specific estimators for each economic examples in the later sections. We find that the estimation approach outlined below has different properties depending on the details of implementation corresponding to each estimand. As a result, different asymptotic distribution results are proceeded for the different versions of this general estimation approach described below. The estimation procedure involves three steps:

1. (Generated Regressors) The generated regressor Λ can be estimated parametrically or nonparametrically, as long as it admits a uniform asymptotic linear representation and its uniform convergence rate satisfies certain conditions, specified in the next section. If the estimator $\hat{\Lambda}$ is a nonparametric regression by a kernel method, let the bandwidth be h_1, the order of the kernel be r_1, and the dimension of the regressors be d_1.
2. (Regression) The second step is the nonparametric regression of the indicator function $\mathbf{1}_{\{Y\le y\}}$ on $(T,\hat{\Lambda})$ and evaluated at (t,λ), i.e.,

$$\hat{F}_{Y|T\hat{\Lambda}}(y|t,\lambda) \equiv \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{\{Y_j\le y\}}K_h(T_j-t)K_h(\hat{\Lambda}_j-\lambda)\Big/\hat{f}_{T\hat{\Lambda}}(t,\lambda)$$
$$\hat{f}_{T\hat{\Lambda}}(t,\lambda) \equiv \frac{1}{n}\sum_{j=1}^{n}K_h(T_j-t)K_h(\hat{\Lambda}_j-\lambda)$$

The product kernel is $K_h(u) \equiv h^{-d_u}\Pi_{l=1}^{d_u}k\big(\frac{u_l}{h}\big)$, where $h=h_2$ is the bandwidth assumed the

[7] We generalize identification by propensity-score weighting from the binary treatment effect literature to the case of continuous treatments. For a discrete treatment, the kernel function degenerates to an indicator function, $\mathbb{E}\Big[\frac{\mathbf{1}_{\{Y\le y\}}\mathbf{1}_{\{T=t\}}}{P(T=t|\Lambda)}W(\Lambda)\Big]$. Propensity-score weighting estimation for the discrete treatment is well-studied in the treatment effect literature, e.g., Cattaneo (2010), Hirano, Imbens, and Ridder (2003), Firpo and Pinto (2011).

same for all the elements of the vector u for simplicity, and k is the r_2-order kernel function satisfying the following Assumption 4. Let the dimension of the regressors at this step be $d_2 = d_t + d_\lambda$.

3. (Partial Mean) The third step is the partial mean, fixing the treatment variable T at level t, $\hat{F}_{Y(t)}(y; \hat{\Lambda}, W) = n^{-1} \sum_{i=1}^{n} \hat{F}_{Y|T\hat{\Lambda}}(y|t, \hat{\Lambda}_i) \cdot W_i$.

The following assumptions will be maintained on the data generating process and the kernel used in the nonparametric regression step described above.

Assumption 3 (Smoothness)

(i) *The data* $\{Y_i, T_i, X_i, Z_i\}$, $i = 1, ..., n$, *is independent and identically distributed (i.i.d.). The random vector* $\Lambda = \Lambda(S_\lambda)$ *is a vector of measurable functions of* S_λ, *a subvector of* $\{T, X, Z\}$.

(ii) *The support of* Λ, $\mathbf{\Lambda}$, *is a compact and convex subset of* $\mathbb{R}^{d_\lambda}$. *The support of* T, $\mathcal{T}$, *is a compact and convex subset of* $\mathbb{R}^{d_t}$. (T, Λ) *has a probability density function* $f_{T\Lambda}(t, \lambda)$, *which is bounded away from zero and is* Δ*-order continuously differentiable with respect to both* t *and* λ, *with uniformly bounded derivatives.*

(iii) *Suppose the unconditional distribution* $F_Y(y)$ *is continuous on a compact support* $\mathcal{Y} \equiv [y_l, y_u] \subset \mathbb{R}$. *The conditional distribution* $F_{Y|T\Lambda}(y|t, \lambda)$ *is* Δ*-order continuously differentiable with respect to both* t *and* λ, *with uniformly bounded derivatives.*

The treatments T or covariates Λ could contain discrete variables and the kernel is replaced by an indicator function, known as the frequency method. For notational convenience, discrete covariates are not allowed for. The smoothness Assumption 3 (ii) requires that the treatment variables cannot have point masses, i.e., $Pr(T = t) = 0$ for $t \in \mathcal{T}$.

Assumption 4 (Kernel)

The kernel function $k(u) : \mathbb{R} \to \mathbb{R}$ *satisfies the following conditions: (i) (r-order)* $\int k(u)du = 1$, $\int u^l k(u)du = 0$ *for* $0 < l < r$, *and* $\int |u^r k(u)|du < \infty$ *for some* $r \geq 2$. *(ii) (bounded support) for some* $L < \infty$, $k(u) = 0$ *for* $|u| > L$. *(iii)* $k(u)$ *is r-times continuously differentiable and the derivatives are uniformly continuous and bounded. (iv) For an integer* Δ_k, *the derivatives of the kernel up to order* Δ_k *exist and are Lipschitz.*[8]

Uniform convergence of the first- and second-step estimators over the range of integration suffices for deriving the properties of the third-step estimator. However, kernel estimation is known to be biased at the boundary of the support. It commonly includes a fixed trimming function that chooses a compact, interior subsupport of (T, Λ) such that the estimator $\hat{F}_{Y|T\Lambda}(y|t, \lambda)$ satisfies the uniform convergence rate in Result A.1 in the Appendix. [9] In this case, the supports $\mathcal{T}$ and $\mathbf{\Lambda}$ are not restricted

[8] (iv) ensures that the estimator takes values in a function space not too complex for the stochastic equicontinuity argument.

[9] In principle, an estimated trimming function could be incorporated into our framework and considered in the asymptotic results. We do not pursue this extension in this paper. The fixed trimming choice allows us to focus on the technical issues associated with estimating the generated regressor and the whole distribution processes. There are two alternative approaches to estimate for the whole support. The first approach assumes a compact support. A generalized kernel or boundary kernel might be used to attain the uniform convergence over the whole compact support, as in Rothe (2010), Darolles, Florens, and Renault (2011), Graham, Imbens, and Ridder (2014). The asymptotic theories and proofs derived in this paper are yet shown to be unchanged. The second approach supposes the support to be unbounded or the density to be zero at the boundary of the support. A random or data-driven trimming is needed for the denominator problem and uniform consistency over the whole support. This approach will complicate the proofs and is beyond the scope of this paper.

to be bounded. The edge problem in Imbens and Newey (2009) is avoided by a fixed trimming. When the first step uses a nonparametric kernel estimation, we adopt another trimming function to trim the boundary of $\mathcal{S}$ such that the estimator for the generated regressor $\hat{\Lambda}(S_\lambda)$ satisfies the convergence rate uniformly over the interior of the support of S_λ. Therefore, the trimmed estimator consistently estimates $F_{Y|T\Lambda}(y|t,\lambda)$ for the subpopulation whose observables S_λ do not take extreme values. Then the third step uses this subsample with the second trimming function. In the following, we suppress the two fixed trimming functions for notational ease. That is, we work on a compact subsupport where the density functions are bounded away from zero, as in Assumption 3 (ii). And the uniform convergence results in Result A.1 hold over these compact integration ranges.

4 Asymptotic Results

We first present the limit theory for estimating the partial mean process when all the regressors are observed. This is for the case to identify the causal distribution of the potential outcome under unconfoundedness, where the Conditional Independence Assumption 1 is satisfied by $\Lambda = X$ observable characteristics. Our weak convergence of the entire distribution process is a nontrivial extension of the partial mean in Newey (1994b). The second subsection considers estimation of the generated regressors $\Lambda(S_\lambda)$. Our main result is a uniform stochastic or Bahadur expansion of the three-step estimator, revealing the influence of estimating the generated regressors of general function form on the final estimator.

To employ empirical process theory as part of the estimator behavior argument, we need to restrict the smoothness and complexity of the conditional cdf of outcomes and the generated regressor. The smoothness class that we will use is defined next. In words, the partial derivatives of these functions are uniformly bounded up to some specified orders. [10]

Definition ($\mathcal{C}_M^\alpha(\mathcal{S})$, van der Vaart and Wellner (1996) (P. 154))
$\mathcal{C}_M^\alpha(\mathcal{S})$ is defined on a bounded set $\mathcal{S}$ in $\mathbb{R}^{d_s}$ as follows: For any vector $q = (q_1, ..., q_d)$ of q_d integers, let D^q denote the differential operator $D^q = \frac{\partial^{q.}}{\partial s_1^{q_1} ... \partial s_d^{q_d}}$. Denote $q. = \sum_{l=1}^d q_l$ and α to be the greatest integer strictly smaller than α. Let $\|g\|_\alpha = \max_{q.\leq \underline{\alpha}} \sup_s |D^q g(s)| + \max_{q.\leq \underline{\alpha}} \sup_{s\neq s'} |D^q g(s) - D^q g(s')| / \|s - s'\|^{\alpha - \underline{\alpha}}$ where $\max_{q.\leq\underline{\alpha}}$ denotes the maximum over $(q_1, ..., q_d)$ such that $q. \leq \underline{\alpha}$ and the suprema are taken over the interior of $\mathcal{S}$. Then $\mathcal{C}_M^\alpha(\mathcal{S})$ is the set of all continuous functions $g : \mathcal{S} \subset \mathbb{R}^{d_s} \mapsto \mathbb{R}$ with $\|g\|_\alpha \leq M$.

4.1 Observable regressors

We estimate the partial mean process with observed regressors Λ by

$$\hat{F}_{Y(t)}(y; \Lambda, W) = \frac{1}{n} \sum_{i=1}^n \hat{F}_{Y|T\Lambda}(y|t, \Lambda_i) \cdot W_i$$

for $y \in \mathcal{Y}$ and $t \in \mathcal{T}$. The estimator is constructed by the second and third steps in Section 3. We begin by stating conditions on the object of estimation that will be used in showing the behavior of

[10] An alternative to the function space $\mathcal{C}_M^\alpha$, we could assume hight-level conditions based on covering number for a general function space as in Lemma A.1 in the Appendix.

the estimator. Let $\|\cdot\|_\infty$ be the sup-norm, i.e., $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$, where $\mathcal{X}$ is the support of X.

Assumption 5 (Complexity - Regression)

(i) *For each fixed $y \in \mathcal{Y}$ and $t \in \mathcal{T}$, $F_{Y|T\Lambda}(y|t,\cdot) \in \mathcal{C}^\alpha_{\mathcal{M}}(\mathbf{\Lambda})$, where $\alpha > d_\lambda/2$.*

(ii) *There exists a universal constant C satisfying a Hölder continuity condition: for any $y_1, y_2 \in \mathcal{Y}$, $t_1, t_2 \in \mathcal{T}$, $\left\|F_{Y|T\Lambda}(y_1|t_1,\cdot) - F_{Y|T\Lambda}(y_2|t_2,\cdot)\right\|_\infty \leq C\|(y_1,t_1)-(y_2,t_2)\|^{1/2}$.*

Assumption 5 implies $\{\Lambda \mapsto F_{Y|T\Lambda}(y|t,\Lambda) : y \in \mathcal{Y}, t \in \mathcal{T}\}$ is Donsker by Lemma A.1 in the Appendix and Example 19.9 in van der Vaart (2000). Assumption 5 (ii) specifies the complexity of the function space in y and t, following Ichimura and Lee (2010) whose unknown function is a process indexed by the parameter of interest. By the assumptions on our estimators, Section B.1 in the Appendix shows that the estimator $\hat{F}_{Y|T\Lambda}(y|t,\Lambda)$ satisfies Assumption 5 with probability approaching one ($w.p.a.1$) for any $y \in \mathcal{Y}$ and $t \in \mathcal{T}$.

Assumption 6 (Bandwidth)

The bandwidth h satisfies (i) $h \to 0$, (ii) $nh^{2r+d_t} \to 0$, and (iii) $nh^{2d+d_t}/\log(n) \to \infty$, as $n \to \infty$.

The following theorem presents the asymptotic linear representation and weak convergence of our estimator.

Theorem 1 (Weak Convergence)

Suppose Assumptions 2, 3, 4, 5, and 6 hold, where $\Delta_k \geq \alpha$ and $\Delta \geq \alpha + r$. Suppose the weight W is uniformly bounded. Suppose the derivatives of $\mathbb{E}[W|\Lambda]$ up to order r exist and are uniformly bounded and continuous. Then uniformly in $y \in \mathcal{Y}$ and $t \in \mathcal{T}$

$$\sqrt{nh^{d_t}}\Big(\hat{F}_{Y(t)}(y;\Lambda,W) - F_{Y(t)}(y;\Lambda,W)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{tin}(y;\Lambda,W) + o_p(1)$$

where the influence function

$$\psi_{tin}(y;\Lambda,W) \equiv \sqrt{h^{d_t}}\,\frac{K_h(T_i - t)}{f_{T|\Lambda}(t|\Lambda_i)} \cdot \Big(\mathbf{1}_{\{Y_i \leq y\}} - F_{Y|T\Lambda}(y|t,\Lambda_i)\Big) \cdot \mathbb{E}\big[W\big|\Lambda = \Lambda_i\big]. \tag{4}$$

Then for any fixed $t \in \mathcal{T}$

$$\sqrt{nh^{d_t}}\Big(\hat{F}_{Y(t)}(\cdot;\Lambda,W) - F_{Y(t)}(\cdot;\Lambda,W)\Big) \Rightarrow \mathbb{G}_t(\cdot;\Lambda,W)$$

the empirical process converges weakly to a Gaussian process $\mathbb{G}_t(\cdot;\Lambda,W)$ with mean zero and the covariance kernel $Cov\big(\mathbb{G}_t(y_1;\Lambda,W), \mathbb{G}_t(y_2;\Lambda,W)\big) = \lim_{n\to\infty}\mathbb{E}\big[\psi_{tin}(y_1;\Lambda,W)\psi_{tin}(y_2;\Lambda,W)\big]$

$$= \mathbb{E}\left[\Big(F_{Y|T\Lambda}(\min\{y_1,y_2\}|t,\Lambda) - F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|t,\Lambda)\Big)\frac{\mathbb{E}[W|\Lambda]^2}{f_{T|\Lambda}(t|\Lambda)}\right]\int K^2(v)dv$$

for any $y_1, y_2 \in \mathcal{Y}$.

The convergence rate of the partial mean estimator is $\sqrt{nh^{d_t}}$ depending on d_t, the dimension of the continuous conditioning variables that are fixed in the outer expectation. It is known that when more conditioning variables are averaged out in the partial mean, the nonparametric estimator converges at a faster rate (Newey, 1994b). A *Full Mean* estimator is the case when all arguments of the regressors

are averaged out by the outer expectation and hence converges at root-n rate; for example, average derivative in Powell, Stock, and Stoker (1989) and Corollary 3 in this paper. This important generic feature of the convergence rate for partial mean and full mean estimators is invariant to the nonparametric estimators considered. For example, Belloni, Chernozhukov, Chetverikov, and Kato (2013) derive uniform rates for linear functionals of series estimators.

Assumption 6 reflects two common features of nonparametric estimation in an intermediate step: curse of dimensionality and under-smoothing by a bias-reducing kernel. The bias in finite sample is larger at the points where the counterfactual distribution has more curvature with respect to t; see Eq. (18) in the Appendix. Although a bandwidth selection procedure is beyond the scope of this paper, a rule of thumb method satisfying the bandwidth Assumptions can be implemented in practice. Su and Ullah (2008) propose a plug-in method by minimizing the asymptotic integrated mean squared error. The stochastic expansion in EJL14 is uniform over the bandwidth which allows the use of a data-driven bandwidth choice procedure.

4.2 Generated regressors

This section presents the asymptotic theory for nonparametric estimation of the partial mean process with generated regressors $\Lambda(S_\lambda) = (X', v_0(T,S)')'$ in Eq. (1), where $S_\lambda = (X', T', S')'$, $v_0(T,S)$ is a vector of measurable functions of observables $S \subset (X, Z)$ and it could contain the treatment T or not. The generated regressor $V \equiv v_0(T,S)$ takes values in $\mathcal{V} \subset \mathbb{R}^{d_v}$. The generated regressor is estimated in the first step, denoting $\hat{V}_i \equiv \hat{v}(T_i, S_i)$. Then for $y \in \mathcal{Y}$ and $t \in \mathcal{T}$,

$$\hat{F}_{Y(t)}\big(y; (X, \hat{V}), W\big) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}_{Y|TX\hat{V}}\big(y|t, X_i, \hat{v}(T_i, S_i)\big) \cdot W_i.$$

The following assumptions require the first-step estimator $\hat{V}$ to converge fast enough and to take values in a function space that is not too complex $w.p.a.1$. There is a tradeoff between the complexity and accuracy assumptions as in MRS12 and EJL14: if the control function is smoother, i.e., it belongs to a less complex function space, then estimation of the generated regressor needs to converge at a faster rate. The complexity of the function space is measured by the cardinality of the covering sets or the packing number, which can be achieved by assuming smoothness of the functions. Primitive sufficient conditions are given in Section B. The key for the asymptotic theory is a stochastic equicontinuity argument from empirical process theory, modified from Lemma 1 in MRS12. So the following high-level assumptions are borrowed from MRS12.

Assumption 7 (Accuracy)
Let the second-step bandwidth $h_2 \sim n^{-\eta}$ with $d_2\eta < 1$. The j-th component $\hat{v}_j$ and v_{0j} of vectors $\hat{v}$ and v_0, respectively, satisfies $\|\hat{v}_j - v_{0j}\|_\infty = o_p(n^{-\delta})$, for some $\delta > 2\eta$ and for all $j = 1, ..., d_v$. [11]

Assumption 8 (Complexity - Generated Regressors)
There exist a sequence of sets of functions $\mathcal{M}_n$ such that

(i) *$Pr(v_{0j} \in \mathcal{M}_n) \to 1$ and $Pr(\hat{v}_j \in \mathcal{M}_n) \to 1$ as $n \to \infty$ for all $j = 1, ..., d_v$.*

[11] We strengthen the Accuracy assumption compared with $\delta > \eta$ in MRS12; see the proof of Lemma **??** in the Appendix.

(ii) *For a constant $C_M > 0$ and a function v_{nj} with $\|v_{nj} - v_{0j}\|_\infty = o(n^{-\delta})$, the set $\bar{\mathcal{M}}_n = \mathcal{M}_n \bigcap \{u_j : \|v_{nj} - u_j\|_\infty \leq n^{-\delta}\}$ can be covered by at most $C_M \exp(\varrho^{-\beta} n^{\xi})$ balls with $\|\cdot\|_\infty$-radius ϱ for all $\varrho \leq n^{-\delta}$, where $0 < \beta < 2$ and $\xi \in \mathbb{R}$.* [12]

Let $f_V(v)$ be the Lebesgue density of V evaluated at $v \in \mathcal{V}$. Let $f_{T|V}(t|v)$ be the density with respect to a σ-finite measure μ_V of T conditional on V, evaluated at $v \in \mathcal{V}$ and $t \in \mathcal{T}$. The influence function of the oracle or infeasible estimator with the true regressor V is Eq. (4) derived in Theorem 1,

$$\psi_{tin}\big(y;(X,V),W\big) \equiv \frac{\sqrt{h_2^{d_t}}\ K_{h_2}\big(T_i - t\big)}{f_{T|XV}(t|X_i,V_i)}\Big(\mathbf{1}_{\{Y_i \leq y\}} - F_{Y|TXV}(y|t,X_i,V_i)\Big) \cdot \mathbb{E}\big[W\big|X = X_i, V = V_i\big].$$

For any two functions $v_1, v_2 \in \bar{\mathcal{M}}_n$, define

$$\Delta_{ARG}\big(y,t,v_1,v_2\big) \equiv \mathbb{E}\big[\big(v_1(T,S) - v_2(T,S)\big)' ARG\big(y,t,X,W,v_2(T,S)\big)\big]$$
$$\Delta_{REG}\big(y,t,v_1,v_2\big) \equiv \mathbb{E}\big[\big(v_1(T,S) - v_2(T,S)\big)' REG\big(y,t,X,S,v_2(T,S)\big)\big]$$

$$\text{where}\quad ARG\big(y,t,X,W,V_2\big) \equiv \nabla_v F_{Y|TXV}(y|t,X,v)\Big|_{v=V_2} \cdot W$$

$$REG\big(y,t,X,S,V_2\big) \equiv \Big\{ - \nabla_v F_{Y|TXV}(y|t,X,v)\Big|_{v=V_2} \cdot \mathbb{E}\big[W\big|X,V=V_2\big]$$
$$+ \Big(F_{Y|TXV}(y|t,X,V_2) - F_{Y|TXS}(y|t,X,S)\Big) \cdot \Big(- \nabla_v \mathbb{E}\big[W\big|X,V=v\big]\Big|_{v=V_2}$$
$$+ \frac{\nabla_v f_{T|XV}(t|X,v)\big|_{v=V_2}}{f_{T|XV}(t|X,V_2)} \cdot \mathbb{E}\big[W\big|X,V=V_2\big]\Big)\Big\}\ \frac{f_{T|XS}(t|X,S)}{f_{T|XV}(t|X,V_2)}$$

and $V_2 = v_2(T,S)$. *ARG* is the influence from estimating V as the *argument*. *REG* comes from estimating V as the *regressor*.[13] Given these assumptions and notations, the following Theorem states the main results.

Theorem 2 (Stochastic Expansion)
Suppose the conditions in Theorem 1 hold. Assume $F_{Y|TXV_1}(y|t,x,v)$ to be well-defined $\forall y \in \mathcal{Y}$ for any $V_1 = v_1(T,S)$ where $v_1 \in \bar{\mathcal{M}}_n$. Suppose Assumptions 7 and 8 hold. Then

(I) *when the generated regressors are not functions of the treatment T, $V = v_0(S)$, uniformly in $y \in \mathcal{Y}$ and $t \in \mathcal{T}$,*

$$\sqrt{nh_2^{d_t}}\Big(\hat{F}_{Y(t)}\big(y;(X,\hat{V}),W\big) - F_{Y(t)}\big(y;(X,V),W\big)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_{tin}\big(y;(X,V),W\big)$$
$$+ \sqrt{nh_2^{d_t}}\ \Big(\Delta_{ARG}\big(y,t,\hat{v}(S),v_0(S)\big) + \Delta_{REG}\big(y,t,\hat{v}(S),v_0(S)\big)\Big)$$
$$+ \sqrt{nh_2^{d_t}}\ R_n + o_p(1)$$

[12] We can allow the bandwidth, convergence rate, and the function space depending on each component of the generated regressors, i.e., η_j, δ_j, $\mathcal{M}_{n,j}$, β_j, and ξ_j for $j = 1, ..., d_v$, as in MRS12. We sacrifice this generality for notational ease.

[13] Note the underlying variables for the estimated control function S and X are allowed to overlap, so it should not be confusing that the conditioning variables for $F_{Y|TXS}$ are T and the union of S and X.

(Ⅱ) *when the generated regressors are functions of the treatment* T, $V = v_0(T,S)$, *uniformly in* $y \in \mathcal{Y}$ *and* $t \in \mathcal{T}$

$$\sqrt{nh_2^{d_t}}\Big(\hat{F}_{Y(t)}\big(y;(X,\hat{V}),W\big) - F_{Y(t)}\big(y;(X,V),W\big)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}\big(y;(X,V),W\big)$$
$$+ \sqrt{nh_2^{d_t}}\ \Big(\Delta_{ARG}\big(y,t,\hat{v}(T,S),v_0(T,S)\big) + \Delta_{REG}\big(y,t,\hat{v}(t,S),v_0(t,S)\big)\Big)$$
$$+ \sqrt{nh_2^{d_t}}\ R_n + o_p(1)$$

where $R_n = O_p\big(n^{-\kappa_1} + n^{-\kappa_2} + n^{-r_2\eta}\big)$, $0 < \kappa_1 < \frac{1}{2}(1-d_2\eta) + (\delta-\eta) - \frac{1}{2}(\delta\beta+\xi)$, *and* $\kappa_2 < \min\big\{1-d_2\eta, 2(\delta-\eta)\big\}$.

The impact of the estimation error from the generated regressor is characterized by $\Delta_{ARG} + \Delta_{REG} = \mathbb{E}\big[\big(\hat{v}(\cdot) - v_0(\cdot)\big)'\big(ARG + REG\big)\big]$, where the first-step estimator for the generated regressor $\hat{v}_0(\cdot)$ is taken as a fixed function given the sample and the expectation is taken over the underlying variables (T,X,S,W). This general expression enables a further derivation by plugging in a linear representation of the estimation error $\hat{v} - v_0$. We emphasize three *elements* of our stochastic expansion that summarize the most important insight of the impact of generated regressors: (i) the partial mean/full mean structure of $\Delta_{ARG} + \Delta_{REG}$ that determines the convergence rate of the estimation error; (ii) *index bias* $F_{Y|TXV} - F_{Y|TXS}$ and *index ratio* $f_{T|XS}/f_{T|XV}$; (iii) the *projection error of the weight* $W - \mathbb{E}[W|X,V]$. These elements are generic for the marginal integration estimators that involve a kernel regression with generated regressors. So it is worth understanding in more detail in the following Remarks.

In Theorem 2 (Ⅱ) when the generated regressors are functions of the treatment T, such as the control variable in the triangular models in Imbens and Newey (2009), the treatment variable T is fixed at t in the influence of estimating V as a regressor in Δ_{REG}. Intuitively, this is because $\mathbb{E}[\mathbf{1}_{\{Y\le y\}}|T=t, v_0(T,S)=v] = \mathbb{E}[\mathbf{1}_{\{Y\le y\}}|T=t, v_0(t,S)=v]$. On the other hand, for estimating the argument $v_0(T,S)$, the expectation in the influence Δ_{ARG} averages over T, capturing variation of T in V. This is a distinctive and subtle feature when the estimated generated regressor is a function of the treatment variables whose values are fixed in the partial mean.

Remark 1 (Convergence rate of full mean and partial mean)
The convergence rate of the impact of the estimation error from the generated regressor $\Delta_{ARG} + \Delta_{REG}$ can be learned by the partial mean/full mean structure as discussed in Section 4.1.

1. **(Root-n rate)** When the randomness of the generated regressor $V = v_0(\cdot)$ comes from all the regressors of $v_0(\cdot)$, $\Delta_{ARG} + \Delta_{REG}$ is an expectation averaging over all the regressors of $\hat{v}(\cdot) - v_0(\cdot)$. So the impact of the estimation error is a full mean and converges at root-n rate. For example, the control variable $v_0(T,Z) = F_{T|Z}(T|Z)$ in Section 5.1. Another case is when the generated regressor is estimated parametrically, its estimation error converges at root-n rate. Consequently, when the partial mean in Eq. (1) is an intermediate step in a multi-step estimator for a semiparametric problem, the impact of generated regressors is of first order.
2. **(Nonparametric rate)** When the randomness of the generated regressor only comes from part of the regressors of $v_0(\cdot)$, the impact of the estimation error $\Delta_{ARG} + \Delta_{REG}$ becomes a partial mean that averages over part of the regressor of $\hat{v}(\cdot) - v_0(\cdot)$. So $\Delta_{ARG} + \Delta_{REG}$ converges at a nonparametric rate, when the generated regressor is estimated nonparametrically. An

example of this case is the generalized propensity score $V = v_0(X) = f_{T|X}(t|X)$ in Section 5.2, where the regressor T is fixed at t. If the generated regressor is estimated by a kernel method and $\Delta_{ARG} + \Delta_{REG} = O_p((nh_1^{d_t})^{-1/2})$, then choosing $h_2 = o(h_1)$ could artificially make the estimation error of smaller order.

Remark 2 (Index)
The generated regressors can be viewed as an index function of obsersables S. The index bias $F_{Y|TXV} - F_{Y|TXS}$ is the impact of the underlying variable S on the outcome distribution that is not captured by the "index" $v_0(t, S)$. The index ratio $f_{T|XS}/f_{T|XV}$ is the impact of S on the treatment density that is not captured by $v_0(t, S)$. For the examples of the control variable in Section 5.1 and the propensity score in the sample selection model in Section 5.3, the index bias is zero. It is also known as *index sufficiency* in Heckman, Urzua, and Vytlacil (2006) in the sense that everywhere S enters the model only through the index $v_0(t, S)$. The index ratio is one when the generated regressor is the generalized propensity score $f_{T|X}(t|X)$ in the continuous treatment effects model in Section 5.2.

The index bias and ratio come from the second-step regression that averages over the underlying variable S of the generated regressor and the regressors (T, X). It results in the distribution functions conditional on S, instead of V that is taken as a given index function v_0. The existing literature has known the index bias. The index ratio is a special feature of the third-step partial mean where the treatment is fixed. Other than this paper, the index ratio is recently found for a discrete treatment in Mammen, Rothe, and Schienle (2013) and Hahn and Ridder (2013).

Remark 3 (Projection of the weight)
Even though the weight W is not estimated, it plays a role in the influence of estimating the generated regressors. The projection of the weight $\mathbb{E}[W|X, V]$ comes from the third-step partial mean, where the *expectation* averages over the arguments (X, V) and the additional variable W. The arguments (X, V) in the second-step kernel regression can be viewed as regressors for W in the third step. Together with ARG, the curvature of the regression function in the generated regressor, $\nabla_v F_{Y|TXV}(y|t, x, v)$, affects the influence of the estimation error through the projection error of the weight $W - \mathbb{E}[W|X, V]$. This is a key element of the third-step summation, which also appears in the semiparametric full mean in Mammen, Rothe, and Schienle (2013), Hahn and Ridder (2013), and EJL14.

Remark 4 (Transformed outcome)
Theorem 2 is readily extended to a transformed dependent variable $A_\alpha(Y)$ replacing the dependent variable $\mathbf{1}_{\{Y \leq y\}}$. Consider the parametric family $\{A_\alpha : \alpha \in \mathcal{A}\}$ composed of uniformly bounded functions A_α indexed by a finite-dimensional parameter $\alpha \in \mathcal{A}$, a compact set in $\mathbb{R}$, and A_α is Lipschitz continuous in α. Theorem 2 provides a stochastic expansion of the partial mean with generated regressors $n^{-1} \sum_{i=1}^n \hat{\mathbb{E}}\big[A_\alpha(Y)|T = t, X = X_i, \hat{V} = \hat{V}_i\big] \cdot W_i$ uniformly over $\alpha \in \mathcal{A}$ and $t \in \mathcal{T}$. This result can be applied to various contexts, for example, semiparametric transformation model in Vanhems and Van Keilegom (2013). Song (2008) derives the uniform convergence rate of series estimators for the regression $\mathbb{E}[A_\alpha(Y)|V(S) = v]$ over an infinite-dimensional α, V, and v. The extension of our result to an infinite-dimensional space $\mathcal{A}$ is possible but beyond the scope of this paper.

More explicitly, the influence function of the infeasible estimator with the true regressor V is

modified from $\psi_{tin}\big(y;(X,V),W\big)$ in Eq. (4) derived in Theorem 1,

$$\frac{\sqrt{h_2^{d_t}}\ K_{h_2}\big(T_i-t\big)}{f_{T|XV}(t|X_i,V_i)}\Big(A_\alpha(Y_i)-\mathbb{E}[A_\alpha(Y)|T=t,X=X_i,V=V_i]\Big)\cdot\mathbb{E}\big[W\big|X=X_i,V=V_i\big].$$

We can replace the regression $F_{Y|TXV}$ with $\mathbb{E}[A_\alpha(Y)|T,X,V]$ in the influence of the generated regressors ARG and REG in Theorem 2. Similarly, Assumption 3 (iii) and Assumption 5 are modified for $\mathbb{E}[A_\alpha(Y)|T=t,\Lambda=\cdot]$.

The main technical challenge is to analyze the estimation error contributed by the regressor role of the generated regressor $n^{-1}\sum_{i=1}^n \hat{F}_{Y|TXV_1}(y|t,x_i,v_i)-\hat{F}_{Y|TXV_2}(y|t,x_i,v_i)$ for $V_1=v_1(S)$ and $V_2=v_2(S)$, $v_1,v_2\in\bar{\mathcal{M}}_n$. Heuristically, we first derive a stochastic equicontinuity argument for the second-step kernel regression estimator in Lemma A.2 in the Appendix. Another stochastic equicontinuity argument is implemented on the third-step partial mean. So the summations in both the second-step regression estimator and the third-step partial mean are replaced with expectations. By Fubini's theorem, interchanging the order of integration extracts the generated regressor from its *regressor* role to the *argument* of the kernel estimator. This is in the spirit of the U-statistic theory (Su and Ullah, 2008; Vanhems and Van Keilegom, 2013), but the empirical process approach allows for general first-step generated regressors. Finally, a standard Taylor series expansion linearizes the partial mean in the estimation error of the generated regressor V_1-V_2.

Lemma A.2 is an argument of stochastic equicontinuity of a Nadaraya-Watson kernel regression estimator in the regressor. Lemma A.2 can be of independent interest under different contexts. Previous work, such as Hahn and Ridder (2013), Song (2008), EJL14, analyzes limiting properties for general regression functions with generated regressor, so they usually impose a high-level continuity assumption on the regression function with respect to the regressors. For example, EJL14 assume a Lipschitz condition: $\sup_v\big|\mathbb{E}[Y|v_1(S)=v]-\mathbb{E}[Y|v_2(S)=v]\big|\leq C\|v_1-v_2\|_\infty$ for some positive constant C. EJL14 derive a stochastic expansion for a full mean by an argument of stochastic equicontinuity in both the regression function and the generated regressor. [14] The similar Lipschitz condition has been used in Chen, Linton, and Van Keilegom (2003) considering a general class of estimators for unknown functions. To implement the theoretical path-derivative approach, Hahn and Ridder (2013) assume the derivative of the regression function $\mathbb{E}[Y|T=t,V=v]$ with respective to the regressor V exists. Song (2012) and Hahn and Ridder (2013) also provide sufficient conditions for this high-level smoothness assumption. In contrast, we derive our stochastic expansion based on a Nadaraya-Watson kernel regression directly. So we only impose continuity assumption on the kernel function, but not on the regression function with respect to the regressor. Note that the stochastic expansion in Theorem 1 is not uniform over the regressor and hence neither is Theorem 2.

MRS12 and Mammen, Rothe, and Schienle (2013) use a weaker Lipschitz continuity assumption based on the index bias instead of the regression function. Our Lemma A.2 serves as a crucial intermediate step for Theorem 2, but does not aim to derive the impact of estimating the generated regressor on the second-step regression estimator. More specifically, we implement a Taylor series expansion only on the third-step partial mean, but not on the second-step regression estimator. Consequently, the in-

[14] EJL14 need this smoothness assumption to obtain the desired covering number to follow the proof of Lemma B.2 in Ichimura and Lee (2010). We use a similar approach only on the second-step regression function in Theorem 1, but not for the generated regressor.

termediate step of Lemma A.2 does not lose information from the approximation error of linearization and needs not the Lipschitz continuity assumption. This distinguishes our results with Theorem 1 in MRS12, which focus on characterizing how the estimation error of the generated regressor affects the second-step *regression estimator*.

5 Examples

The usefulness of the general estimation procedure and asymptotic theory is illustrated by, but not limited to, three economic examples in this section. The three elements — partial mean/full mean, index, and projection of the weight — critically determine the influence of estimating the generated regressors. The generated regressors can be estimated by (semi)parametric or nonparametric methods satisfying the Assumptions and with some uniform linear representation, although we focus on Nadaraya-Watson kernel estimators for these examples. The primitive Assumptions 10, 11, and 12 are in the Appendix.

5.1 Control variables in triangular models

Consider the nonseparable outcome equation $Y = \phi(T, \epsilon)$, where the treatment vector of interest $T = (T_1, T_2')'$ contains a single endogenous variable T_1. The remaining treatment subvector T_2 is exogenous. Assume a valid control variable V for T_1 that satisfies the Conditional Independence Assumption 1. Assume a nonseparable first stage equation for the treatment variable: $T_1 = g(Z, e)$, where the function g is strictly monotonic in the second argument. The instrumental vector Z is independent of (ϵ, e). The disturbance e is a continuously distributed scalar with cdf strictly increasing on the support of e. [15] [16] Imbens and Newey (2009) construct a control variable by $V = F_{T_1|Z}(T_1|Z)$. For a separable first stage equation, $T_1 = g(Z) + e$, Newey, Powell, and Vella (1999) use the reduced form residual $V = T_1 - \mathbb{E}[T_1|Z]$. Lee (2007) specifies a quantile regression first stage: $T_1 = Q_\tau(T_1|Z) + e$, so a valid control variable is the quantile regression residual.

The proposed estimator is applied to the triangular simultaneous equations models in Newey, Powell, and Vella (1999) and Imbens and Newey (2009). In these examples, the weak convergence of the partial mean estimator is first-order equivalent to the Gaussian process as if the true control variable was observed. The insight on the full mean of the estimation error and the stochastic expansion in the following Corollary 1 are novel to the literature. Our uniform expansion of the estimated control variables can be applied to the case when the partial mean is an intermediate step in a more complicated estimation procedure. The exclusion assumption of the instrumental variable implies the index bias is zero $F_{Y|TZ}(y|t, Z) = F_{Y|TV}(y|t, v_0(t_1, Z))$. Theorem 2 (II) implies the influence of the estimation error

[15] In general, we can include additional covariates in the outcome equation $\phi(T, X, \epsilon)$ such that $T \perp \epsilon|(X, V)$. There can be endogenous observables $X_1 \subset X = (X_1', X_2')'$ in the outcome equation ϕ, but are excluded from the first stage equation of T_1. And the exogenous X_2 is a subvector of $Z = (X_2', Z_T')'$, where Z_T is the excluded exogenous instrumental vector for T_1. The results in Corollary 1 hold with additional conditioning variables X.

[16] As discussed in Section 3, the first trimming function in the second step (Regression) is based on the compact subsupport of Z. The second trimming function in the third step (Parital Sum) is based on the compact subsupport of (T, V). So the subpopulation is selected so that their values of instrumental variables, treatments, characteristics, and the unobservable in the first stage equation do not take extreme values. Using the two fixed trimming functions, the identification argument is still valid for the subpopulation.

of the generated regressors $ARG + REG$ denoted by

$$A(y,t,T,Z,W) \equiv \nabla_v F_{Y|TV}(y|t,v)\Big|_{v=v_0(T_1,Z)} \cdot W - \nabla_v F_{Y|TV}(y|t,v)\Big|_{v=v_0(t_1,Z)} \cdot \mathbb{E}\big[W\big|V = v_0(t_1,Z)\big] \cdot \frac{f_{T|Z}(t|Z)}{f_{T|V}(t|v_0(t_1,Z))}.$$

Corollary 1 (Control Variable)

Assume the conditions in Theorem 2 and Assumptions 10 and 11 hold.

(I) *The control variable is estimated by a nonparametric kernel method using the r_1-order kernel satisfying Assumption 4. The bandwidth $h_1 \sim n^{-g} \to 0$ satisfies $g < (2\xi+\beta)/(d_1\beta + 2d_s)$.* [17]

1. *For $V = v_0(T_1,Z) = F_{T_1|Z}(T_1|Z)$ in Imbens and Newey (2009), the smoothed kernel estimator $\hat{V}_i = \hat{v}(T_{1i},Z_i) = \hat{F}_{T_1|Z}(T_{1i}|Z_i) = n^{-1}\sum_{j=1}^n G_{h_1}(T_{1j}-T_{1i})K_{h_1}(Z_j - Z_i)\big/\hat{f}_Z(Z_i)$ where $G_{h_1}(u) \equiv \int^u K_{h_1}(v)dv$.* [18] *The influence from estimating the control variable in Theorem 2 (II) is $\Delta_{ARG} + \Delta_{REG} =$*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\Big(\mathbf{1}_{\{T_{1i}\le T_1\}} - F_{T_1|Z}(T_1|Z_i)\Big)\cdot A(y,t,T,Z_i,W)\Big|Z = Z_i\Big] + O_p(h_1^{r_1} + \|R_n^v\|_\infty)$$
$$= O_p(n^{-1/2} + h_1^{r_1} + \|R_n^v\|_\infty)$$

uniformly in $y \in \mathcal{Y}$ and $t \in \mathcal{T}$, where $\|R_n^v\|_\infty = O_p\big(\big((\log n/nh_1^{d_z}\big)^{1/2} + h_1^{r_1}\big)^2\big)$.

2. *For $V = v_0(T_1,Z) = T_1 - \mathbb{E}[T_1|Z]$ in Newey, Powell, and Vella (1999), consider the estimator $\hat{V}_i = \hat{v}(T_{1i},Z_i) = T_{1i} - \hat{\mathbb{E}}[T_1|Z_i] = T_{1i} - n^{-1}\sum_{j=1}^n T_{1j}K_{h_1}(Z_j - Z_i)/\hat{f}_Z(Z_i)$. Then the influence from estimating the control variable is $\Delta_{ARG}+\Delta_{REG} = -n^{-1}\sum_{i=1}^n \big(T_{1i} - \mathbb{E}[T_1|Z = Z_i]\big)\cdot\mathbb{E}\big[A(y,t,T,Z_i,W)\big|Z=Z_i\big] + O_p(h_1^{r_1} + \|R_n\|_\infty) = O_p(n^{-1/2} + h_1^{r_1} + \|R_n^v\|_\infty)$ uniformly in $y \in \mathcal{Y}$ and $t \in \mathcal{T}$, where $\|R_n^v\|_\infty = O_p\big(\big((\log n/nh_1^{d_z})^{1/2} + h_1^{r_1}\big)^2\big)$.*

The smaller order terms $O_p\big(h_1^{r_1} + \|R_n^v\|_\infty\big) = o_p(n^{-1/2})$ can be controlled by assuming $(2r_1)^{-1} < g < (2d_1)^{-1}$.

(II) *Consider the models of Imbens and Newey (2009) and Newey, Powell, and Vella (1999) in (I). Consider the cases when $\hat{V}$ is a (i) parametric estimator; or (ii) a nonparametric kernel estimator using the r_1-order kernel satisfying Assumption 4. The bandwidth $h_1 \sim n^{-g} \to 0$ satisfies $\frac{1-\eta d_t}{2r_1} < g < \min\big\{\frac{1+\eta d_t}{2d_1}, (2\xi+\beta)/(d_1\beta+2d_s)\big\}$.* [19] *Then uniformly in $y \in \mathcal{Y}$ and $t \in \mathcal{T}$*

$$\sqrt{nh^{d_t}}\Big(\hat{F}_{Y(t)}\big(y;\hat{V},W\big) - F_{Y(t)}\big(y;V,W\big)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{tin}\big(y;V,W\big) + o_p(1).$$

For any $t \in \mathcal{T}$, $\sqrt{nh^{d_t}}\big(\hat{F}_{Y(t)}(\cdot;\hat{V},W) - F_{Y(t)}(\cdot;V,W)\big) \Rightarrow \mathbb{G}_t(\cdot;V,W)$ a Gaussian process defined in Theorem 1.

[17] This is a sufficient condition for Assumption 10 (iii) to ensure the first-step estimator $\hat{V}$ converging to $\mathcal{M}_n$ *w.p.a.*1 by Result A.1. Also $\delta = \min\big\{(1-gd_1)/2, gr_1\big\}$. The first stage equation is additive in Newey, Powell, and Vella (1999), $d_s = d_z$. So the condition is weaker than the nonseparable first stage equation where $d_s = d_z + 1$ in Imbens and Newey (2009).

[18] Because the generated regressor is a function of both T_1 and Z, the stochastic equicontinuity argument requires the cdf estimator to be smooth in both T_1 and Z.

[19] The smaller order terms $O_p\big(h_1^{r_1} + \|R_n^v\|_\infty\big) = o_p\big((nh_2^{d_t})^{-1/2}\big)$ by assuming $\frac{1-\eta d_t}{2r_1} < g < \frac{1+\eta d_t}{2d_1}$.

In Corollary 1 (I), the full mean of the influence from estimating the control variables results in a sample average of *i.i.d.* mean zero random variables and hence converges at root-n rate. For estimating the triangular model in Newey, Powell, and Vella (1999), Su and Ullah (2008) and MRS12 also find that the estimation error of the reduced form residual is first-order ignorable. We recognize it is essentially the full mean structure that leads to this result. In contrast to the marginal integration estimators, Newey, Powell, and Vella (1999) use a two-step series estimator to exploit their additive structural function and characterize the estimation error in the asymptotic variance.

5.2 Generalized propensity score

Under the unconfoundedness assumption, Hirano and Imbens (2004) show that regressing on the generalized propensity score (GPS) $f_{T|X}(t|X)$ is sufficient for estimating continuous treatment effects. The propensity score in the binary treatment case is often used for dimension-reduction to avoid the need to match units on the values of all covariates. It also results in a weaker smoothness assumption on the distribution functions and the kernel. In practice, it is easier to check the common support assumption by projection on the GPS than on the support of the covariates X, as discussed in Flores, Flores-Lagunes, Gonzalez, and Neumann (2012).

Consider one continuous treatment variable $d_t = 1$. Define $v_0(t,x) = f_{T|X}(t|x)$ and $\Lambda = V = v_0(t,X)$. The insight from Theorem 2 is that the estimation error of the GPS is a partial mean. The key to the following results is the property of the GPS $V = f_{T|X}(t|X) = f_{T|V}(t|v_0(X))$ and hence the index ratio $f_{T|X}/f_{T|V} = 1$ and $\frac{\partial}{\partial v} f_{T|V}(t|v) = 1$. We first present the limiting property when the nonparametric estimation of the GPS is not first-order ignorable. Second, when the GPS is estimated parametrically or nonparametrically with a faster convergence rate, the first-order asymptotic property is the same as if the true GPS was observed.

Corollary 2 (Generalized Propensity Score)
Suppose the conditions in Theorem 2 and Assumptions 10 and 11 hold.

(I) *Consider $\hat{V} = \hat{v}(X) = \hat{f}_{T|X}(t|X)$ to be a nonparametric kernel estimator with order $r_1 = r_2 = r$ and $h_1 = h_2 = h \sim n^{-\eta}$ satisfying Assumption 12 (ii).*

1. *For the overall distribution $F_{Y(t)}(y)$, Assumption 1 holds with $\Lambda = X$. Then uniformly over $y \in \mathcal{Y}$ and $t \in \mathcal{T}$*

$$\sqrt{nh}\Big(\hat{F}_{Y(t)}\big(y;\hat{V},W=1\big) - F_{Y(t)}(y)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}(y|X,W=1) + o_p(1).$$

For any $t \in \mathcal{T}$, $\sqrt{nh}\big(\hat{F}_{Y(t)}(\cdot;\hat{V},W=1) - F_{Y(t)}(\cdot)\big) \Rightarrow \mathbb{G}_t(\cdot;X,W=1)$ a Gaussian process defined in Theorem 1.

2. *For $W \neq 1$, uniform in $y \in \mathcal{Y}$, $t \in \mathcal{T}$*

$$\begin{aligned}
&\sqrt{nh}\Big(\hat{F}_{Y(t)}\big(y;\hat{V},W\big) - F_{Y(t)}\big(y;V,W\big)\Big) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}\big(y|V,W\big) + \sqrt{h}K_h(T_i - t)\nabla_v F_{Y|TV}(y|t,V_i)\Big(\mathbb{E}[W|X_i] - \mathbb{E}[W|V_i]\Big) \\
&+ \sqrt{h}K_h(T_i - t)\Big(\frac{\mathbb{E}[W|V_i]}{f_{T|X}(t|X_i)} - \nabla_v\mathbb{E}[W|V_i]\Big)\Big(F_{Y|TV}(y|t,V_i) - F_{Y|TX}(y|t,X_i)\Big) + o_p(1)
\end{aligned}$$

(Ⅱ) *Consider the cases* $\hat{v}(X) = \hat{f}_{T|X}(t|X)$ *is (i) a (semi)parametric estimator; or (ii) a nonparametric kernel estimator with* h_1 *satisfying* $h_2 = o(h_1)$ *and Assumption 12 (i). Then uniformly in* $y \in \mathcal{Y}$ *and* $t \in \mathcal{T}$,

$$\sqrt{nh_2}\Big(\hat{F}_{Y(t)}\big(y;\hat{V},W\big) - F_{Y(t)}\big(y;V,W\big)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}(y|V,W) + o_p(1).$$

For any $t \in \mathcal{T}$, $\sqrt{nh_2}\big(\hat{F}_{Y(t)}(\cdot;\hat{V},W) - F_{Y(t)}(\cdot;V,W)\big) \Rightarrow \mathbb{G}_t(\cdot;V,W)$ *a Gaussian process defined in Theorem 1.*

Remark 5

For estimating the overall distribution $F_{Y(t)}(y)$ using the GPS, Corollary 2 implies

1. Regression on the nonparametrically estimated GPS is first-order asymptotically equivalent to regressing on X, using the same bandwidth and kernel for both estimators. Using the GPS allows a weaker smoothness condition on the distribution functions.
2. In comparison with the estimator using the nonparametrically estimated GPS in (I) or the whole set of covariates X, the estimator in (Ⅱ) that regresses on a (semi)parametrically estimated GPS, or the nonparametric estimator with $h_2 = o(h_1)$, or the true GPS has (i) a larger asymptotic variance, (ii) a slower convergence rate, and (iii) a weaker regularity condition.

For estimating the effects for the treated population, $F_{Y(t)|T}(y|\bar{t})$, the above results do not hold for the weight $W = f_{T|X}(\bar{t}|X)/f_T(\bar{t})$. [20]

Remark 5-2 (i) comes from the fact that the whole set of observables X provides finer conditioning variables than its index $v_0(X)$. When the index bias is not zero, i.e., there exists x such that $F_{Y|TV}(y|t,v_0(x)) \neq F_{Y|TX}(y|t,x)$, the inequality between the corresponding asymptotic variances is strict. Lemma D.1 in the Appendix provides a formal and general result comparing expectations of the conditional variances given the whole set of observables X and given the index $v_0(X)$, respectively; Heckman, Ichimura, and Todd (1998) discuss the binary treatment case.

For Remark 5-2 (ii) and (iii), when the dimension of the regressors is smaller, Assumption 12 (i) requires less smooth distribution functions and lower-order kernels. Then it follows that the bandwidth converges to zero faster, which results in a slower convergence rate ($\sqrt{nh}$). A weaker regularity condition is the benefit of dimension reduction using the GPS, which is also noted in Song (2014) for the single-index nuisance parameters in the semiparametric models. However, a slower convergence rate is the additional cost for the nonparametric partial mean.

5.3 Nonparametric sample selection models with endogeneity

Consider the sample selection model with endogeneity in Das, Newey, and Vella (2003), the latent variable $Y^* = \phi(T) + \epsilon$, $T = g(Z) + e$, and $d = \mathbf{1}_{\{\alpha(Z)-\eta \geq 0\}}$, assuming Z is exogenous. The observed outcome is $Y = dY^*$. Das, Newey, and Vella (2003) show that the propensity score for selection and reduced form residuals lead to a control function method to account for both selection and endogeneity. Specifically, the generated regressor is $\hat{V}_i = \big(\hat{\mathbb{E}}[d|Z=Z_i], T_i - \hat{\mathbb{E}}[T|Z=Z_i]\big)'$. The partial mean with

[20] Hirano and Imbens (2004) do not show the identification for the cdf for the treated $\bar{t}$, $\mathbb{E}[\mathbf{1}_{\{Y(t)\leq y\}}|T=\bar{t}]$. We show it is identified by $F_{Y(t)}(y;v_0(t,X),W)$ by modifying the proof of Theorem 3.1 in Hirano and Imbens (2004).

generated regressor $n^{-1}\sum_{i=1}^{n}\hat{\mathbb{E}}[Y|d=1,T=t,\hat{V}=\hat{V}_i]$ consistently estimates the structural equation $\phi(t)$ up to an additive constant.

Theorem 2 implies the convergence rate is $(nh_2^{d_t})^{1/2}$, irrelevant of the dimension of Z. The estimation error of estimating the propensity score and the control function converges at a root-n rate and hence is first-order ignorable as in Section 5.1. Furthermore, the index bias $\mathbb{E}[Y|d=1,T=t,V(t,Z)]-\mathbb{E}[Y|d=1,T=t,Z]$ is zero and there is no additional weight ($W=1$). In contrast, Das, Newey, and Vella (2003) use a series estimator to utilize the additive structure. Their asymptotic results show that the estimation errors of the propensity score and control function affect the limiting distribution of their estimator.

6 Semiparametric Models

The stochastic equicontinuity arguments for Theorem 2 can be applied to a full mean process with generated regressors and a partial mean process for discrete multi-valued treatments. The important elements in the nonparametric partial mean, *index* and *projection of the weight* in Remarks 2 and 3, persist in all the following semiparametric models. We present the stochastic expansion of the three-step estimator modified from the procedure in Section 3. By further assuming regularity conditions, such as Donsker property, we could similarly derive the asymptotic normality and other distributional treatment effects for discrete treatments.

Because the semiparametric estimators converge at a faster root-n rate, the regularity conditions are more restrictive than in the nonparametric partial mean case. See Section B in the Appendix for more discussion. Mammen, Rothe, and Schienle (2013) derive sharp bounds on weighted integrals of the remainder terms instead of controlling its supremum norm as in MRS12 and this paper. Their sharp bounds allow weaker regularity assumptions. They also allow the regression function to depend on the finite-dimensional parameter of interest in general. We do not explore these extensions since the semiparametric models are not the focus.

6.1 Full mean

Consider the parameter of interest defined by a moment condition $\theta\equiv\mathbb{E}\big[\rho\big(m(X,V),W\big)\big]$, where ρ is a known uniformly bounded function of the regression $m(x,v)\equiv\mathbb{E}[Y|X=x,V=v)]$ and variables W. Hahn and Ridder (2013) use Newey (1994a)'s path-derivative method to derive the influence function for θ when the generated regressor V is estimated. We propose an estimator $\hat{\theta}=n^{-1}\sum_{i=1}^{n}\rho\big(\hat{\mathbb{E}}[Y|X=X_i,\hat{V}=\hat{V}(S_i)],W_i\big)$ following the three-step procedure described in Section 3.

Corollary 3 (Semiparametric GMM estimation))
Let the conditions in Theorem 2 hold with $d_t=0$. Assume the derivatives of $\mathbb{E}[\frac{\partial}{\partial m}\rho(m(X,V),W)|X=$

$x, V = v]$ *with respect to* (x, v) *up to order* r *exist and are uniformly bounded and continuous. Then*

$$\sqrt{n}\Big(\hat{\theta} - \theta\Big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg(\rho\big(m(X_i, V_i), W_i\big) - \theta + \Big(Y_i - \mathbb{E}[Y|X_i, V_i)]\Big) \cdot \mathbb{E}\Bigg[\frac{\partial \rho\big(m(X, V), W\big)}{\partial m} \Bigg| X = X_i, V = V_i \Bigg] \Bigg) + \sqrt{n}\ \Delta_\rho\big(\hat{v}(S)\big) + \sqrt{n} R_n + o_p(1)$$

where $\Delta_\rho\big(v_1(S)\big) \equiv \mathbb{E}\big[\big(v_1(S) - v_0(S)\big)' A(X, S, W)\big]$ *for any* $v_1 \in \bar{\mathcal{M}}_n$ *and* $A(X, S, W) \equiv$

$$\Bigg(\frac{\partial \rho\big(m(X, v_0(S)), W\big)}{\partial m} - \mathbb{E}\Bigg[\frac{\partial \rho\big(m(X, V), W\big)}{\partial m} \Bigg| X, v_0(S) \Bigg] \Bigg) \cdot \nabla_v \mathbb{E}[Y|X, V = v]\Big|_{v = v_0(S)} - \Big(\mathbb{E}[Y|X, v_0(S)] - \mathbb{E}[Y|X, S] \Big) \cdot \nabla_v \mathbb{E}\Bigg[\frac{\partial \rho\big(m(X, V), W\big)}{\partial m} \Bigg| X, V = v \Bigg] \Bigg|_{v = v_0(S)}.$$

As in Remark 4 in Hahn and Ridder (2013), consider the parameter of interest β *defined by the moment condition* $\theta = \mathbb{E}\big[\rho\big(\mathbb{E}[Y|X, v_0(S)], W, \beta\big)\big] = 0$. *Then the GMM estimator* $\hat{\beta}$ *follows* $\sqrt{n}\big(\hat{\beta} - \beta\big) = \Big(\mathbb{E}\Big[\frac{\partial}{\partial \beta'} \rho\big(\mathbb{E}[Y|X, v_0(S)], W, \beta\big)\Big]\Big)^{-1} \sqrt{n}\big(\hat{\theta} - \theta\big) + o_p(1)$.

The stochastic expansion is ready to derive the influence function by plugging in the linear representation of the estimation error $\hat{v}(S) - v(S)$ from a semi/non-parametric estimation. The influence function of the proposed estimator coincides with the theoretical findings in Hahn and Ridder (2013). The derivative $\frac{\partial}{\partial m} \rho\big(m(X, V), W\big)$ plays the role of the weight W in the full mean. As discussed in Section 4.2, the influence function from the generated regressor contains the projection error of the weight and the index bias. Because all the arguments are averaged out in the full mean, there is no index ratio as in the partial mean case. When there is no weight or ρ is linear in the regression m, the estimation error of the generated regressor is first-order ignorable.

Remark (Comparison with Escanciano, Jacho-Chávez, and Lewbel (2014))
The full mean $\mathbb{E}[\mathbb{E}[Y|X, V] \cdot W]$ is a special case of $\rho(m, w) = m \cdot w$. EJL14 study sums of weighted regression residuals and derive a uniform expansion over the weights, generated regressors, and bandwidth. Their result allows for random trimming and data-driven bandwidth choice. They handle the generated regressors and the estimated weight at the same time under the assumption of zero index bias, $\mathbb{E}[Y|X, v_0(S)] - \mathbb{E}[Y|X, S] = 0$. In contrast, we focus on the estimation error of the generated regressors given the known weight. Corollary 3 complements EJL14 by the following two points: First, we show the generated regressor does affect the limiting distribution through its *regressor* role by the term associated with the projection of the weight $\mathbb{E}[W|X, V]$. This point is not previously recognized. Second, we characterize the influence of estimating the generated regressors when the index bias is not zero. The reason why we reach the same results with EJL14 is that when the index bias is zero, the estimation error of the weight has no effect on the limiting distribution of their full mean of the residuals. [21] As discussed in Remark 3, the weight plays a role in the influence of the generated regressors, even though the weight is not estimated. So it is worthwhile to deal with estimating the weight and

[21] EJL14 analyze the full mean of the residuals $n^{-1} \sum_{i=1}^{n} \big(Y_i - \hat{\mathbb{E}}[Y|V = V_i]\big) \cdot W_i$ which is $n^{-1} \sum_{i=1}^{n} Y_i \cdot W_i - \mathbb{E}\big[\mathbb{E}[Y|V] \cdot W\big]$ subtracted by our weighted full mean, $n^{-1} \sum_{i=1}^{n} \hat{\mathbb{E}}[Y|V = V_i] \cdot W_i - \mathbb{E}\big[\mathbb{E}[Y|V] \cdot W\big]$. Some simple algebra shows our stochastic expansion for estimating the generated regressors $V = v_0(X)$ coincide when assuming the index bias is zero.

the generated regressor separately when the index bias is not zero. In such case, the estimation errors of the generated regressors and the weight both contribute additional terms to the influence function.

Remark (Policy effects in Rothe (2010) and Imbens and Newey (2009))
The proposed estimator is applicable to the policy effects studied in Rothe (2010) and Imbens and Newey (2009) when a control variable is the generated regressor to correct for endogeneity. Corollaries 3 and 1 suggest an interesting finding that the estimation error of the control variables constructed in Newey, Powell, and Vella (1999) or Imbens and Newey (2009) is first-order ignorable. This non-trivial finding is implied by the elements of our stochastic expansion: (i) the estimation error of the control variables is a full mean and hence converges at root-n, same as the policy effect estimators; (ii) the *index bias* is zero by the exclusion assumption of the instruments; (iii) *the projection error of the weight* is zero. More specifically, Rothe (2010) provides an inference method for the distributional policy effect of counterfactually changing the covariate distribution to some known $f_{X^*}(X)$, i.e., $\int F_{Y|X}(y|x)f_{X^*}(x)dx - F_Y(y)$. The full mean process is the counterfactual outcome distribution for some known covariate distribution $f_{X^*}(x)$, when choosing the weight to be $W = f_{X^*}(X)/f_X(X)$, i.e., $\mathbb{E}[F_{Y|X}(y|X)W] = \int F_{Y|X}(y|x)f_{X^*}(x)dx$. When there is no generated regressors, our stochastic expansion coincides with the asymptotic results in Rothe (2010). Rothe (2010) discusses to relax the exogenous assumption of the covariates by including an additional control variable. Then the estimator is modified by $n^{-1}\sum_{i=1}^n \hat{F}_{Y|X\hat{V}}(y|X_i^*, \hat{V}_i)$, where $\{X_i^*\}_{i=1,\dots,n}$ is an *i.i.d.* random sample from the given density f_{X^*}. Another application is the average policy effect for a known function l such that $X_i^* = l(X_i)$ in Imbens and Newey (2009). The proof of Proposition 1 in Rothe (2010) implies that this estimator is asymptotically equivalent to our estimator for $\mathbb{E}[F_{Y|XV}(y|X,V)W]$ with a known weight $W = f_{X^*V}(X,V)/f_{XV}(X,V)$. We present the stochastic expansion for this full mean special case in Corollary 6 in the Appendix.

6.2 Partial mean for discrete treatments

This section considers the partial mean when the treatment is a discrete random vector, taking values in a finite set $\mathcal{T}$. Theorem 2 applies to the discrete treatments case directly by setting $d_t = 0$, replacing the kernel $K_h(T_i - t)$ with $\mathbf{1}_{\{T_i=t\}}$, and replacing $f_{T|XV}(t|x,v)$ with the propensity score (PS) $P_t(x,v) \equiv Pr(T = t|X = x, V = v)$; we prove this result along with the following Corollary 4 in the Appendix. So we skip the repetition and present an important application on the regression estimator for the binary treatment effects using an estimated PS as the generated regressor $V = P_t(X)$.

Consider the generated regressor estimated by $\hat{V}_i = \hat{P}_t(X_i) = n^{-1}\sum_{j=1}^n \mathbf{1}_{\{T_j=t\}}K_{h_1}(X_j - X_i)/\hat{f}_X(X_i)$, where $\hat{f}_X(X_i) \equiv n^{-1}\sum_{j=1}^n K_{h_1}(X_j - X_i)$. Denote $p_{\bar{t}} \equiv Pr(T = \bar{t})$. The result in the following Corollary 4 for estimating the ATE where $W = 1$ coincides with Hahn and Ridder (2013) and Mammen, Rothe, and Schienle (2013). For the ATT, $\mathbb{E}[Y(t) - Y(\bar{t})|T = \bar{t}]$ or $F_{Y(t)|T}(y|\bar{t}) - F_{Y(\bar{t})|T}(y|\bar{t})$, choose the weight $W = \mathbf{1}_{\{T=\bar{t}\}}/p_{\bar{t}}$. That is, the ATT is estimated by $\hat{\mathbb{E}}[Y(t) - Y(\bar{t})|T = \bar{t}] = n^{-1}\sum_{i=1}^n \big(\hat{\mathbb{E}}[Y|T = t, \hat{V} = \hat{V}_i] - Y_i\big)W_i$.

Consider the case when the weight $W = W(X)$ is estimated by $\hat{W}(X)$. The following heuristic argument shows the estimation error of the weight has no effect on the limiting distribution, especially when the index bias is zero. A similar linearization and stochastic equicontinuity argument shows that the estimation error in $\hat{W}(X)$ is dominated by $n^{-1}\sum_{i=1}^n \big(Y_i - \mathbb{E}[Y|V = v_0(X_i)]\big)\cdot\big(\hat{W}(X_i) - W(X_i)\big) = \mathbb{E}\big[\big(Y - \mathbb{E}[Y|V]\big)\cdot\big(\hat{W}(X) - W(X)\big)\big] + o_p(1) = \mathbb{E}\big[\big(Y - \mathbb{E}[Y|X]\big)\cdot\big(\hat{W}(X) - W(X)\big)\big] + o_p(1) = o_p(1)$. The second equality is by assuming the index bias is zero, $\mathbb{E}[Y|V = v_0(x)] = \mathbb{E}[Y|X = x]$ for all $x \in \mathcal{X}$. The last equality is by the orthogonality of the error and functions of X.

Corollary 4 (Discrete Treatment - Propensity Score)
Suppose the conditions in Theorem 2 hold with $d_t = 0$. Suppose the bandwidth Assumption 12 (i) holds with $d_t = 0$. Then for $t \in \mathcal{T}$, uniformly in $y \in \mathcal{Y}$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\hat{F}_{Y|T\hat{V}}(y|t,\hat{V}_i)W_i - \mathbb{E}\big[F_{Y|TV}(y|t,V)W\big]\Big) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(F_{Y|TV}(y|t,V_i)W_i - \mathbb{E}\big[F_{Y|TV}(y|t,V)W\big] \\
&\qquad + \frac{\mathbf{1}_{\{T_i=t\}}}{P_t(X_i)}\Big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|TV}(y|t,V_i)\Big)\cdot\mathbb{E}\big[W\big|V=V_i\big]\Bigg) \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{T_i=t\}} - P_t(X_i)\Big)\Bigg\{\frac{\partial}{\partial v}F_{Y|TV}(y|t,v)\Big|_{v=V_i}\cdot\Big(\mathbb{E}[W|X=X_i] - \mathbb{E}\big[W\big|V=V_i\big]\Big) \\
&\qquad + \Big(F_{Y|TV}(y|t,V_i) - F_{Y|TX}(y|t,X_i)\Big)\Bigg(-\frac{\partial}{\partial v}\mathbb{E}\big[W\big|V=v\big]\Big|_{v=V_i} + \frac{\mathbb{E}\big[W\big|V=V_i\big]}{P_t(X_i)}\Bigg)\Bigg\} + o_p(1)
\end{aligned}$$

For the binary treatment case, $\mathbb{E}[W|V] = P_{\bar{t}}(X)/p_{\bar{t}} = \mathbb{E}[W|X]$ implies the projection error of the weight is zero and $\nabla_v\mathbb{E}[W|V=v] = -1/p_{\bar{t}}$. Therefore, the limiting property of our estimator coincides with the theoretical result in Hahn and Ridder (2013) that the regression estimator using a nonparametrically estimated PS for the ATT is efficient. The semiparametric efficient bounds for the binary ATE and ATT have been studied in Hahn (1998), Hirano, Imbens, and Ridder (2003), Firpo (2007), Chen, Hong, and Tarozzi (2008), Cattaneo (2010), among others. Our results confirm that regression on X and the nonparametrically estimated PS reaches the semiparametric efficiency bound for both the ATE and ATT. Regression on the parametrically estimated PS or the true PS is less efficient.

Furthermore, Corollary 4 suggests a new finding: the efficiency result might not hold when the discrete treatment variable takes on more than two values. This is because the projection of the weight $\mathbb{E}[W|V] = \mathbb{E}\big[P_{\bar{t}}(X)/p_{\bar{t}}\big|P_t(X)\big]$ is generally unknown for a multi-valued treatment. Corollary 4 suggests additional terms from the projection of the weight. For the continuous treatment, our estimator for the treatment effect on the treated shares the same property in Remark 5.

7 Inference for the Treatment Effects

Often the objects of ultimate interest are policy effects or inequality measures. Such objects can be expressed as functionals of the potential outcome distributions identified by the partial mean $F_{Y(t)}(y;\Lambda,W)$ and estimated in previous sections. The key to the distribution theory for a class of smooth functionals is the functional delta method for Hadamard-differentiable functionals. The results are illustrated by the mean and quantile operators. In this section, we let $\theta_t(y) = F_{Y(t)}(y;\Lambda,W)$ by suppressing (Λ,W) in the notation for brevity. The corresponding asymptotic theorem derived in previous sections provides the influence function and weak convergence: denoting as $\sqrt{nh^{d_t}}\big(\hat{\theta}_t - \theta_t\big) = n^{-1/2}\sum_{i=1}^{n}\psi_{tin} + o_p(1)$ and converges weakly to a Gaussian process $\mathbb{G}_t$. For the semiparametric models in Section 6, assuming the Donsker property of the influence function is sufficient for the functional

delta method. For example, we can conduct inference on policy effects or quantile treatment effects for discrete treatments in Corollary 4,

Assumption 9
The functional Γ defined over the distribution functions of potential outcomes is Hadamard differentiable. [22]

These Hadamard-differentiable functionals can be highly nonlinear functionals of the cdf, but admit a linear functional derivative. Weak convergence of the estimators will be implied by the functional delta method in empirical process theory. Assumption 9 is a high-level assumption that could impose restrictions or smoothness on the distribution functions of potential outcomes. In particular, when Γ is the τ-quantile operator on $\theta_t(y) = F_{Y(t)}(y; \Lambda, W)$, Γ is a generalized inverse $\theta_t^{-1} : (0,1) \to \mathcal{Y}$ given by $\theta_t^{-1}(\tau) = \inf\{y : \theta_t(y) \geq \tau\}$. Then Assumption 9 means $\theta_t(y)$ is continuously differentiable at the τth-quantile, with the derivative being strictly positive and bounded over a compact neighborhood. Additional assumptions might be needed for different policy functionals. For instance, Bhattacharya (2007) gives regularity conditions for Hadamard-differentiability of Lorenz and Gini functionals.

Theorem 3
Assume the conditions in the asymptotic theorem for $\hat{\theta}_t$ hold. Consider the parameter θ as an element of a parameter space $D_\theta \subset l^\infty(\mathcal{Y})$ with D_θ containing the true value θ_t. Suppose a functional $\Gamma(\theta)$ mapping D_θ to $l^\infty(\mathcal{W})$ is Hadamard differentiable in θ at θ_t with derivative Γ'_θ. Then

1. *(Functional Delta Method)*

$$\left|\sqrt{nh^{d_t}}\big(\Gamma(\hat{\theta}_t)(w) - \Gamma(\theta_t)(w)\big) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma'_\theta(\psi_{tin})(w)\right| = o_p(1)$$
$$\sqrt{nh^{d_t}}\big(\Gamma(\hat{\theta}_t)(w) - \Gamma(\theta_t)(w)\big) \Rightarrow \Gamma'_\theta(\mathbb{G}_t)(w) \equiv G(w)$$

   where G is a Gaussian process indexed by $w \in \mathcal{W}$ in $l^\infty(\mathcal{W})$, with mean zero and covariance kernel defined by the limit of the second moment of $\Gamma'_\theta(\psi_{tin})$.

2. *(Causal effects)*

$$\sqrt{nh^{d_t}}\begin{pmatrix}\hat{\theta}_t(\cdot) - \theta_t(\cdot)\\ \hat{\theta}_{\bar{t}}(\cdot) - \theta_{\bar{t}}(\cdot)\end{pmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\psi_{tin}(\cdot)\\ \psi_{\bar{t}in}(\cdot)\end{pmatrix} + o_p(1) \Rightarrow \mathbb{G}_{t\bar{t}}(\cdot)$$

   a Gaussian process with zero mean. The diagonal elements of the covariance matrix are the covariance matrix of $\mathbb{G}_t$ and $\mathbb{G}_{\bar{t}}$. And the off-diagonal terms are zero. Theorem 3 implies

$$\sqrt{nh^{d_t}}\Big(\Gamma(\hat{\theta}_t) - \Gamma(\hat{\theta}_{\bar{t}}) - \big(\Gamma(\theta_t) - \Gamma(\theta_{\bar{t}})\big)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\Gamma'_\theta(\psi_{tin}) - \Gamma'_\theta(\psi_{\bar{t}in})\Big) + o_p(1) \Rightarrow \mathbb{G}^{\Gamma}_{t\bar{t}}$$

[22] See, for example, van der Vaart (2000) for dentition: let Γ be a Hadamard-differentiable functional mapping from $\mathcal{F}$ to some normed space $\mathcal{E}$, with derivative Γ'_f, a continuous linear map $\mathcal{F} \mapsto \mathcal{E}$. For every $h_n \to h$ and $f \in \mathcal{F}$,

$$\lim_{u\to 0}\frac{1}{u}\Big(\Gamma(f + uh_n) - \Gamma(f)\Big) = \Gamma'_f(h).$$

a mean-zero Gaussian process with the covariance kernel is the summation of the covariance of $\Gamma'_\theta(\mathbb{G}_t)$ *and the covariance of* $\Gamma'_\theta(\mathbb{G}_{\bar{t}})$.

The above result gives the policy/inequality treatment effects of shifting the treatment from $\bar{t}$ to t, $\Gamma\big(\theta_t\big) - \Gamma\big(\theta_{\bar{t}}\big)$. The estimators of the distributional features at different treatment levels t and $\bar{t}$, $\Gamma\big(\theta_t\big)$ and $\Gamma\big(\theta_{\bar{t}}\big)$, are asymptotically uncorrelated.

7.1 Examples: mean and quantile

The mean for the cdf θ_t is $\Gamma(\theta_t) = \int_{\mathcal{Y}} u d\theta_t(u)$, which has the Hadamard derivative $\Gamma'(\theta) = \int u d\theta(u)$. Then the estimator is $\int y \, d\hat{\theta}_t(y)$ by replacing the dependent variable $\mathbf{1}_{\{Y \leq y\}}$ with Y in the estimation procedure described in Section 3. Alternatively, we can use the transformed outcome in Remark 4. Theorems 1 and 2 provide the asymptotic theory of estimating the mean $\mathbb{E}[Y(t)]$ by simply replacing $F_{Y|T\Lambda}(y|t, \Lambda_i)$ with $\mathbb{E}[Y|T = t, \Lambda = \Lambda_i]$ in the influence function $\psi_{tin}(\cdot; \Lambda, W)$ in Eq. (4), *ARG* and *REG* in Section 4.2.

The unconditional quantile function is inverted directly from the unconditional cdf. For the quantile process $\{Q_\tau : \tau \in (0,1)\}$ of the cdf θ_t, $Q_\tau \equiv \inf\{y : \theta_t(y) \geq \tau\}$. The following corollary gives the asymptotic theory of estimating unconditional quantile function of $Y(t)$ for the whole population assuming unconfoundedness and using control variables.

Corollary 5 (Quantile Process)
Assume the conditions in Theorem 3. Suppose $\sqrt{nh^{d_t}}\big(\hat{\theta}_t(\cdot) - \theta_t(\cdot)\big) = n^{-1/2}\sum_{i=1}^n \psi_{tin}(\cdot) + o_p(1) \Rightarrow \mathbb{G}_t(\cdot)$. *Assume* θ_t *is continuously differentiable with strictly positive derivative* $\frac{\partial}{\partial y}\theta_t(y)\big|_{y=Q_\tau} \equiv \theta'_t(Q_\tau)$. *Then the influence function for estimating the quantile process is* $\psi^Q_{tin}(\tau) \equiv -\psi_{tin}\big(Q_\tau\big)\big/\theta'_t(Q_\tau)$. *Therefore,*

$$\sqrt{nh^{d_t}}\big(\hat{Q}_\cdot - Q_\cdot\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi^Q_{tin}(\cdot) + o_p(1) \Rightarrow -\mathbb{G}_t\big(Q_\cdot\big)\Big/\theta'_t(Q_\cdot) \equiv \mathbb{G}^Q_t(\cdot)$$

where $\mathbb{G}^Q_t$ *is a Gaussian process indexed by* $\tau \in [a,b] \subset (0,1)$ *in the metric space* $l^\infty([a,b])$. *The Gaussian process* $\mathbb{G}^Q_t$ *has zero mean and covariance kernel, for any* $\tau_1 < \tau_2 \in [a,b]$, $Cov\big(\mathbb{G}^Q_t(\tau_1), \mathbb{G}^Q_t(\tau_2)\big) = \lim_{n\to\infty}\mathbb{E}\big[\psi^Q_{tin}(\tau_1)\psi^Q_{tin}(\tau_2)\big]$.

Remark (Quantile Structural Function)
Consider the τth-quantile function of $Y(t)$, $Q_\tau = Q_\tau(Y(t)) = F^{-1}_{Y(t)}(\tau)$. The conditioning variables are $\Lambda(S_\lambda) = V$, where the control variables V is estimated in Section 5.1. Corollaries 1 and 5 imply

$$\sqrt{nh^{d_t}}\big(\hat{Q}_\cdot(Y(t)) - Q_\cdot(Y(t))\big) \Rightarrow \mathbb{G}^Q_t(\cdot)$$

a Gaussian process with mean zero and covariance $Cov(\mathbb{G}^Q_t(\tau_1), \mathbb{G}^Q_t(\tau_2)) \equiv$

$$\mathbb{E}\bigg[\frac{1}{f_{T|V}(t|V)}\Big(F_{Y|TV}(Q_{\tau_1}|t,V) - F_{Y|TV}(Q_{\tau_1}|t,V)F_{Y|TV}(Q_{\tau_2}|t,V)\Big)\bigg]\frac{\int K^2(v)dv}{f_{Y(t)}(Q_{\tau_1})\; f_{Y(t)}(Q_{\tau_2})}.$$

7.2 Inference

The asymptotic theorems in the previous sections can be used to calculate pointwise confidence intervals. The pointwise influence function can be estimated by replacing unknown functions with consistent estimators. Then the covariance matrix can be estimated by the sample variance of the estimated influence functions. Alternatively, the covariance matrix can be estimated by a plug-in method that is a sample analogue with consistently estimated unknown functions. The procedure is standard and omitted for brevity.

Besides pointwise inference, we might be interested in testing a hypothesis involving a policy on the whole distribution: constant effect or stochastic dominance. We suggest using a multiplier method to simulate the empirical processes defined in Theorem 1, Corollary 1, and Corollary 2. The multiplier method has been used in Donald, Hsu, and Barrett (2012) and Donald and Hsu (2014) to simulate a distribution process. It is easy to perform asymptotically valid inference on distributional features defined by the Hadamard-differentiable functionals. Let $\{U_i\}_{i=1}^n$ be a sequence of *i.i.d.* random variables with mean zero and variance one, for example, $\mathcal{N}(0,1)$, independent of the data. The influence function ψ_t for the estimator $\hat{\theta}_t$ is estimated consistently by some estimator $\hat{\psi}_t$. Here, we require the estimator for the conditional outcome distribution in $\hat{\psi}_t$ to be monotone in y by using second-order kernel or a monotone transformation. The following theorem shows that $n^{-1/2}\sum_{i=1}^n U_i\hat{\psi}_{tin}(\cdot)$ simulates the asymptotic distribution of the estimator.

Theorem 4 (Multiplier CLT)
Assume the conditions in Theorem 1 or Corollary 1 or 2 which gives $\sqrt{nh^{d_t}}\big(\hat{\theta}_t(\cdot)-\theta_t(\cdot)\big) = n^{-1/2}\sum_{i=1}^n \psi_{tin}(\cdot)+ o_p(1) \Rightarrow \mathbb{G}_t(\cdot)$. *Then*

$$\mathbb{G}^M_{tin}(\cdot) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\hat{\psi}_{tin}(\cdot) \Rightarrow \mathbb{G}_t(\cdot)$$

conditional on sample path with probability approaching 1. For the Hadamard-differentiable functional Γ,

$$\Gamma'\Big(\mathbb{G}^M_{tin}(\cdot)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\Gamma'\big(\hat{\psi}_{tin}(\cdot)\big) \Rightarrow \Gamma'\big(\mathbb{G}_t(\cdot)\big).$$

8 Conclusion

We derive a stochastic expansion showing how the presence of generated regressors affects the limiting behavior of the three-step nonparametric estimator of the partial mean process defined in Eq. (1). The influence of estimating the generated regressors has three important elements: partial mean/full mean structure, index, and the projection of the weight. These elements provide insights on conditions under which the estimation error is ignorable. We explicitly estimate the average and quantile structural functions where the endogeneity is corrected by control variables or generalized propensity score. The uniform expansion and weak convergence theorems derived in this paper are readily applied to many inequality measures, such as the Gini coefficient and the Theil index. Our stochastic expansion accounting for the estimation error of the generated regressor is uniform over the treatment value t and the distributional threshold value y. Therefore, these results can be extended or applied to more complicated estimation procedure where the partial mean is an intermediate step, for example, test

for stochastic dominance in Lee, Linton, and Whang (2009) and Rothe (2010), the transformation model in Vanhems and Van Keilegom (2013), and the compensating variation in Bhattacharya (2013). Another interesting extension to conduct uniform inference on the partial mean process indexed by both y and t using the strong approximation results established in Chernozhukov, Lee, and Rosen (2013).

References

Agüero, J., M. Carter, and I. Woolard (2010). The impact of unconditional cash transfers on nutrition: The south african child support grant. working paper.

Altonji, J. G. and R. L. Matzkin (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–1102.

Andrews, D. W. (1994). Chapter 37 empirical process methods in econometrics. Volume 4 of *Handbook of Econometrics*, pp. 2247 – 2294. Elsevier.

Andrews, D. W. K. and X. Shi (2013). Inference based on conditional moment inequalities. *Econometrica 81*(2), 609–666.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2013). On the asymptotic theory for least squares series: pointwise and uniform results. Working paper.

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics 137*(2), 674–707.

Bhattacharya, D. (2013). Nonparametric welfare analysis for discrete choice. Working paper.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources 8*(4), pp. 436–455.

Blundell, R. and J. L. Powell (2003). *Endogeneity in Nonparametric and Semiparametric Regression Models*, Volume II. Cambridge University Press, Cambridge, U.K.

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

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics 36*(2), pp. 808–843.

Chen, X., O. Linton, and I. Van Keilegom (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica 81*(6), 2205–2268.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013). Intersection bounds: Estimation and inference. *Econometrica 81*(2), 667–737.

Darolles, S., J.-P. Florens, and E. M. Renault (2011). Nonparametric instrumental regression. *Econometrica 79*(5), 1541–1565.

Das, M., W. K. Newey, and F. Vella (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies 70*(1), 33–58.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica 64*(5), 1001–44.

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, Part 3*(0), 383–397.

Donald, S. G., Y.-C. Hsu, and G. F. Barrett (2012). Incorporating covariates in the measurement of welfare and inequality: methods and applications. *The Econometrics Journal 15*(1), C1–C30.

Escanciano, J. C., D. T. Jacho-Chávez, and A. Lewbel (2014). Uniform convergence of weighted sums of non and semiparametric residuals for estimation and testing. *Journal of Econometrics 178*(3), 426 – 443.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica 75*(1), 259–276.

Firpo, S. and C. Pinto (2011). Identification and estimation of distributional impacts of interventions using changes in inequality measures. Working paper.

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.

Flores, C. A. (2007). Estimation of dose-response functions and optimal doses with a continuous treatment. Technical report.

Flores, C. A., A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *The Review of Economics and Statistics 94*(1), 153–171.

Graham, B. S., G. W. Imbens, and G. Ridder (2014). Complementarity and aggregate implications of assortative matching: A nonparametric analysis. *Quantitative Economics 5*(1), 29–66.

Guerre, E., I. Perrigne, and Q. Vuong (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica 68*(3), 525–574.

Haerdle, W., P. Janssen, and R. Serfling (1988). Strong uniform consistency rates for estimators of conditional functionals. *The Annals of Statistics 16*(4), 1428–1449.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–332.

Hahn, J. and G. Ridder (2013). The asymptotic variance of semi-parametric estimators with generated regressors. *Econometrica 81*(1), 315–340.

Heckman, J. J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies 65*(2), 261–94.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006, August). Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics 88*(3), 389–432.

Hirano, K. and G. Imbens (2004). *The Propensity Score with Continuous Treatments*, Chapter 7. John Wiley and Sons.

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.

Ichimura, H. and S. Lee (2010). Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics 159*, 252–266.

Imbens, G. W. and W. K. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

Kluve, J., H. Schneider, A. Uhlendorff, and Z. Zhao (2012). Evaluating continuous training programs using the generalized propensity score. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(2), 587–617.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer: New York.

Lee, S. (2007). Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics 141*(2), 1131 – 1158.

Lee, S., O. Linton, and Y.-J. Whang (2009). Testing for stochastic monotonicity. *Econometrica 77*(2), 585–602.

Mammen, E., C. Rothe, and M. Schienle (2012). Nonparametric regression with nonparametrically generated covariates. *Annals of Statistics 40*, 1132–1170.

Mammen, E., C. Rothe, and M. Schienle (2013). Semiparametric estimation with generated covariates. working paper.

Masry, E. (1996). Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications 65*(1), 81 – 101.

Matzkin, R. (2007). *Nonparametric Identification*, Volume 6B, pp. 5307–5368. Amsterdam: Elsevier.

Newey, W. (1994a). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–1382.

Newey, W. K. (1994b). Kernel estimation of partial means and a general variance estimator. *Econometric Theory 10*(02), 1–21.

Newey, W. K., J. L. Powell, and F. Vella (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica 67*(3), 565–604.

Oaxaca, R. (1973, October). Male-female wage differentials in urban labor markets. *International Economic Review 14*(3), 693–709.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*(5), 1027–57.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Conference Board of the Mathematical Science: NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*(6), 1403–30.

Rothe, C. (2010). Nonparametric estimation of distributional policy effects. *Journal of Econometrics 155*, 56–70.

Sherman, R. (1994). Maximal inequalities for degenerate u-processes with applications to optimization estimators. *The Annals of Statistics 22*(1), 439–459.

Song, K. (2008). Uniform convergence of series estimators over function spaces. *Econometric Theory 24*, 1463–1499.

Song, K. (2012). On the smoothness of conditional expectation functionals. *Statistics & Probability Letters 82*(5), 1028–1034.

Song, K. (2014). Semiparametric models with single-index nuisance parameters. *Journal of Econometrics 178*(3), 471–483.

Sperlich, S. (2009). A note on non-parametric estimation with predicted variables. *Econometrics Journal 12*(2), 382–395.

Su, L. and A. Ullah (2008). Local polynomial estimation of nonparametric simultaneous equations models. *Journal of Econometrics 144*(1), 193–218.

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

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: with Application to Statistics*. New York: Springer-Verlag.

Vanhems, A. and I. Van Keilegom (2013). Semiparametric transformation model with endogeneity: a control function approach. TSE Working Paper 11-243.

White, H. and K. Chalak (2013). Identification and identification failure for treatment effects using structural systems. *Econometric Reviews 32*(3), 273–317.

Appendix

NOTATION. Let $(Z_1, Z_1, ..., Z_n)$ be an *i.i.d.* sequence of random variables taking values in a probability space $(\mathcal{Z}, \mathcal{B})$ with distribution P. For some measurable function $\phi : \mathcal{Z} \to \mathbb{R}$, define $\mathbb{E}\phi = \int \phi dP$ and $G_n\phi = \sqrt{n}(n^{-1}\sum_i \phi(Z_i) - \mathbb{E}\phi)$ for the empirical process at ϕ. Let $N(\epsilon, \mathcal{V}, \|\cdot\|)$ be the covering number with respect to the semimetric $\|\cdot\|$ and $N_{[\cdot]}(\epsilon, \mathcal{V}, \|\cdot\|)$ be the bracketing number. Let $\bar{O}_p(a_n)$ and $\bar{o}_p(a_n)$ be $O_p(a_n)$ and $o_p(a_n)$ uniformly in $y \in \mathcal{Y}$, $t \in \mathcal{T}$. Denote $f_{T|\Lambda}(t|\lambda) \equiv f_{t|\lambda}$ and $F_{Y|T\Lambda}(y|t,\Lambda) \equiv F_{y|t,\Lambda}$ for simplicity. Let C denote a generic constant.

A Preliminaries

We start with a heuristic sketch of the proof for Theorem 2 which gives an outline of the Appendix. The estimator can be decomposed to each step in the estimation procedure: $n^{-1}\sum_{i=1}^n \hat{F}_{Y|T\hat{\Lambda}}(y|t,\hat{\Lambda}_i)W_i - \mathbb{E}\big[F_{Y|T\Lambda}(y|t,\Lambda)W\big]$

$$
\begin{aligned}
&= n^{-1}\sum_{i=1}^n F_{Y|T\Lambda}(y|t,\Lambda_i)W_i - \mathbb{E}\big[F_{Y|T\Lambda}(y|t,\Lambda)W\big] \\
&+ n^{-1}\sum_{i=1}^n \Big(\hat{F}_{Y|T\Lambda}(y|t,\Lambda_i) - F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)W_i \qquad (5) \\
&+ n^{-1}\sum_{i=1}^n \Big(\hat{F}_{Y|T\hat{\Lambda}}(y|t,\Lambda_i) - \hat{F}_{Y|T\Lambda}(y|t,\Lambda_i)\Big)W_i \qquad (6) \\
&+ n^{-1}\sum_{i=1}^n \nabla_v F_{Y|T\Lambda}(y|t,v)'\big|_{v=\Lambda_i}(\hat{\Lambda}_i - \Lambda_i)W_i + \text{smaller order terms.} \qquad (7)
\end{aligned}
$$

The first line above is the estimation error of the third-step summation, which is $\sqrt{n}$-consistent. The second line (5) is the estimation error of the second-step regression with observable regressors, derived in Theorem 1 whose proof is in Section C. The third term (6) is the estimation error from generated regressors for the *regressor* role, analyzed in Section D.1.2. The term (7) is from estimating the argument in the regression function, which is extracted by the directional derivative in Section D.1.1. Section D collects the proofs related to generated regressors in Sections 4.2, 5, and 6. Section B provides primitive conditions for the smoothness assumptions and the nonparametric tuning parameters. The proofs for the inference for the treatment effects in Section 7 are in Section E.

The rest of Section A presents lemmas of stochastic equicontinuity arguments, whose proofs are in Section F. Lemma A.1 is used in Theorem 1 for the partial mean process with observable regressors in (5). We use the bracketing central limit theorem and modify Lemma B.2 in Ichimura and Lee (2010) for the complexity of the function space. [23]

Lemma A.3 is the key stochastic equicontinuity argument to deal with the *regressor* role of the generated regressor in (6), followed by Lemma A.2, Lemma A.4, and Lemma A.5. Lemma A.3 is based on Lemma 1 in MRS12. Lemma A.2 is a stochastic expansion of the second-step regression estimator using generated regressors around the estimator using the true regressors. Lemma A.2 is similar to Theorem 1 in MRS12 which characterizes how the estimation error of the generated regressor $\hat{v} - v_0$ affects the *regression estimator*. In contrast, Lemma A.2 serves as a crucial intermediate step for Theorem 2, but does not aim to extract the impact of the estimation error on the second-step

[23] Lemma 1 in Rothe (2010) shares the same stochastic equicontinuity result. We add a formal argument that the estimator $\hat{F}_{Y|T\Lambda}$ belongs to the function space in probability as sample size n goes to infinity. Ichimura and Lee (2010) also note that this result might not hold for fixed n.

regression estimation.[24] To extract the estimation error, we only implement a Taylor series expansion to linearize the stochastic expansion on the third-step partial mean, but not on the second-step regression estimation. This distinguishes our results with MRS12. We suppress the term $K_h(X_j - x)$ for notational brevity. Lemma A.4 is similar to deriving the uniform convergence rate for kernel regression, for examples, Masry (1996).

Lemma A.5 is for the generated regressors playing the *argument* role in (7) and shares the same idea with Lemma 1 in Chen, Linton, and Van Keilegom (2003). Their small order term is controlled by $o_p(n^{-1/2})$ using Donsker theorem. In contrast, we follow MRS12 to use a larger function space $\mathcal{M}_n$ in the sense that for $v \in \mathcal{M}_n$, $v/n^{\xi^*} \in \mathcal{C}_M^\alpha$ in Assumption 10 in Section B. So the bracketing CLT cannot be directly applied. When $\xi^* = 0$ ($\xi = 0$), the results coincide.

The following Result A.1 is from Lemma B.3 in Newey (1994b) and Theorem 3.2 in Haerdle, Janssen, and Serfling (1988). Define $g_{YT\Lambda}(y,t,\lambda) \equiv F_{Y|T\Lambda}(y|t,\lambda) f_{T\Lambda}(t,\lambda)$ and $\hat{g}_{YT\hat{\Lambda}}(y,t,\lambda) \equiv n^{-1}\sum_{j=1}^n \mathbf{1}_{\{Y_j \le y\}} K_h(T_j - t) K_h(\hat{\Lambda}_j - \lambda)$.

Result A.1
Suppose the bandwidth $h \to 0$ and $\log n/(nh^d) \to 0$, where the dimension of the regressors is $d = d_t + d_\lambda$. Suppose Assumptions 3 and 4 hold. For the first four results below, assume $\Delta \ge r$ and $\Delta_k \ge 0$.

1. $\sup_{(y,\lambda,t)\in\mathcal{Y}\times\mathbf{\Lambda}\times\mathcal{T}} \left|\hat{g}_{YT\Lambda}(y,t,\lambda) - g_{YT\Lambda}(y,t,\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^d}\big)^{1/2} + h^r\Big)$
2. $\sup_{(\lambda,t)\in\mathbf{\Lambda}\times\mathcal{T}} \left|\hat{f}_{T\Lambda}(t,\lambda) - f_{T\Lambda}(t,\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^d}\big)^{1/2} + h^r\Big)$
3. $\sup_{(\lambda,t)\in\mathbf{\Lambda}\times\mathcal{T}} \left|\hat{f}_{T|\Lambda}(t|\lambda) - f_{T|\Lambda}(t|\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^d}\big)^{1/2} + h^r\Big)$
4. $\sup_{(y,\lambda,t)\in\mathcal{Y}\times\mathbf{\Lambda}\times\mathcal{T}} \left|\hat{F}_{Y|T\Lambda}(y|t,\lambda) - F_{Y|T\Lambda}(y|t,\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^d}\big)^{1/2} + h^r\Big)$
5. *Now assume $\Delta \ge r+q$ and $\Delta_k \ge q$. Then* $\sup_{(y,\lambda,t)\in\mathcal{Y}\times\mathbf{\Lambda}\times\mathcal{T}} \left|\frac{\partial^q}{\partial t^q}\hat{F}_{Y|T\Lambda}(y|t,\lambda) - \frac{\partial^q}{\partial t^q}F_{Y|T\Lambda}(y|t,\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^{d+2q}}\big)^{1/2} + h^r\Big)$.

Lemma A.1 (Stochastic Equicontinuity - Regression)
Define $\mathcal{F}$ to be a class of uniformly bounded functions $f : \mathcal{Y}\times\mathcal{T}\times\mathbf{\Lambda} \mapsto \mathbb{R}$ such that (i) for each fixed $\bar{y} \in \mathcal{Y}$ and $\bar{t} \in \mathcal{T}$, the subclass $\{f(\bar{y},\bar{t},\cdot) \in \mathcal{F}\}$ is $\mathcal{M}$, where the class $\mathcal{M}$ is a class of functions such that $\log N(\epsilon,\mathcal{M},\|\cdot\|_\infty) \le C\epsilon^{-\nu}$ for some $\nu < 2$. (ii) There exists a universal constant C satisfying a Hölder continuity condition: for any $f \in \mathcal{F}$,

$$\|f(y_1,t_1,\cdot) - f(y_2,t_2,\cdot)\|_\infty \le C\|(y_1,t_1) - (y_2,t_2)\|^{1/2}. \tag{8}$$

Suppose $F_{Y|T\Lambda} \in \mathcal{F}$ and $\hat{F}_{Y|T\Lambda} \in \mathcal{F}$ w.p.a.1. Suppose the weight function W is uniformly bounded. Then

$$\sup_{y\in\mathcal{Y},t\in\mathcal{T}} \Bigg| \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(\hat{F}_{Y|T\Lambda}(y|t,\Lambda_i) - F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)W_i - \sqrt{n}\,\mathbb{E}\Big[\Big(\hat{F}_{Y|T\Lambda}(y|t,\Lambda) - F_{Y|T\Lambda}(y|t,\Lambda)\Big)W\Big]\Bigg| = o_p(1).$$

Lemma A.2 (Stochastic Equicontinuity - Generated Regressors)
Suppose Assumptions 3, 4, 6, 7, and 8 hold. For $y \in \mathcal{Y}, t, \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{V}, v_1, v_2 \in \bar{\mathcal{M}}_n$, $V_1 =$

[24] Note that in Lemma A.2, $F_{Y|TXS}(y|t,X,S) - F_{Y|TXV}(y|t,x,u)$ is not zero when the index bias is zero.

$v_1(T,S)$, $V_2 = v_2(T,S)$,

$$\begin{aligned}
&\hat{F}_{Y|TXV_1}(y|t,x,u) - \hat{F}_{Y|TXV_2}(y|t,x,u) \\
&= \mathbb{E}\Big[\frac{f_{T|XS}(t|X,S)}{f_{TXV}(t,x,u)}\Big(F_{Y|TXS}(y|t,X,S) - F_{Y|TXV}(y|t,x,u)\Big) \\
&\qquad K_h(X-x)\Big(K_h(v_1(t,S)-u) - K_h(v_2(t,S)-u)\Big)\Big] + R_n
\end{aligned}$$

where $\sup_{y\in\mathcal{Y},t,\in\mathcal{T},x\in\mathcal{X},u\in\mathcal{V},v_1,v_2\in\bar{\mathcal{M}}_n}|R_n| = O_p\big(n^{-\kappa_1} + n^{-\kappa_2} + n^{-r_2\eta}\big)$, $\kappa_2 < \min\{1-d_2\eta, 2(\delta-\eta)\}$, *and* $0 < \kappa_1 < \frac{1}{2}(1-d_2\eta) + (\delta-\eta) - \frac{1}{2}(\delta\beta+\xi)$.

Lemma A.3 (Lemma 1 in MRS12)
Suppose the conditions of Lemma A.2 hold. Then

$$\begin{aligned}
&\sup_{\substack{t\in\mathcal{T},u\in\mathcal{V},y\in\mathcal{Y}\\ v_1,v_2\in\mathcal{M}_n}}\Big|\frac{1}{n}\sum_{j=1}^n \mathbf{1}_{\{Y_j\le y\}}K_h(T_j-t)\Big(K_h(v_1(T_j,S_j)-u) - K_h(v_2(T_j,S_j)-u)\Big) \\
&\qquad - \mathbb{E}\Big[\mathbf{1}_{\{Y\le y\}}K_h(T-t)\Big(K_h(v_1(T,S)-u) - K_h(v_2(T,S)-u)\Big)\Big]\Big| = O_p(n^{-\kappa_1})
\end{aligned}$$

$$\begin{aligned}
&\sup_{\substack{t\in\mathcal{T},u\in\mathcal{V}\\ v_1,v_2\in\mathcal{M}_n}}\Big|\frac{1}{n}\sum_{j=1}^n K_h(T_j-t)\Big(K_h(v_1(T_j,S_j)-u) - K_h(v_2(T_j,S_j)-u)\Big) \\
&\qquad - \mathbb{E}\Big[K_h(T-t)\Big(K_h(v_1(T,S)-u) - K_h(v_2(T,S)-u)\Big)\Big]\Big| = O_p(n^{-\kappa_1})
\end{aligned}$$

where $0 < \kappa_1 < \frac{1}{2}(1-d_2\eta) + (\delta-\eta) - \frac{1}{2}(\delta\beta+\xi)$.

Lemma A.4
Assume the conditions of Lemma A.2 hold. Suppose a function $A(y,t,x,s;W,X,V)$ *is uniformly bounded in all its arguments and uniformly continuous in* (y,t,x,s). *When the kernel function is of r-order, assume* $A(y,t,x,s;W,X,V)$ *to be r-order continuous differentiable with respect to* (W,X,V), *with uniformly bounded derivatives. Then*

$$\begin{aligned}
&\sup_{\substack{y\in\mathcal{Y},t\in\mathcal{T}\\ v_1,v_2\in\mathcal{M}_n}}\sup_{\substack{x\in\mathcal{X}\\ s\in\mathcal{S}}}\Big|\frac{1}{n}\sum_{i=1}^n A(y,t,x,s;W_i,X_i,V_i)K_h(x-X_i)\big(K_h(v_1(t,s)-V_i) - K_h(v_2(t,s)-V_i)\big) \\
&- \mathbb{E}_{WXV}\Big[A(y,t,x,s;W,X,V)K_h(x-X)\big(K_h(v_1(t,s)-V) - K_h(v_2(t,s)-V)\big)\Big]\Big| = O_p(n^{-\kappa_{11}})
\end{aligned}$$

where $0 < \kappa_{11} < \frac{1}{2}(1-(d_2-d_t)\eta) + (\delta-\eta) - \frac{1}{2}(\delta\beta+\xi)$ *and* $\mathbb{E}_{WXV}$ *denotes the expectation with respect to the joint density of* (W,X,V).

Lemma A.5
Assume the conditions of Lemma A.2 hold. Suppose a function $A(y,t;W,V)$ *is uniformly bounded in all its arguments and uniformly continuous in* (y,t). *Then*

$$\sup_{\substack{y\in\mathcal{Y},t\in\mathcal{T}\\ v_1,v_2\in\mathcal{M}_n}}\Big|\frac{1}{n}\sum_{i=1}^n\Big(v_1(T_i,S_i) - v_2(T_i,S_i)\Big)'A(y,t;W_i,V_i) - \mathbb{E}\Big[\Big(v_1(T,S) - v_2(T,S)\Big)'A(y,t;W,V)\Big]\Big|$$

$= O_p(n^{-\kappa_{12}})$, *where* $\kappa_{12} < \frac{1}{2} + \delta - \frac{1}{2}(\delta\beta+\xi)$.

B Primitive Conditions

B.1 Observable regressors

This section shows Assumptions 5, 6, and the conditions in Theorem 1 are sufficient for the high-level conditions in Lemma A.1. By Theorem 2.7.1 of van der Vaart and Wellner (1996), there exists a constant C depending only on $M, \alpha, diam(\mathbf{\Lambda}), d$ such that $\log N(\epsilon, \mathcal{C}_M^\alpha, \|\cdot\|_\infty) \leq C\epsilon^{-d/\alpha}$ for a bounded convex $\mathbf{\Lambda}$. So choosing $\mathcal{C}_M^\alpha(\mathbf{\Lambda})$ with $\alpha > d/2$ to be the function space $\mathcal{M}$ satisfies the conditions in Lemma A.3. Assumption 5 implies $F_{Y|T\Lambda}(y|t,\Lambda) \in \mathcal{F}$. The condition of $P\big(\forall y \in \mathcal{Y}, \forall t \in \mathcal{T}, \hat{F}_{Y|T\Lambda}(y|t,\cdot) \in \mathcal{F}\big) \to 1$ hold by the following arguments:

1. The condition that $P\big(\forall y \in \mathcal{Y}, \forall t \in \mathcal{T}, \hat{F}_{Y|T\Lambda}(y|t,\cdot) \in \mathcal{M} = \mathcal{C}_M^\alpha(\mathbf{\Lambda})\big) \to 1$ is checked by the uniform convergence of the qth derivative of $\hat{F}_{Y|T\Lambda}$ for $q. \leq \underline{\alpha}$

$$\Big\|D^q \hat{F}_{Y|T\Lambda}(y|t,\cdot) - D^q F_{Y|T\Lambda}(y|t,\cdot)\Big\|_\infty = O_p\Big(\sqrt{\frac{\log n}{nh^{d_2+2q}}} + h^r\Big) = o_p(1)$$

   by Result A.1, Assumptions 5, and 6. Similar primitive conditions have been discussed in MRS12, Appendix C of EJL14 and in footnote 11 of Ichimura and Lee (2010).
2. Since the estimator $\hat{F}_{Y|T\Lambda}(y|t,\cdot)$ is smooth in t by construction, it suffices to show the Hölder continuity in y. We check the following sufficient high-level assumption modifying Assumption 3.4 in Ichimura and Lee (2010): For any $\epsilon > 0$ and $\delta > 0$, there exists n_0 such that for all $n \geq n_0$, for any $y_1, y_2 \in \mathcal{Y}$, $t_1, t_2 \in \mathcal{T}$,

$$Pr\Bigg\{\Big\|\hat{F}_{Y|T\Lambda}(y_1|t,\cdot) - \hat{F}_{Y|T\Lambda}(y_2|t,\cdot) - \big(F_{Y|T\Lambda}(y_1|t_1,\cdot) - F_{Y|T\Lambda}(y_2|t_2,\cdot)\big)\Big\|_\infty \leq \delta\big\|(y_1,t_1) - (y_2,t_2)\big\|^{1/2}\Bigg\} \geq 1-\epsilon. \tag{9}$$

   (9) and the Hölder continuity of $F_{Y|T\Lambda}$ imply $\|\hat{F}_{Y|T\Lambda}(y_1|t,\cdot) - \hat{F}_{Y|T\Lambda}(y_2|t,\cdot)\|_\infty \leq C_L|y_1-y_2|^{1/2}$, $w.p.a.1$ for $t = t_1, t_2$. (9) is satisfied by Chebyshev's inequality and the mean-square-errors of our kernel estimator for the regressor $\mathbb{E}[\mathbf{1}_{\{y_2<Y\leq y_1\}}|T=t,\Lambda]$, assuming $y_1 > y_2$. The continuity in t is implied by the kernel Assumption 4.
   Because the estimator $\hat{F}_{Y|T\Lambda}(y|t,\Lambda)$ is nonsmooth in y, the function space $\mathcal{F}$ is allowed to be less smooth in y by assuming a Hölder continuity (8). Alternatively, as discussed in Ichimura and Lee (2010), a smoothed cdf estimator is needed if a stronger Lipschitz continuity assumption is made on $\mathcal{F}$.

B.2 Generated regressors

When the generated regressors are specified and estimated parametrically, $\delta = 1/2$ and Complexity Assumption 8 is satisfied by Example 19.7 in van der Vaart (2000) for a Donsker parametric function. The following primitive conditions are sufficient for Assumption 8.

Assumption 10 (Complexity)
For any $j = 1, ..., d_v$,

(i) *Let $\mathcal{M}_n$ be the set of functions defined on some compact and convex sets $\mathcal{S} \subset \mathbb{R}^{d_s}$. For any $v \in \mathcal{M}_n$, $v/n^{\xi^*} \in \mathcal{C}_M^\alpha(\mathcal{S})$, for some $\xi^* \geq 0$, $\alpha > d_s/2$, and $M > 0$.*
(ii) $v_{0j} \in \mathcal{C}_M^\alpha(\mathcal{S})$
(iii) $\|D^\alpha \hat{v}_j - D^\alpha v_{0j}\|_\infty = o_p(n^{\xi^*})$

Assumption 10 (i) assumes $\mathcal{M}_n$ to be the set of functions whose partial derivatives up to order α exists and are uniformly bounded by some multiple of n^{ξ^*}. By Corollary 2.7.2 in van der Vaart and Wellner

(1996), Assumption 10 (i) implies Assumption 8 (ii) by letting $\beta \equiv d_s/\alpha$ and $\xi \equiv \xi^* d_s/\alpha = \xi_j^* \beta$. Then the complexity of the function space is controlled by the uniform bound ξ^* and the differentiability α. Assumptions 10 (ii) and (iii) are sufficient for Assumption 8 (i). The advantage of the uniform bound ξ^* is that by Assumptions 10 (iii), when a stronger smoothness assumption is required (a larger α), a larger ξ^* can be a leverage to avoid restrictive assumption on the first-step estimation of the generated regressor $\hat{v}$. If the function space is more restrictive or less complex (smaller ξ^* or larger α), the first-step of estimation needs to be more accurate to ensure $\hat{v}$ belongs to $\mathcal{M}_n$ *w.p.a.*1. This is the additional cost of assuming a smoother function space.

The following assumption for the second-step bandwidth is sufficient for the conditions in Theorem 2 and makes the remainder terms of smaller order, i.e., $\sqrt{nh_2^{d_t}} R_n = o_p(1)$. The stochastic equicontinuity argument contributes a term of order $n^{-\kappa_1}$ to the remainder term R_n. The smaller-order terms from linearizing the estimator are governed by $O_p(n^{-\kappa_2})$.

Assumption 11 (2nd-step Bandwidth)
The bandwidth for the second-step regression $h_2 \sim n^{-\eta}$ satisfies

$$\frac{1}{2r_2 + d_t} < \eta < \min\Big\{\frac{1}{2d_2 + d_t}, \frac{\delta(2-\beta) - \xi}{d_2 + 2 - d_t}, \frac{\delta}{2}\Big\}. \tag{10}$$

When $d_t \leq 4$, assume $\delta > 1/4$. When $d_t < 4$, assume $\eta < \frac{4\delta - 1}{4 - d_t}$. When $d_t > 4$ and $\delta < 1/4$, assume $\eta > \frac{1 - 4\delta}{d_t - 4}$.

The smaller-order term from linearizing the estimator $\|\hat{V} - V\|_\infty^2/h^2 = o_p\big((nh^{d_t})^{-1/2}\big)$ implies $\delta > \frac{1}{4}(1 + \eta(4 - d_t))$. When the dimension of continuous treatments is larger than four, δ is allowed to be smaller than $1/4$. The convergence rate of the second-stage regression is also allowed to be slower than $n^{-1/4}$.

For the cases of full mean, discrete treatment, and semiparametric estimation, the conditions are modified by setting $d_t = 0$. Assumption 11 then reflects that the regularity conditions are more restrictive in the semiparametric models. Both the generated regressor and second-step regression are estimated at a rate faster than $n^{-1/4}$. It is a common condition in semiparametric models with infinite dimensional nuisance parameters, for example, Newey (1994a), Chen, Linton, and Van Keilegom (2003), EJL14, and Mammen, Rothe, and Schienle (2013).

The following primitive bandwidth condition is for the nonparametrically estimated GPS in Section 5.2, where $d_2 = 2$, $d_1 = d_x + d_t$, and $d_s = d_x$.

Assumption 12 (Bandwidth - GPS)

(i) *The bandwidth for the second step regression $h_2 \sim n^{-\eta}$ and the bandwidth for the first step GPS estimation $h_1 \sim n^{-g}$ satisfy*

$$\frac{1}{2r_1 + d_t} < g < \min\Big\{\frac{1}{2d_1}, \frac{2\xi + \beta}{\beta d_1 + 2d_x}\Big\}$$
$$\frac{1}{2r_2 + d_t} < \eta < \min\Big\{\frac{1}{4}\big(1 - d_1 g\big), \frac{1 - 2d_1 g}{4 - d_t}, \frac{(1 - \beta/2)(1 - d_1 g) - \xi}{4 - d_t}, \frac{1}{4 + d_t}\Big\}$$

(ii) *The second step regression and the first step GPS estimation use the same kernel function with order r and bandwidth $h_1 = h_2 \sim n^{-\eta}$ satisfies*

$$\frac{1}{2r + 1} < \eta < \min\Big\{\frac{1}{2d_1 + 4 - d_t}, \frac{1 - \xi - \beta/2}{d_1(1 - \beta/2) + 4 - d_t}, \frac{\xi + \beta/2}{d_1\beta/2 + d_x}\Big\}.$$

C Proof of Theorem 1 (Observable regressors)

The decomposition and linearization follows the proof of Theorem 1 in Rothe (2010).

$$\begin{aligned}\sqrt{nh^{d_t}}(\hat{F}_{Y(t)}(y;\Lambda,W)-F_{Y(t)}(y;\Lambda,W)) &= \sqrt{nh^{d_t}}\big(\frac{1}{n}\sum_{i=1}^{n}\hat{F}_{Y|T\Lambda}(y|t,\Lambda_i)W_i - \mathbb{E}\big[F_{Y|T\Lambda}(y|t,\Lambda)W\big]\big)\\ &= \sqrt{h^{d_t}}\ G_n\big[\hat{F}_{Y|T\Lambda}(y|t,\Lambda_i)W_i - F_{Y|T\Lambda}(y|t,\Lambda_i)W_i\big] \quad (11)\\ &+ \sqrt{h^{d_t}}\ G_n\big[F_{Y|T\Lambda}(y|t,\Lambda_i)W_i\big] \quad (12)\\ &+ \sqrt{nh^{d_t}}\ \mathbb{E}\big[\hat{F}_{Y|T\Lambda}(y|t,\Lambda)W - F_{Y|T\Lambda}(y|t,\Lambda)W\big]. \quad (13)\end{aligned}$$

The first term (11) is $\bar{o}_p(\sqrt{h^{d_t}})$ by Lemma A.1. The second term (12) is $\bar{O}_p(\sqrt{h^{d_t}}) = \bar{o}_p(1)$, by the Donsker property of $\{(\Lambda,W)\mapsto F_{Y|T\Lambda}(y|t,\Lambda)W : y\in\mathcal{Y}, t\in\mathcal{T}\}$. The asymptotic distribution is dominated by the third term (13). The influence function is not standard Donsker and contains kernels. The weight W is projected on the regressors $\mathbb{E}[W|\Lambda=\lambda]\equiv W_\Lambda(\lambda)$. In (13), $\mathbb{E}\big[\hat{F}_{Y|T\Lambda}(y|t,\Lambda)W - F_{Y|T\Lambda}(y|t,\Lambda)W\big]$

$$\begin{aligned}&= \int \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\lambda)\big)\ K_h(t-T_i)K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{\hat{f}_{T\Lambda}(t,\lambda)}dF_\Lambda(\lambda) \quad (14)\\ &= \int \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\lambda)\big)\ K_h(t-T_i)K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{f_{T\Lambda}(t,\lambda)}dF_\Lambda(\lambda) \quad (15)\\ &- \int \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\lambda)\big)\big(\hat{f}_{t\lambda}-f_{t\lambda}\big)K_h(t-T_i)K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{f_{t\lambda}^2}dF_\Lambda(\lambda) \quad (16)\\ &+ \bar{O}_p(\|\hat{f}_{T\Lambda}-f_{T\Lambda}\|_\infty^2).\end{aligned}$$

Because $|\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\lambda)|\le 1$ and integration takes over a compact set, the last term is made $o_p((nh^{d_t})^{-1/2})$ by Assumption 6 (iii). We show (15) contributes the main influence function $\psi_{tin}(y)$ and (16) is of smaller order by the U-process theory.

Define $D_i(\lambda) = \big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\lambda)\big)W_\Lambda(\lambda)f_\Lambda(\lambda)/f_{T\Lambda}(t,\lambda)$. By the standard algebra for kernel, i.e., change of variables, Taylor expansion, and the dominated convergence theorem,

$$\int K_h(\lambda-\Lambda_i)D_i(\lambda)d\lambda = \int K(v)\ D_i(\Lambda_i+vh)dv = D_i(\Lambda_i) + O_p(h^r). \quad (17)$$

Then (15) becomes $n^{-1}\sum_{i=1}^n \psi_{tin}(y;\Lambda,W)h^{-d_t/2} + o_p((nh^{d_t})^{-1/2})$ by Assumption 6 (ii). The bias is dominated by the bias of the influence function. Consider $d_t=1$ for simplicity,

$$\begin{aligned}&h^{-1/2}\mathbb{E}\psi_{tin}(y) = \mathbb{E}\big[\big(F_{Y|T\Lambda}(y|T,\Lambda) - F_{Y|T\Lambda}(y|t,\Lambda)\big)K_h\big(T-t\big)W_\Lambda(\Lambda)/f_{T|\Lambda}(t|\Lambda)\big]\\ &= \mathbb{E}\Big[\Big(\frac{\partial^r}{\partial t^r}\Big(F_{Y|T\Lambda}(y|t,\Lambda)f_{T|\Lambda}(t|\Lambda)\Big) - F_{Y|T\Lambda}(y|t,\Lambda)\frac{\partial^r}{\partial t^r}f_{T|\Lambda}(t|\Lambda)\Big)\frac{h^r}{r!}\int u^rK(u)du\frac{W_\Lambda(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\Big]\\ &= Ch^r\mathbb{E}\Big[\frac{\partial^r}{\partial t^r}F_{Y|T\Lambda}(y|t,\Lambda)\cdot W_\Lambda(\Lambda)\Big]. \quad (18)\end{aligned}$$

We now show (16) is $\bar{o}_p((nh^{d_t})^{-1/2})$. Define the convolution kernel

$$\begin{aligned}\bar{K}_h(\Lambda_i-\Lambda_j) &\equiv \frac{1}{h^{d_\lambda}}\bar{K}\Big(\frac{\Lambda_i-\Lambda_j}{h}\Big) = \frac{1}{h^{d_\lambda}}\int K(v)K\Big(v-\frac{\Lambda_i-\Lambda_j}{h}\Big)dv\\ &= \frac{1}{h^{2d_\lambda}}\int K\Big(\frac{\Lambda_i-\lambda}{h}\Big)K\Big(\frac{\Lambda_j-\lambda}{h}\Big)d\lambda. \quad (19)\end{aligned}$$

When $s=t$, $\bar{K}(0)=\int K^2(u)du$. Define here $D_i(\lambda)=\big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|T\Lambda}(y|t,\lambda)\big)W_\Lambda(\lambda)f_\Lambda(\lambda)/f^2_{T\Lambda}(t,\lambda)$. We first calculate

$$\int K_h(\lambda-\Lambda_i)K_h(\lambda-\Lambda_j)D_i(\lambda)d\lambda=\int K(v)\ K_h(\Lambda_i-\Lambda_j+hv)\ D_i(\Lambda_i+hv)dv$$
$$=D_i(\Lambda_i)\bar{K}_h(\Lambda_i-\Lambda_j)+\int D_i'(\Lambda_i)vK(v)K\Big(v+\frac{\Lambda_i-\Lambda_j}{h}\Big)dv+\text{small order terms}$$

The second term is $o_p(h)$ for $i\neq j$ and is $D_i'(\Lambda_i)\int vK^2(v)dv$ for $i=j$ which contributes a smaller order term in (16): $n^{-2}\sum_{i=1}^n K_h^2(t-T_i)D_i'(\Lambda_i)\int vK^2(v)dv=\bar{o}_p((nh^{d_t})^{-1/2})$. Then

$$(16)=\frac{1}{n}\sum_{i=1}^n K_h(T_i-t)$$
$$\times\int\Big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|T\Lambda}(y|t,\lambda)\Big)\frac{W_\Lambda(\lambda)}{f^2_{T\Lambda}(t,\lambda)}K_h(\Lambda_i-\lambda)\Big(\frac{1}{n}\sum_{j=1}^n K_h(\Lambda_j-\lambda)K_h(T_j-t)-f_{T\Lambda}(t,\lambda)\Big)dF_\Lambda(\lambda)$$
$$=\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n D_i(\Lambda_i)K_h(T_i-t)\big(\bar{K}_h(\Lambda_i-\Lambda_j)K_h(T_j-t)-f_{T\Lambda}(t,\Lambda_i)\big)+\bar{o}_p\Big(\frac{1}{\sqrt{nh^{d_t}}}\Big).$$

For the projection of the U-process, define $\bar{f}_{T\Lambda}(t,\Lambda_i)\equiv\mathbb{E}\big[\bar{K}_h(\Lambda_i-\Lambda_j)\cdot K_h(T_j-t)\big|Z_i\big]$

$$=\int\int\frac{1}{h^{d_t}}K\Big(\frac{T-t}{h}\Big)\cdot\frac{1}{h^{2d_\lambda}}\int K\Big(\frac{\Lambda_i-v}{h}\Big)K\Big(\frac{\Lambda-v}{h}\Big)dv\cdot f_{T\Lambda}(T,\Lambda)d\Lambda dT$$
$$=\int\frac{1}{h^{d_\lambda}}K\Big(\frac{\Lambda_i-v}{h}\Big)\int\int K(r)K(s)\ f_{T\Lambda}(t+rh,v+sh)drds\ dv$$
$$=\int K(w)f_{T\Lambda}(t,\Lambda_i+wh)dw+O_p(h^r)=f_{T\Lambda}(t,\Lambda_i)+O_p(h^r).$$

Define $H(Z_i,Z_j;y,t,h)\equiv$

$$D_i(\Lambda_i)K_h(T_i-t)\cdot\big(\bar{K}_h(\Lambda_i-\Lambda_j)\cdot K_h(T_j-t)-\bar{f}_{T\Lambda}(t,\Lambda_i)\big).$$

Then (16) becomes

$$\frac{1}{n^2}\sum_i\sum_{j\neq i}H(Z_i,Z_j;y,t,h)+\frac{1}{n^2}\sum_i H(Z_i,Z_i;y,t,h)+\bar{o}_p\Big(\frac{1}{\sqrt{nh^{d_t}}}\Big).\tag{20}$$

The second term in (20) is $\bar{o}_p(1/\sqrt{nh^{d_t}})$, because its second part is $n^{-1}(15)$ and its first part is smaller than $n^{-2}\sum_i u(\Lambda_i)K_h^2(T_i-t)h^{-d_\lambda}\int K^2(v)dv=o_p((nh^{d_t})^{-1/2})$. The first term in (20) is a degenerate second order U-process. By applying Corollary 4 (ii) in Sherman (1994),

$$\sup_{y\in\mathcal{Y},t\in\mathcal{T},h>0}\Big|\frac{1}{n^2}\sum_i\sum_{j\neq i}h^{2d_t+d_\lambda}H(Z_i,Z_j;y,t,h)\Big|=O_p\Big(\frac{1}{n}\Big).$$

Therefore, by Assumption 6,

$$\sup_{y\in\mathcal{Y},t\in\mathcal{T}}\Big|\frac{1}{n^2}\sum_i\sum_{j\neq i}H(Z_i,Z_j;y,t,h)\Big|=O_p\Big(\frac{1}{nh^{2d_t+d_\lambda}}\Big)=o_p\Big(\frac{1}{\sqrt{nh^{d_t}}}\Big).$$

Applying Corollary 4 in Sherman (1994)
The class of P-degenerate functions of order two $\mathcal{H}\equiv\{h^{2d_t+d_\lambda}H(Z_i,Z_j;y,t,h):y\in\mathcal{Y},t\in\mathcal{T},h>0\}$ has an envelope $F(\Lambda_i,\Lambda_j)=u(\Lambda_i)$. Let $\mathcal{H}$ be a real-valued functions on $\mathcal{S}^2=\mathcal{S}\otimes\mathcal{S}$. And $P^2=P\otimes P$

denotes the product measure. We then show $\mathcal{H}$ is Euclidean for this envelope F satisfying $\mathbb{E}F^2 = P^2F^2 = \mathbb{E}[u^2(\Lambda)] < \infty$.

$\{\mathbf{1}_{\{Y_i \le y\}} : y \in \mathcal{Y}\}$ and $\{F_{Y|T\Lambda}(y|t,\Lambda_i) : y \in \mathcal{Y}, , t \in \mathcal{T}\}$ are manageable by the fact that they are monotone increasing in y (p.221 in Kosorok (2008)) and assumed to belong to C_M^α. By Example 2.10 and Lemma 2.14 in Pakes and Pollard (1989), $\{K((T-t)/h) : h > 0, t \in \mathcal{T}\}$ and hence $\mathcal{H}$ are Euclidean.

Lemma C.1 (Functional Central Limit Theorem)
The process $n^{-1/2}\sum_{i=1}^n \psi_{tin}(\cdot)$ weakly converges to a Gaussian process $\mathbb{G}_t(\cdot)$ defined in Theorem 1.

Proof of Lemma C.1
For all $\omega \in \Omega$, the triangular array $f_{ni}(\omega, y) \equiv n^{-1/2}\psi_{tin}(y) = \big(\mathbf{1}_{\{Y_i(\omega)\le y\}} - F_{Y|T\Lambda}(y|t,\Lambda_i(\omega))\big) \cdot (nh^{d_t})^{-1/2} K_h\big(T_i(\omega) - t\big)\frac{W_\Lambda(\Lambda_i(\omega))}{f_{T|\Lambda}(t|\Lambda_i(\omega))}$ are independent within rows. Define the $n \times 1$ vector $f_n(\omega, y) \equiv \big(f_{n1}(\omega, y), ..., f_{nn}(\omega, y)\big)'$ and the random set $\mathcal{F}_{n\omega} \equiv \big\{f_n(\omega, y) : y \in \mathcal{Y}\big\}$. We skip the subscript t for notational ease without loss of clarity. We check the conditions for the functional CLT, Theorem 10.6 in Pollard (1990).

(i) The triangular array processes $\{f_{ni}(\omega, y)\}$ are manageable with respect to the envelopes $F_{ni}(\omega) \equiv (nh^{d_t})^{-1/2} K_h\big(T_i(\omega) - t\big)\frac{W_\Lambda(\Lambda(\omega)}{f_{T|\Lambda}(t|\Lambda_i(\omega))}$. First, $\{\mathbf{1}_{\{Y_i\le y\}} : y \in \mathcal{Y}, i = 1, ..., n\}$ and $\{F_{Y|T\Lambda}(y|t,\Lambda_i) : y \in \mathcal{Y}, i = 1, ..., n\}$ are manageable by the fact that they are monotone increasing in y (p.221 in Kosorok (2008). And $F_n(\omega) \equiv (F_{n1}, ..., F_{nn})^\top$ is a $\mathbb{R}^n$-valued function on the underlying probability space. Then (i) is proved by applying Lemma E1 in Andrews and Shi (2013).

Before we proceed to check the next conditions, it will be convenient to calculate the following expectations. Define $V(y, T, \Lambda) \equiv \big(F_{Y|T\Lambda}(y|T,\Lambda) - F_{Y|T\Lambda}(y|t,\Lambda)\big) \cdot f_{T|\Lambda}(T|\Lambda)$. By assumption, $\frac{\partial^r}{\partial T^r}V(y, T, \Lambda)$ is bounded uniformly over y, T, Λ, and $f_{T|\Lambda}$ is uniformly bounded away from zero. Then $\mathbb{E}f_{ni}(y)$

$$= \sqrt{\frac{h^{d_t}}{n}}\mathbb{E}\left[\int K(u)\Big(F_{Y|T\Lambda}(y|t+uh,\Lambda) - F_{Y|T\Lambda}(y|t,\Lambda)\Big)\cdot f_{T|\Lambda}(t+uh|\Lambda)du\ \frac{W_\Lambda(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\right]$$
$$= \sqrt{\frac{h^{d_t}}{n}}\frac{h^r}{r!}\int K(u)u^r du \cdot \mathbb{E}\left[\frac{\partial^r}{\partial T^r}V(y, T, \Lambda)\Big|_{T=t} \cdot \frac{W_\Lambda(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\right] = \bar{O}\Big(h^r\sqrt{\frac{h^{d_t}}{n}}\Big).$$

For any $t, s \in \mathcal{T}$, $\mathbb{E}\big[f_{ni}(y_1, t)f_{ni}(y_2, s)\big]$

$$= \frac{1}{n}\mathbb{E}\left[\Big(F_{y_1|T\Lambda} - F_{y_1|t\Lambda}F_{y_2|T\Lambda} - F_{y_1|T\Lambda}F_{y_2|s\Lambda} + F_{y_1|t\Lambda}F_{y_2|s\Lambda}\Big)\frac{1}{h^{d_t}}K\Big(\frac{T-t}{h}\Big)K\Big(\frac{T-s}{h}\Big)\frac{W_\Lambda^2(\Lambda)}{f_{t|\Lambda}f_{s|\Lambda}}\right]$$
$$= \frac{1}{n}\mathbb{E}\Bigg[\int \Big(F_{Y|T\Lambda}(y_1|t+uh,\Lambda) - F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|t+uh,\Lambda) - F_{Y|T\Lambda}(y_1|t+uh,\Lambda)F_{Y|T\Lambda}(y_2|s,\Lambda)$$
$$+ F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|s,\Lambda)\Big)K(u)K\big(u + \frac{t-s}{h}\big)f_{T\Lambda}(t+uh,\Lambda)du\ \frac{W_\Lambda^2(\Lambda)}{f_{t|\Lambda}f_{s|\Lambda}}\Bigg]$$
$$= \frac{1}{n}\mathbb{E}\left[\Big(F_{Y|T\Lambda}(y_1|t,\Lambda) - F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|t,\Lambda)\Big)\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}(s|\Lambda)} + O(h)\right]\bar{K}\big(\frac{s-t}{h}\big) \tag{21}$$

uniformly in $y_1 \le y_2 \in \mathcal{Y}$. The convolution kernel $\bar{K}$ is defined in (19).

(ii) Define $\mathcal{Z}_n(y) = \sum_{i=1}^n \big(f_{ni}(y) - \mathbb{E}f_{ni}(y)\big)$. Let $y_1 \le y_2 \in \mathcal{Y}$. Using (21), the covariance kernel of the limiting Gaussian process is

$$\lim_{n\to\infty}\mathbb{E}\mathcal{Z}_n(y_1)\mathcal{Z}_n(y_2) = \lim_{n\to\infty}\mathbb{E}\big[\psi_{tin}(y_1)\psi_{tin}(y_2)\big] - \mathbb{E}\big[\psi_{tin}(y_1)\big]\mathbb{E}\big[\psi_{tin}(y_2)\big]$$
$$= \mathbb{E}\left[\Big(F_{Y|T\Lambda}(y_1|t,\Lambda) - F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|t,\Lambda)\Big)\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\right]\int K^2(v)dv. \tag{22}$$

(iii) Using (21),

$$\sum_{i=1}^{n}\mathbb{E}F_{ni}^2=\mathbb{E}\Big[\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}+O(h)\Big]\int K^2(v)dv.$$

(iv) For any $\epsilon>0$, $\sum_{i=1}^n\mathbb{E}F_{ni}^2\ \mathbf{1}\big\{F_{ni}>\epsilon\big\}\to 0$ holds. This is because $\mathbf{1}\big\{F_{ni}>\epsilon\big\}=0$ for n large enough, by assuming K is bounded, $f_{T|\Lambda}$ is bounded away from zero, and $\sqrt{nh^{d_t}}\to\infty$.

(v) Uniform in y_1,y_2, $n\mathbb{E}\big|f_{ni}(y_1)-f_{ni}(y_2)\big|^2\longrightarrow$

$$\rho(y_1,y_2)^2\equiv\int K^2(u)du\mathbb{E}\left[\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\Big(F_{y_2|t,\Lambda}-F_{y_1|t,\Lambda}-\big(F_{y_2|t,\Lambda}-F_{y_1|t,\Lambda}\big)^2\Big)\right].$$

Therefore, uniformly in y_1,y_2, $\rho_n(y_1,y_2)\equiv\big(\sum_{i=1}^n\mathbb{E}\big|f_{ni}(y_1)-f_{ni}(y_2)\big|^2\big)^{1/2}\to\rho(y_1,y_2)$.

D Generated Regressors

Section D.1.1 presents the proof of Theorem 2. Section D.2 collects the proofs for Corollaries for the examples in Section 5. Section D.3 collects the proofs for Corollaries for the semiparametric models in Section 6.

D.1 Proof of Theorem 2

D.1.1 Functional directional derivative

Let $\Lambda=V=v_0(S)$ for simplicity without lost of clarity and its estimator $\hat{V}_i\equiv\hat{v}(S_i)$. The first part of the proof uses directional derivative to decompose the variation from estimating the generated regressor for its two roles, the argument and the regressor, following the proof of Corollary 6 in MRS12.

The true functions $\bar{f}=(\bar{f}_1,\bar{f}_2)=\big(\mathbb{E}\big[\mathbf{1}_{\{Y\le y\}}\big|T=t,V=v\big],v_0(S)\big)$. Denote $f_2=f_2(S)$ and $f_1=f_1(t,v)$. Define the functional $S_n(f)\equiv\frac{1}{n}\sum_{i=1}^n f_1\big(t,f_2(S_i)\big)W_i-\theta_0$. For any two functions of S, $f(S)$ and $g(S)$, denote $[f+g](S)\equiv f(S)+g(S)$ and $f^{(v)}\equiv\nabla_v f(t,v)$. The directional derivative

$$\begin{aligned}
\dot{S}_n(\bar{f})[f-\bar{f}]&=\lim_{s\to0}\frac{1}{s}\Big(S_n\big(\bar{f}+s(f-\bar{f})\big)-S_n(\bar{f})\Big)\\
&=\lim_{s\to0}\frac{1}{s}\frac{1}{n}\sum_{i=1}^n\Big\{\Big[\bar{f}_1+s(f_1-\bar{f}_1)\Big]\Big(t,\big[\bar{f}_2+s(f_2-\bar{f}_2)\big](S_i)\Big)W_i-\bar{f}_1\big(t,\bar{f}_2(S_i)\big)W_i\\
&\quad-\bar{f}_1\Big(t,\big[\bar{f}_2+s(f_2-\bar{f}_2)\big](S_i)\Big)W_i+\bar{f}_1\Big(t,\big[\bar{f}_2+s(f_2-\bar{f}_2)\big](S_i)\Big)W_i\Big\}\\
&=\frac{1}{n}\sum_{i=1}^n\big[f_1-\bar{f}_1\big]\big(t,\bar{f}_2(S_i)\big)W_i+\frac{1}{n}\sum_{i=1}^n\bar{f}_1^{(v)}\big(t,\bar{f}_2(S_i)\big)\cdot\big[f_2-\bar{f}_2\big](S_i)W_i.
\end{aligned}$$

Define

$$\begin{aligned}
T_{1,n}(\hat{f})&\equiv\frac{1}{n}\sum_{i=1}^n\big(\hat{F}_{Y|T\hat{V}}(y|t,v_0(S_i))-F_{Y|TV}(y|t,v_0(S_i))\big)W_i\\
T_{2,n}(\hat{f})&\equiv\frac{1}{n}\sum_{i=1}^n\nabla_vF_{Y|TV}(y|t,v_0(S_i))'\big(\hat{v}(S_i)-v_0(S_i)\big)W_i.
\end{aligned}$$

$T_{1,n}$ is the estimation error from the second-step nonparametric regression in (5) and the first-step generated regressor in (6). $T_{2,n}$ contributes (7). By Lemma A.5 with $A(y,t;W,V)=\nabla_vF_{Y|TV}(y|t,v)\big|_{v=V}$.

W,

$$T_{2,n} = \mathbb{E}\big[\nabla_v F_{Y|TV}(y|t, v_0(S))'\big(\hat{v}(S) - v_0(S)\big) \cdot W\big] + O_p\big(n^{-\kappa_{12}}\big) \tag{23}$$
$$= \mathbb{E}\big[\nabla_v F_{Y|TV}(y|t, v_0(S))'\big(\hat{v}(S) - v_0(S)\big) \cdot \mathbb{E}[W|S]\big] + O_p\big(n^{-\kappa_{12}}\big) \tag{24}$$

The second expression (24) is useful when replacing the estimation error $\hat{v} - v_0$ by its linear representation; see Proof of Corollary 2 for example.

The smaller order terms

$$S_n(f) - S_n(\bar{f}) - \dot{S}_n(\bar{f})[f - \bar{f}] = \frac{1}{n}\sum_{i=1}^n \hat{F}_{Y|T\hat{V}}(y|t, \hat{V}_i) - \frac{1}{n}\sum_{i=1}^n F_{Y|TV}(y|t, V_i) - T_{1,n}(f) - T_{2,n}(f)$$

$$= \frac{1}{n}\sum_{i=1}^n \big[f_{1,A}^{(v)} - \bar{f}_1^{(v)}\big]\big(t, \bar{f}_2(S_i)\big) \cdot \big(f_2(S_i) - \bar{f}_2(S_i)\big)W_i + O_p(\|f_{1,B}\|_\infty) + O_p\Big(||f_2 - \bar{f}_2||_\infty^2\Big) \equiv so1$$

for any $f_1 = f_{1,A} + f_{1,B}$. By Lemma A.2,

$$\hat{f}_1 - \bar{f}_1 = \big[\hat{F}_{Y|T\hat{V}} - \hat{F}_{Y|TV} + \hat{F}_{Y|TV} - F_{Y|TV}\big](y|t, v)$$
$$= \frac{1}{f_{TV}(t, v)}\mathbb{E}\Big[f_{T|S}(t|S)\big(F_{Y|TS}(y|t, S) - F_{Y|TV}(y|t, v)\big)\big(K_h(\hat{v}(S) - v) - K_h(v_0(S) - v)\big)\Big]$$
$$+ \hat{F}_{Y|TV}(y|t, v) - F_{Y|TV}(y|t, v) + O_p(R_n).$$

Let $f_{1,B} = O_p(R_n)$ and the rest leading terms be $f_{1,A} - \bar{f}_1$. Together with Result A.1 and a similar calculation in (32),

$$\big\|f_{1,A}^{(v)} - \bar{f}_1^{(v)}\big\|_\infty = O(h^{-2})\big\|\hat{V} - V\big\|_\infty + O_p\Big(\sqrt{\log n/(nh^{d_2+2})} + h^r\Big).$$

So by the bandwidth assumption,

$$\big|so1\big| \le O_p\Big(\big\|f_{1,A}^{(v)} - \bar{f}_1^{(v)}\big\|_\infty \cdot \big\|\hat{V} - V\big\|_\infty\Big) + O_p(R_n) + O_p\Big(\big\|\hat{V} - V\big\|_\infty^2\Big)$$
$$= O_p\Big(\frac{1}{h^2}\|\hat{V} - V\|_\infty^2 + \sqrt{\frac{\log n}{nh^{d_2+2}}}\|\hat{V} - V\|_\infty + R_n\Big) = O_p\Big(R_n\Big) = o_p((nh^{d_t})^{-1/2}).$$

D.1.2 The regressor role of the generated regressor

The first part of $T_{1,n}(\hat{f})$, $\frac{1}{n}\sum_{i=1}^n W_i\Big(\hat{F}_{Y|TX\hat{V}}(y|t, X_i, V_i) - \hat{F}_{Y|TXV}(y|t, X_i, V_i)\Big)$ comes from the generated regressor used as a regressor in the estimator, and the argument is evaluated at the true $V \in \mathcal{V}$.

By Lemma A.2 and Lemma A.4, we claim

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} W_i\Big(\hat{F}_{Y|TXV_1}(y|t,X_i,V_i) - \hat{F}_{Y|TXV_2}(y|t,X_i,V_i)\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{W_i}{f_{TXV}(t,X_i,V_i)}\mathbb{E}_{XS}\Big[f_{T|XS}(t|X,S)\Big(F_{Y|TXS}(y|t,X,S) - F_{Y|TXV}(y|t,X_i,V_i)\Big) \\
&\qquad\qquad K_h(X-X_i)\Big(K_h(v_1(t,S)-V_i) - K_h(v_2(t,S)-V_i)\Big)\Big] + \bar{O}_p\big(R_n\big) \\
&\equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{XS}\Big[A(y,t,X,S;W_i,X_i,V_i)K_h(X-X_i)\Big(K_h(v_1(t,S)-V_i) - K_h(v_2(t,S)-V_i)\Big)\Big] + \bar{O}_p\big(R_n\big) \\
&= \mathbb{E}_{W_iX_iV_i}\Big[\mathbb{E}_{XS}\Big[A(y,t,X,S;W_i,X_i,V_i)K_h(X-X_i)\Big(K_h(v_1(t,S)-V_i) - K_h(v_2(t,S)-V_i)\Big)\Big]\Big] \\
&\quad + \bar{O}_p\big(R_n\big) \qquad (25)
\end{aligned}
$$

where $\mathbb{E}_{W_iX_iV_i}$ denotes the expectation by the joint density of the random variables (W,X,V) and the subscript i is to distinguish from the random variables (X,S) in the second stage regression. Define

$$
A(y,t,x,s;W_i,X_i,V_i) \equiv \frac{W_i f_{T|XS}(t|x,s)}{f_{TXV}(t,X_i,V_i)}\Big(F_{Y|TXS}(y|t,x,s) - F_{Y|TXV}(y|t,X_i,V_i)\Big).
$$

The stochastic equicontinuity (25) is implied by Lemma A.4 and the fact that the sup-norm implies the L_1-norm:

$$
\begin{aligned}
&\sup_{\substack{y\in\mathcal{Y},t\in\mathcal{T}\\ v_1,v_2\in\bar{\mathcal{M}}_n}} \mathbb{E}_{XS}\Big[\Big|\frac{1}{n}\sum_{i=1}^{n} A(y,t,X,S;W_i,X_i,V_i)K_h(X-X_i)\Big(K_h(v_1(t,S)-V_i) - K_h(v_2(t,S)-V_i)\Big) \\
&\qquad - \mathbb{E}_{W_iX_iV_i}\Big[A(y,t,X,S;W_i,X_i,V_i)K_h(X-X_i)\Big(K_h(v_1(t,S)-V_i) - K_h(v_2(t,S)-V_i)\Big)\Big]\Big|\Big] \\
&\qquad = O_p\big(n^{-\kappa_1}\big).
\end{aligned}
$$

The stochastic equicontinuity argument replaces the summation with the expectation over the arguments (X_i,V_i) and weight W_i in the third step. We calculate the outer expectation for (W_i,X_i,V_i) first, instead of the inner expectation for the regressor (X,S). The interchangeability of integrations is by Fubini's theorem with the existence of expectations. This is the key step of the proof. Denote

$W_{XV}(x,u) \equiv \mathbb{E}[W|X=x,V=u]$. Then in (25),

$$
\begin{aligned}
&\mathbb{E}_{W_iX_iV_i}\Big[\mathbb{E}_{XS}\Big[A(y,t,X,S;W_i,X_i,V_i)K_h(X-X_i)\Big(K_h(v_1(t,S)-V_i)-K_h(v_2(t,S)-V_i)\Big)\Big]\Big] \\
&= \mathbb{E}_{W_iX_iV_i}\Big[\frac{W_i}{f_{TXV}(t,X_i,V_i)}\mathbb{E}_{XS}\Big[f_{T|XS}(t|X,S)\Big(F_{Y|TXS}(y|t,X,S)-F_{Y|TXV}(y|t,X_i,V_i)\Big) \\
&\quad K_h(X-X_i)\Big(K_h(v_1(t,S)-V_i)-K_h(v_2(t,S)-V_i)\Big)\Big]\Big] \\
&= \mathbb{E}_{XS}\Big[\mathbb{E}_{X_iV_i}\Big[\frac{W_{XV}(X_i,V_i)}{f_{TXV}(t,X_i,V_i)}\cdot f_{T|XS}(t|X,S)\Big(F_{Y|TXS}(y|t,X,S)-F_{Y|TXV}(y|t,X_i,V_i)\Big) \\
&\quad K_h(X-X_i)\Big(K_h(v_1(t,S)-V_i)-K_h(v_2(t,S)-V_i)\Big)\Big]\Big] \\
&= \mathbb{E}_{XS}\mathbb{E}_{X_i}\Big[\int \frac{W_{XV}(X_i,u)}{f_{TXV}(t,X_i,u)}\Big(F_{Y|TXS}(y|t,X,S)-F_{Y|TXV}(y|t,X_i,u)\Big) \\
&\quad K_h(X-X_i)\big(K_h(v_1(t,S)-u)-K_h(v_2(t,S)-u)\big)f_{V|X}(u|X_i)du\ f_{T|XS}(t|X,S)\Big] \\
&= \mathbb{E}_{XS}\Big[\Big(\frac{W_{XV}(X,v_1(t,S))}{f_{T|XV}(t|X,v_1(t,S))}\Big(F_{Y|TXS}(y|t,X,S)-F_{Y|TXV}(y|t,X,v_1(t,S))\Big) \\
&\quad -\frac{W_{XV}(X,v_2(t,S))}{f_{T|XV}(t|X,v_2(t,S))}\Big(F_{Y|TXS}(y|t,X,S)-F_{Y|TXV}(y|t,X,v_2(t,S))\Big)\Big)\ f_{T|XS}(t|X,S)\Big]+O_p(h^r) \\
&= \mathbb{E}_{XS}\Big[\Big\{F_{Y|TXS}(y|t,X,S)\Big(-\frac{W_{XV}(X,v_2(t,S))}{f_{T|XV}(t|X,v_2(t,S))}\nabla_v f_{T|XV}(t|X,v_2(t,S))+\nabla_v W_{XV}(X,v_2(t,S))\Big) \\
&\quad -W_{XV}(X,v_2(t,S))\nabla_v F_{Y|TXV}(y|t,X,v_2(t,S)) \\
&\quad +W_{XV}(X,v_2(t,S))F_{Y|TXV}(y|t,X,v_2(t,S))\frac{\nabla_v f_{T|XV}(t|X,v_2(t,S))}{f_{T|XV}(t|X,v_2(t,S))} \\
&\quad -F_{Y|TXV}(y|t,X,v_2(t,S))\nabla_v W_{XV}(X,v_2(t,S))\Big\}'\Big(v_1(t,S)-v_2(t,S)\Big)\frac{f_{T|XS}(t|X,S)}{f_{T|XV}(t|X,v_2(t,S))}\Big] \\
&\quad +O\big(\|v_1-v_2\|_\infty^2\big)+O_p(h^r)
\end{aligned}
$$

where the second equality is followed by the conditional expectation by $f_{W|XV}(W_i|X_i,V_i)$. The last equality is by a Taylor expansion for any $v_1, v_2 \in \bar{\mathcal{M}}_n$. Together with the other term in $T_{1,n}$ $\frac{1}{n}\sum_{i=1}^n \hat{F}_{Y|TV}(y|t,V_i) - F_{Y|TV}(y|t,V_i)$ by Theorem 1 and $T_{2,n}$ in (23), Theorem 2 is derived.

D.2 Examples: Proofs of Corollaries 1 and2

The stochastic expansions in Theorem 2 are for general generated regressor satisfying regularity conditions. To apply these results to the economic examples, it is useful to express the estimation error of the generated regressors by the linear representation of the form $\hat{v}(T,S) - v(T,S) = n^{-1}\sum_{i=1}^n \vartheta_{in}(T,S) + R_n^v(T,S)$ for some ϑ_{in} with $\mathbb{E}[\vartheta_{in}(t,s)] = O_p(R_n^b) = o_p(1)$ and $\mathbb{E}[\|\vartheta_{in}(t,s)\|^2] < \infty$ for all t,s. The remainder term R_n^v is controlled to be of smaller order. Then the linear representation implies the influence form estimating the generated regressor

$$
\Delta_{ARG}\big(y,t,\hat{v}(T,S),v_0(T,S)\big) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\vartheta_{in}(T,S)\cdot ARG(y,t,X,W,v_0(T,S))\big] + \bar{O}_p(\|R_n^v\|_\infty).
$$

Δ_{REG} is expressed similarly.

D.2.1 Proof of Corollary 1

The linear representation for the kernel regression follows the same decomposition in (31).

$$\hat{F}_{T_1|Z}(t_1|z) - F_{T_1|Z}(t_1|z) = \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{1}_{\{T_{1i}\le t_1\}} - F_{T_1|Z}(t_1|z)\big)K_{h_1}(Z_i - z)/f_Z(z) + R_n^v(t_1, z)$$

$$\mathbb{E}[T_1|Z=z] - \hat{\mathbb{E}}[T_1|Z=z] = -\frac{1}{n}\sum_{i=1}^{n}\big(T_{1i} - \mathbb{E}[T_1|Z=z]\big)K_{h_1}(Z_i - z)/f_Z(z) + R_n^v(t_1, z)$$

where $\|R_n^v\|_\infty = O_p\big(\big(\sqrt{\log n/(nh_1^{d_z})} + h_1^{r_1}\big)^2\big)$. A standard algebra for kernel gives the results.

D.2.2 Proof of Corollary 2

The influence of estimating the GPS is $\mathbb{E}\big[A(y,X,W)\cdot\big(\hat{f}_{T|X}(t|X) - f_{T|X}(t|X)\big)\big]$, where $A(y,X) = \nabla_v F_{Y|TV}(y|t,v_0(X))\big(W - \mathbb{E}[W|v_0(X)]\big) + \big(F_{Y|TV}(y|t,v_0(X)) - F_{Y|TX}(y|t,X)\big)\cdot\big(\mathbb{E}[W|v_0(X)]/f_{T|X}(t|X) - \nabla_v\mathbb{E}[W|V=v]\big|_{v=v_0(X)}\big)$. The linear representation is

$$\hat{f}_{T|X}(t|x) - f_{T|X}(t|x) = \frac{1}{n}\sum_{i=1}^{n}\Big(K_{h_1}(T_i - t) - f_{T|X}(t|x)\Big)K_{h_1}(X_i - x)/f_X(x) + R_n^v(t,x)$$

$$\text{where } \sup_{t,x}\big|R_n^v\big| = O_p\Big(\|\hat{f}_X - f_X\|_\infty\|\hat{f}_{T|X} - f_{T|X}\|_\infty\Big) = O_p\bigg(\Big(\sqrt{\frac{\log n}{nh_1^{d_x}}} + h_1^{r_1}\Big)\Big(\sqrt{\frac{\log n}{nh_1^{d_x+1}}} + h_1^{r_1}\Big)\bigg).$$

Then $\mathbb{E}\big[A(y,X,W)\cdot\big(\hat{f}_{T|X}(t|X) - f_{T|X}(t|X)\big)\big]$

$$\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\Big(K_{h_1}(T_i - t) - f_{T|X}(t|X)\Big)\mathbb{E}\big[A(y,X_i,W)\big|X = X_i\big] + O_p(h_1^{r_1} + \|R_n^v\|_\infty)\\
&= O_p\big((nh_1)^{-1/2} + n^{-1/2} + h_1^{r_1} + \|R_n^v\|_\infty\big).
\end{aligned}$$

Assume $(2r_1+1)^{-1} < g < (2d_x)^{-1}$ so that $O_p\big(h_1^{r_1} + \|R_n^v\|_\infty\big) = o_p\big((nh_1)^{-1/2}\big)$.

Therefore, if $h_2 = O(h_1)$,

$$\begin{aligned}
&\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_{Y|T\hat{V}}(y|t,\hat{v}(X_i))W(X_i) - \mathbb{E}[F_{Y|TV}(y|t,v_0(X))W(X)]\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}(y;V,W)\\
&+ \sqrt{\frac{h_2}{h_1}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{\sqrt{h_1}}K\Big(\frac{T_i - t}{h_1}\Big)\bigg\{\nabla_v F_{Y|TV}(Y|t,v_0(X_i))\big(\mathbb{E}[W|X_i] - \mathbb{E}[W|v_0(X_i)]\big)\\
&+ \Big(F_{Y|TV}(y|t,v_0(X_i)) - F_{Y|TX}(y|t,X_i)\Big)\Big(\frac{\mathbb{E}[W|v_0(X_i)]}{f_{T|X}(t|X_i)} - \nabla_v\mathbb{E}[W|V=v]|_{v=v_0(X_i)}\Big)\bigg\} + o_p(1).
\end{aligned}$$

When $h_1 = h_2$ and $W = 1$, the influence function is reduced to $n^{-1/2}\sum_{i=1}^{n}\psi_{tin}(y;X,W=1)$.

Lemma D.1
Suppose $A(X)$ is a positive function of X. Let B be any measurable function of Y such that the following moments exist. Then

$$\mathbb{E}\big[var\big(B(Y)\big|T=t,X\big)\cdot A(X)\big] \le \mathbb{E}\big[var\big(B(Y)\big|T=t,V(X)\big)\cdot A(X)\big].$$

Equality holds if and only if $\mathbb{E}[B(Y)|T=t,V(X)] = \mathbb{E}[B(Y)|T=t,X]$ almost surely, when $A(X) > 0$ almost surely.

Proof of Lemma D.1
First note that $\mathbb{E}\big[var\big(B(Y)\big|T=t,X\big)\cdot A(X)\big] = \mathbb{E}\Big[var\big(B(Y)\big|T=t,X\big)\cdot A(X)\cdot \frac{f_T(t)}{f_{T|X}(t|X)}\Big|T=t\Big]$. And $\frac{f_T(t)}{f_{T|X}(t|X)}$ is a function of X. So we could abuse the notation of $A(X)$ and prove $\mathbb{E}\big[var\big(B(Y)\big|T=t,V(X)\big)\cdot A(X)\big|T=t\big] \geq \mathbb{E}\big[var\big(B(Y)\big|T=t,X\big)\cdot A(X)\big|T=t\big]$.

By the law of iterated expectations,

$$\begin{aligned}&\mathbb{E}\big[var\big(B(Y)\big|T=t,V(X)\big)\cdot A(X)\big|T=t\big]\\&=\mathbb{E}\big[\mathbb{E}\big[\big(B(Y)-\mathbb{E}[B(Y)|T=t,V(X)]\big)^2\big|T=t,X\big]\cdot A(X)\big|T=t\big].\end{aligned}$$

We could skip conditioning on $T=t$ for notational ease and observe that

$$\begin{aligned}&\mathbb{E}\big[\big(B(Y)-\mathbb{E}[B(Y)|V(X)]\big)^2\big|X\big]-\mathbb{E}\big[\big(B(Y)-\mathbb{E}[B(Y)|X]\big)^2\big|X\big]\\&=\mathbb{E}[B^2(Y)|X]-2\mathbb{E}[B(Y)|X]\cdot\mathbb{E}[B(Y)|V(X)]+\mathbb{E}[B(Y)|V(X)]^2-\mathbb{E}[B^2(Y)|X]+\mathbb{E}[B(Y)|X]^2\\&=\big(\mathbb{E}[B(Y)|V(X)]-\mathbb{E}[B(Y)|X]\big)^2\geq 0.\end{aligned}$$

D.3 Semiparametric models

We start with a simple *full mean* that averages out all the arguments in the second-step regression function, defined as $\mathbb{E}\big[F_{Y|\Lambda}(y|\Lambda)\cdot W\big] \equiv \int \mathbb{E}\big[\mathbf{1}_{\{Y\leq y\}}\big|\Lambda=\lambda\big]\cdot w\, dF_{\Lambda W}(\lambda,w)$. The estimator follows the procedure described in Section 3 by dropping the treatment variable T. That is, $n^{-1}\sum_{i=1}^n \hat{F}_{Y|\hat{\Lambda}}(y|\hat{\Lambda}_i)\cdot W_i$ where $\Lambda=(X',V')'$, the generated regressor $V=v_0(S)$, and $\hat{V}_i=\hat{v}(S_i)$.

Corollary 6 (Full mean with generated regressors)
Assume the conditions in Theorem 2 hold with $d_t=0$. Then uniformly in $y\in\mathcal{Y}$,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(\hat{F}_{Y|X\hat{V}}(y|X_i,\hat{V}_i)W_i-\mathbb{E}\big[F_{Y|XV}(y|X,V)W\big]\Big)\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(F_{Y|XV}(y|X_i,V_i)W_i-\mathbb{E}\big[F_{Y|XV}(y|X,V)W\big]\\&\qquad\qquad+\Big(\mathbf{1}_{\{Y_i\leq y\}}-F_{Y|XV}(y|X_i,V_i)\Big)\cdot\mathbb{E}\big[W\big|X=X_i,V=V_i\big]\Big)\\&\quad+\sqrt{n}\,\Delta\big(\hat{v}(S)\big)+\sqrt{n}\,R_n+o_p(1)\end{aligned}$$

where for any $v_1\in\bar{\mathcal{M}}_n$

$$\begin{aligned}\Delta(v_1(S))\equiv\mathbb{E}\Big[\Big(v_1(S)-v_0(S)\Big)'\Big\{\Big(W-\mathbb{E}[W|X,v_0(S)]\Big)\cdot\nabla_v F_{Y|XV}(y|X,v)\Big|_{v=v_0(S)}\\-\Big(F_{Y|XV}(y|X,v_0(S))-F_{Y|XS}(y|X,S)\Big)\cdot\nabla_v\mathbb{E}[W|X,V=v]\Big|_{v=v_0(S)}\Big\}\Big]\end{aligned}$$

D.3.1 Proof of Corollary 6 (Full mean)

First consider the estimator with observable regressors.

$$\sqrt{n}\big(n^{-1}\sum_{i=1}^{n}\hat{F}_{Y|\Lambda}(y|\Lambda_i)W_i - \mathbb{E}\big[F_{Y|\Lambda}(y|\Lambda)W\big]\big) = G_n\big[\hat{F}_{Y|\Lambda}(y|\Lambda_i)W_i - F_{Y|\Lambda}(y|\Lambda_i)W_i\big]$$
$$+\ G_n\big[F_{Y|\Lambda}(y|\Lambda_i)W_i\big] \tag{26}$$
$$+\ \sqrt{n}\ \mathbb{E}\big[\hat{F}_{Y|\Lambda}(y|\Lambda)W - F_{Y|\Lambda}(y|\Lambda)W\big]. \tag{27}$$

The first term is $\bar{o}_p(1)$ by the stochastic equicontinuity result in Lemma A.1. The second term (26) is $\bar{O}_p(1)$, by the Donsker property of $\{(\Lambda, W) \mapsto F_{Y|\Lambda}(y|\Lambda)W : y \in \mathcal{Y}\}$. The third term (27) contributes the influence from estimating the conditional CDF.

$$\mathbb{E}\big[\hat{F}_{Y|\Lambda}(y|\Lambda)W - F_{Y|\Lambda}(y|\Lambda)W\big] = \int \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|\Lambda}(y|\Lambda_i)\Big)\ K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{\hat{f}_\Lambda(\lambda)}dF_\Lambda(\lambda)$$
$$+\int \frac{1}{n}\sum_{i=1}^{n}\Big(F_{Y|\Lambda}(y|\Lambda_i) - F_{Y|\Lambda}(y|\lambda)\Big)\ K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{\hat{f}_\Lambda(\lambda)}dF_\Lambda(\lambda).$$

The proof is the same as the proof of Theorem 1 in Section C by dropping $K_h(t-T_i)$ and replacing $f_{T\Lambda}$ with f_Λ. The influence function of (27) is dominated by $\psi_{in}(y) \equiv \big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|\Lambda}(y|\Lambda_i)\big)W_\Lambda(\Lambda_i)$.

We now show the remainder terms as in (16) is $\bar{o}_p(n^{-1/2})$. We only note the difference in the following. This term $u'(\Lambda_i)\int vK^2(v)dv$ for $i=j$ contributes to a smaller order term: $n^{-2}\sum_{i=1}^{n}D_i'(\Lambda_i)\int vK^2(v)dv = \bar{O}_p(n^{-1}) = \bar{o}_p(n^{-1/2})$. For the projection of the U-process, define

$$\bar{f}_\Lambda(\Lambda_i) \equiv \mathbb{E}\Big[\bar{K}_h(\Lambda_i-\Lambda_j)\Big|Z_i\Big] = \int \frac{1}{h^{2d_\lambda}}\int K\Big(\frac{\Lambda_i - v}{h}\Big)K\Big(\frac{\Lambda - v}{h}\Big)dv\cdot f_\Lambda(\Lambda)d\Lambda$$
$$= \int \frac{1}{h^{d_\lambda}}K\Big(\frac{\Lambda_i - v}{h}\Big)\int K(s)\ f_\Lambda(v+sh)ds\ dv = f_\Lambda(\Lambda_i) + O_p(h^r)$$
$$H(Z_i, Z_j; y, h) \equiv D_i(\Lambda_i)\cdot\Big(\bar{K}_h(\Lambda_i-\Lambda_j) - \bar{f}_\Lambda(\Lambda_i)\Big).$$

Then consider (20)

$$\frac{1}{n^2}\sum_i\sum_{j\ne i}H(Z_i,Z_j;y,h) + \frac{1}{n^2}\sum_i H(Z_i,Z_i;y,h) + \bar{o}_p(n^{-1/2}).$$

The second term is $\bar{o}_p(1/\sqrt{n})$, because its second part is $n^{-1}(15)$ and its first part is smaller than $n^{-2}\sum_i u(\Lambda_i)h^{-d_\lambda}\int K^2(v)dv = \bar{O}_p((nh^{d_\lambda})^{-1}) = \bar{o}_p(n^{-1/2})$. The first term in (20) is a degenerate second order U-process. By Corollary 4 (ii) in Sherman (1994),

$$\sup_{y\in\mathcal{Y}}\Big|\frac{1}{n^2}\sum_i\sum_{j\ne i}h^{d_\lambda}H(Z_i,Z_j;y,h)\Big| = O_p\Big(\frac{1}{n}\Big).$$

Therefore, by the bandwidth assumption,

$$\sup_{y\in\mathcal{Y}}\Big|\frac{1}{n^2}\sum_i\sum_{j\ne i}H(Z_i,Z_j;y,h)\Big| = O_p\Big(\frac{1}{nh^{d_\lambda}}\Big) = o_p\Big(\frac{1}{\sqrt{n}}\Big).$$

Now, consider the estimation error from the generated regressors. All the stochastic equicontinuity Lemmas in Section A carry over by dropping the variables T.

D.3.2 Proof of Corollary 3 (GMM)

Following similar steps in proving Corollary 6, first for the observable regressors

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n}\rho(\hat{\mathbb{E}}[Y|\Lambda_i],W_i)-\mathbb{E}[\rho(\mathbb{E}[Y|\Lambda],W)]\Big)=G_n\big[\rho(\hat{\mathbb{E}}[Y|\Lambda_i],W_i)-\rho(\mathbb{E}[Y|\Lambda_i],W_i)\big]$$
$$+\;G_n\big[\rho(\mathbb{E}[Y|\Lambda_i],W_i)\big]+\sqrt{n}\,\mathbb{E}\big[\rho(\hat{\mathbb{E}}[Y|\Lambda],W)-\rho(\mathbb{E}[Y|\Lambda],W)\big].$$

The last term $\sqrt{n}\,\mathbb{E}\big[\rho(\hat{\mathbb{E}}[Y|\Lambda],W)-\rho(\mathbb{E}[Y|\Lambda],W)\big]=\sqrt{n}\,\mathbb{E}\big[\partial_m\rho(\mathbb{E}[Y|\Lambda],W)\,(\hat{\mathbb{E}}[Y|\Lambda]-\mathbb{E}[Y|\Lambda])\big]+o_p(1)$ by assuming $\sup_\lambda|\hat{\mathbb{E}}[Y|\Lambda=\lambda]-\mathbb{E}[Y|\Lambda=\lambda]|=o_p(n^{-1/4})$ implied by Assumption 11 with $d_t=0$. Therefore, the proof is the same as for (27) by replacing the weight W with $\partial_m\rho(\mathbb{E}[Y|\Lambda],W)$.

Now, consider the estimation error from the generated regressors. Modify the functional in Section D.1.1 to be $S_n(f)\equiv n^{-1}\sum_{i=1}^n\rho(f_1\big(t,f_2(S_i)\big),W_i)-\theta_0$. The directional derivative

$$\dot{S}_n(\bar{f})[f-\bar{f}]=\lim_{s\to0}\frac{1}{s}\Big(S_n\big(\bar{f}+s(f-\bar{f})\big)-S_n(\bar{f})\Big)$$
$$=\lim_{s\to0}\frac{1}{s}\frac{1}{n}\sum_{i=1}^{n}\Big\{\rho\Big(\Big[\bar{f}_1+s(f_1-\bar{f}_1)\Big]\Big(t,\big[\bar{f}_2+s(f_2-\bar{f}_2)\big](S_i)\Big),W_i\Big)-\rho\Big(\bar{f}_1\big(t,\bar{f}_2(S_i)\big),W_i\Big)$$
$$-\rho\Big(\bar{f}_1\Big(t,\big[\bar{f}_2+s(f_2-\bar{f}_2)\big](S_i)\Big),W_i\Big)+\rho\Big(\bar{f}_1\Big(t,\big[\bar{f}_2+s(f_2-\bar{f}_2)\big](S_i)\Big),W_i\Big)\Big\}$$
$$=\frac{1}{n}\sum_{i=1}^{n}\partial_m\rho(\bar{f}_1(\bar{f}_{2i}),W_i)\big[f_1-\bar{f}_1\big]\big(t,\bar{f}_2(S_i)\big)+\partial_m\rho(\bar{f}_1(\bar{f}_{2i}),W_i)\bar{f}_1^{(v)}\big(t,\bar{f}_2(S_i)\big)\cdot\big[f_2-\bar{f}_2\big](S_i)$$

Define

$$T_{1,n}(\hat{f})\equiv\frac{1}{n}\sum_{i=1}^{n}\partial_m\rho\big(\mathbb{E}[Y|V=v_0(S_i)],W_i\big)\cdot\Big(\hat{\mathbb{E}}[Y|\hat{V}=v_0(S_i)]-\mathbb{E}[Y|V=v_0(S_i)]\Big)$$
$$T_{2,n}(\hat{f})\equiv\frac{1}{n}\sum_{i=1}^{n}\partial_m\rho\big(\mathbb{E}[Y|V=v_0(S_i)],W_i\big)\cdot\nabla_v\mathbb{E}\big[Y\big|T=t,V=v\big]\big|'_{v=v_0(S_i)}\Big(\hat{v}(S_i)-v_0(S_i)\Big).$$

The rest of the proof is the same as Corollary 6 by replacing the weight W with $\partial_m\rho(\mathbb{E}[Y|V],W)$.

D.3.3 Proof of Corollary 4 (Discrete Treatment)

First for the estimator with observable regressors, we follow the proof of Theorem 1 by replacing the kernel $K_h(T-t)$ with $\mathbf{1}_{\{T=t\}}$ and replacing $f_{T|\Lambda}$ with $P_t(\Lambda)$ for discrete treatment T. The modification follows the proof of Corollary 6 for the full mean in the previous section.

The influence function of $\sqrt{n}\;\mathbb{E}\big[\hat{F}_{Y|T\Lambda}(y|t,\Lambda)W-F_{Y|T\Lambda}(y|t,\Lambda)W\big]$ is dominated by $\big\{\psi_{tin}(y)\equiv\big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|T\Lambda}(y|t,\Lambda_i)\big)\frac{\mathbf{1}_{\{T_i=t\}}}{P_t(\Lambda_i)}W_\Lambda(\Lambda_i):y\in\mathcal{Y}\big\}$ which is assumed to be Donsker.

We now show the remainder terms as in (16) is $o_p(n^{-1/2})$. We note the difference in the following. For the projection of the U-process, define $\bar{f}_{T\Lambda}(t,\Lambda_i)\equiv\mathbb{E}\big[\bar{K}_h(\Lambda_i-\Lambda_j)\mathbf{1}_{\{T_j=t\}}\big|Z_i\big]$

$$=\int\frac{1}{h^{2d_\lambda}}\int K\Big(\frac{\Lambda_i-v}{h}\Big)K\Big(\frac{\Lambda-v}{h}\Big)dv\cdot f_{T\Lambda}(t,\Lambda)d\Lambda$$
$$=\int\frac{1}{h^{d_\lambda}}K\Big(\frac{\Lambda_i-v}{h}\Big)\int K(s)\;f_{T\Lambda}(t,v+sh)ds\;dv=f_{T\Lambda}(t,\Lambda_i)+O_p(h^r).$$

Define

$$H(Z_i,Z_j;y,h)\equiv D_i(\Lambda_i)\cdot\mathbf{1}_{\{T_i=t\}}\big(\bar{K}_h(\Lambda_i-\Lambda_j)\mathbf{1}_{\{T_j=t\}}-\bar{f}_\Lambda(\Lambda_i)\big).$$

Then the rest of the proof is the same as the proof of Corollary 6 in the previous section.

Now, consider the estimation error from the generated regressors. All the stochastic equicontinuity Lemmas in Section A carry over by replacing $K_h(T-t)$ with $\mathbf{1}_{\{T=t\}}$ and $f_{T|XS}$ with $P_t(X,S)$.

Consider the special case when the generated regressor is the PS estimated by a kernel regression, $\hat{P}_t(x) - P_t(x) = n^{-1}\sum_{i=1}^n \big(\mathbf{1}_{\{T_i=t\}} - P_t(x)\big)K_{h_1}(X_i - x)/f_X(x) + O_p\big(\|R_n^v\|_\infty\big)$, where $\|R_n^v\|_\infty = O_p\big(\big(\sqrt{\log n/(nh_1^{d_x})} + h_1^{r_1}\big)^2\big)$. Then

$$\begin{aligned}\Delta_{ARG} &= \mathbb{E}\big[(\hat{P}_t(X) - P_t(X))\frac{\partial}{\partial v}F_{Y|TV}(y|t,v)\big|_{v=P_t(X)} \cdot W\big] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big(\mathbf{1}_{\{T_i=t\}} - P_t(X)\big)\frac{K_{h_1}(X_i - X)}{f_X(X)}\frac{\partial}{\partial v}F_{Y|TV}(y|t,v)\big|_{v=P_t(X)} \cdot \mathbb{E}[W|X]\Big] + O_p(\|R_n^v\|_\infty).\end{aligned}$$

A simple algebra expanding $\Delta_{ARG} + \Delta_{REG}$ gives the result. Note that $\hat{F}_{Y(\bar{t})|T}(y|\bar{t}) = \hat{F}_{Y|T}(y|\bar{t}) = n^{-1}\sum_{i=1}^n \mathbf{1}_{\{Y_i\le y\}}W_i$ contributes to the influence function by $n^{-1}\sum_{i=1}^n \mathbf{1}_{\{Y_i\le y\}}W_i - F_{Y(\bar{t})|T}(y|\bar{t})$.

E Inference for the Treatment Effects

E.1 Proof of Theorem 3

By the functional delta method (e.g., Theorem 3.9.4 in van der Vaart and Wellner (1996)) and the linearity of the Hadamard derivative, the weak convergence to a Gaussian process is implied. Using the results in (21), for the diagonal term $t \neq s$, $\lim_{n\to\infty}\mathbb{E}\big[\psi_{tin}(y_1)\psi_{si}(y_2)\big] = 0$. It follows $Cov = \lim_{n\to\infty}\mathbb{E}\big[\big(\Gamma'_\theta(\psi_{tin}) - \Gamma'_\theta(\psi_{\bar{t}i})\big)^2\big] = \lim_{n\to\infty}\mathbb{E}\big[\big(\Gamma'_\theta(\psi_{tin})^2\big] + \mathbb{E}\big[\Gamma'_\theta(\psi_{\bar{t}i})^2\big]$.

Mean

Using integration by parts, $\int_{\mathcal{Y}} y d\mathbf{1}_{\{Y\le y\}} = Y$. So $\Gamma'\big(\mathbf{1}_{\{Y\le y\}} - F_{Y|TX}(y|t,X)\big) = Y - \mathbb{E}[Y|t,X]$.

Quantile processes

The Hadamard derivative is shown in Example 3.9.24 in van der Vaart and Wellner (1996).

E.2 Proof of Theorem 4 (Multiplier method)

Decompose

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\hat{\psi}_{tin}(y) = \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\psi_{tin}(y) + \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\big(\hat{\psi}_{tin}(y) - \psi_{tin}(y)\big)$$

We use the functional CLT by checking the conditions of Theorem 10.6 in Pollard (1990), as in the proof of Lemma C.1. We show the first term $n^{-1/2}\sum_{i=1}^n U_i\psi_{tin}(\cdot) \Rightarrow \mathbb{G}_t(\cdot)$. The second empirical process is $\bar{o}_p(1)$ by showing its weak convergence to a Gaussian process with zero covariance kernel.

Define $f_{ni}^u(y) = U_i f_{ni}(y) = U_i n^{-1/2}\psi_{tin}(y)$ whose envelope is $F_{ni}^u = U_i F_{ni} = U_i n^{-1}h^{d_t/2}K_h\big(T_i - t\big)W_\Lambda(\Lambda_i)/f_{t|\Lambda_i}$. Then (i) holds. $\mathbb{E}f_{ni}^u(y) = 0$ and $\mathcal{Z}_n^u(y) = \sum_{i=1}^n f_{ni}^u(y)$.

(ii) $\mathbb{E}[\mathcal{Z}_n^u(y_1)\mathcal{Z}_n^u(y_2)] = \mathbb{E}\big[\sum_{i=1}^n f_{ni}^u(y_1)f_{ni}^u(y_2)\big] = n^{-1}\sum_{i=1}^n \psi_{tin}(y_1)\psi_{tin}(y_2)\cdot\mathbb{E}U_i^2 \xrightarrow{p} H(y_1,y_2)$ defined in (22), by the weak law of large number.

(iii) $\sum_{i=1}^n \mathbb{E}{F_{ni}^u}^2 = n^{-1}\sum_{i=1}^n h^{d_t}K_h^2\big(T_i - t\big)W_\Lambda(\Lambda_i)^2/f_{t|\Lambda_i}^2 \to \int K^2(u)du\cdot\mathbb{E}[W_\Lambda(\Lambda)^2 f_{t|\Lambda}^{-1}]$.

(iv) $B \equiv \inf_{(\Lambda_i,T_i)} f_{t|\Lambda_i}\big(W_\Lambda(\Lambda_i)^2h^{d_t}K_h\big(T_i - t\big)\big)^{-1}$ exists, because $f_{T|\Lambda}$ is bounded away from zero,

the weight and the kernel are uniformly bounded. For any $\epsilon > 0$,

$$\sum_{i=1}^{n} \mathbb{E}{F^u_{ni}}^2 \, \mathbf{1}\{F_{ni} > \epsilon\} = \frac{h^{d_t}}{n} \sum_{i=1}^{n} K_h^2\big(T_i - t\big) \frac{W_\Lambda(\Lambda_i)^2}{f^2_{t|\Lambda_i}} \cdot \mathbb{E}\Big[U_i^2 \, \mathbf{1}\Big\{U_i > \frac{\sqrt{n}\epsilon f_{t|\Lambda_i}}{\sqrt{h^{d_t}} K_h\big(T_i - t\big) W_\Lambda(\Lambda_i)}\Big\}\Big]$$

$$\leq \frac{h^{d_t}}{n} \sum_{i=1}^{n} K_h^2\big(T_i - t\big) \frac{W_\Lambda(\Lambda_i)^2}{f^2_{t|\Lambda_i}} \cdot \mathbb{E}\Big[U_i^2 \, \mathbf{1}\Big\{U_i > \sqrt{nh^{d_t}}\epsilon B\Big\}\Big] \to \int K^2(u)du \cdot \mathbb{E}\big[W_\Lambda(\Lambda_i)^2 f^{-1}_{t|\Lambda}\big] \cdot 0.$$

(v) Denote $F_{Y|T\Lambda}(y|t,\Lambda_i) = F_y$ and $\mathbf{1}_{\{Y_i \leq y\}} = \mathbf{1}_y$. Then for any $y_1 \leq y_2$, $\rho^u_n(y_1,y_2)^2$

$$= \sum_{i=1}^{n} \mathbb{E}\Big(f^u_{ni}(y_1) - f^u_{ni}(y_2)\Big)^2 = \frac{1}{n} \sum_{i=1}^{n} \Big(\psi^2_{tin}(y_1) + \psi^2_{tin}(y_2) - 2\psi_{tin}(y_1)\psi_{tin}(y_2)\Big)$$

$$= \frac{h^{d_t}}{n} \sum_{i=1}^{n} K_h^2\big(T_i - t\big) W_\Lambda(\Lambda_i)^2 \Big[\mathbf{1}_{y_1} - 2\mathbf{1}_{y_1}F_{y_1} + F^2_{y_1} + \mathbf{1}_{y_2} - 2\mathbf{1}_{y_2}F_{y_2} + F^2_{y_2} \tag{28}$$

$$- 2\mathbf{1}_{y_1} + 2\mathbf{1}_{y_1}F_{y_2} + 2\mathbf{1}_{y_2}F_{y_1} - 2F_{y_1}F_{y_2}\Big] \Big/ f^2_{t|\Lambda_i} \tag{29}$$

$$= \frac{h^{d_t}}{n} \sum_{i=1}^{n} \frac{K_h^2\big(T_i - t\big)}{f^2_{t|\Lambda_i}} W_\Lambda(\Lambda_i)^2 \Big[\mathbf{1}_{y_1}\big(-1 - 2F_{y_1} + 2F_{y_2}\big) + \mathbf{1}_{y_2}\big(1 - 2F_{y_2} + 2F_{y_1}\big) + \big(F_{y_1} - F_{y_2}\big)^2\Big]$$

$$\to \int K^2(u)du \cdot \mathbb{E}\Big[\frac{W_\Lambda(\Lambda_i)^2}{f_{t|\Lambda}}\Big(F_{y_1}(-1 - 2F_{y_1} + 2F_{y_2}) + F_{y_2}(1 - 2F_{y_2} + 2F_{y_1}) + (F_{y_1} - F_{y_2})^2\Big)\Big]$$

$$= \int K^2(u)du \cdot \mathbb{E}\Big[\frac{W_\Lambda(\Lambda)^2}{f_{t|\Lambda}}(F_{y_2} - F_{y_1})(1 - F_{y_2} + F_{y_1})\Big] \equiv \rho^u(y_1,y_2)^2.$$

It remains to show that for all deterministic sequences $\{y_{1n}\}$ and $\{y_{2n}\}$ such that $\rho^u(y_{1n},y_{2n}) \to 0$, $\rho^u_n(y_{1n},y_{2n}) \to 0$. Using the same argument in Lemma C.1, $\sqrt{nh^{d_t}}\big\{n^{-1}\sum_{i=1}^n \psi^2_{tin}(y_1) - \int K^2(u)du \cdot \mathbb{E}\big[W_\Lambda(\Lambda)^2\big(F_{y_1} - F^2_{y_1}\big)/f_{t|\Lambda}\big]\big\}$ converges to a Gaussian process of y_1. So the first part $n^{-1}\sum_{i=1}^n \big(\psi^2_{tin}(y_1) + \psi^2_{tin}(y_2)\big)$ in (28) converges uniformly in y_1, y_2.
For the second part $-2n^{-1}\sum_{i=1}^n \psi_{tin}(y_1)\psi_{tin}(y_2)$ in (29) indexed by both y_1 and y_2, we focus on one of the terms, defining $A_n(y_1,y_2) \equiv n^{-1}\sum_{i=1}^n h^{d_t}K_h^2\big(T_i - t\big)W_\Lambda(\Lambda_i)^2\mathbf{1}_{y_1}F_{y_2}/f^2_{t|\Lambda_i}$ and $A(y_1,y_2) \equiv \int K^2(u)du \cdot \mathbb{E}\big[F_{y_1}F_{y_2}W_\Lambda(\Lambda)^2/f_{t|\Lambda}\big]$. It suffices to show that for all deterministic sequences $\{y_{1n}\}$ and $\{y_{2n}\}$ such that $A(y_{1n},y_{2n}) \to 0$, $A_n(y_{1n},y_{2n}) \to 0$. By assumption, $W_\Lambda(\Lambda)^2/f_{t|\Lambda} < \delta < \infty$. $A(y_{1n},y_{2n}) \to 0$ means that for any $\epsilon > 0$, there exists an integer N_0 such that for $n > N_0$,

$$A(y_{1n},y_{2n}) \leq \int K^2(u)du \cdot \delta \cdot \mathbb{E}\big[F_{Y|T\Lambda}(y_{1n}|t,\Lambda)F_{Y|T\Lambda}(y_{2n}|t,\Lambda)\big]$$
$$\leq C \cdot \mathbb{E}\big[F_{Y|T\Lambda}\min\{y_{1n},y_{2n}\}|t,\Lambda)\big] < \epsilon \tag{30}$$

defining $C = \int K^2(u)du \cdot \delta$ for notational ease. Actually, $\min\{y_{1n},y_{2n}\}$ can be either y_{1n} or y_{2n}. It's not required both the deterministic sequence to converge.
Since $F_{Y|T\Lambda}$ is increasing in y, there exists y_0 such that $C \cdot \mathbb{E}[F_{Y|T\Lambda}(y_0|t,\Lambda)] = \epsilon$. Then for any $y < y_0$, $C \cdot \mathbb{E}[F_{Y|T\Lambda}(y|t,\Lambda)] \leq \epsilon$. Then (30) implies either $y_{1n} < y_0$ or $y_{1n} < y_0$ or both for $n > N_0$.
First, note that $n^{-1}\sum_{i=1}^n h^{d_t}K_h^2\big(T_i - t\big)\mathbf{1}_{\{Y_i \leq y_0\}}/f_{t|\Lambda_i} \to \int K^2(u)du\mathbb{E}\big[F_{y_0|t\Lambda}\big]$, i.e., for any $\epsilon_1 > 0$, there exists an integer N_1 such that $\big|\frac{1}{n}\sum_{i=1}^n h^{d_t}K_h^2\big(T_i - t\big)\mathbf{1}_{\{Y_i \leq y_0\}}/f_{t|\Lambda_i} - \int K^2(u)du\mathbb{E}\big[F_{y_0|t\Lambda}\big]\big| <$

ϵ_1, for $n > N_1$. For the case $y_{1n} < y_0$, for $n > \max\{N_1, N_0\}$,

$$A_n(y_{1n}, y_{2n}) \leq \delta \frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2(T_i - t)\mathbf{1}_{y_{1n}} F_{y_{2n}} \Big/ f_{t|\Lambda_i} \leq \delta \frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2(T_i - t)\mathbf{1}_{\{Y_i \leq y_0\}} \Big/ f_{t|\Lambda_i}$$
$$\leq \delta \int K^2(u)du\mathbb{E}\big[F_{y_0|t\Lambda}\big] + \delta\epsilon_1 = \epsilon + \delta\epsilon_1.$$

For the other case $y_{2n} < y_0$, use the similar argument by $\frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2(T_i - t)F_{y_0} \Big/ f_{t|\Lambda_i} \to \int K^2(u)du\mathbb{E}\big[F_{y_0|t\Lambda}\big]$. Then it's shown $A_n(y_{1n}, y_{2n}) \to 0$. The same argument applies to other terms in (29).

Therefore, the FCLT implies $\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i \psi_{tin}(\cdot) \Rightarrow \mathbb{G}_t(\cdot)$. Next we need to show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\Big(\hat{\psi}_{tin}(y) - \psi_{tin}(y)\Big) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\sqrt{h^{d_t}} K_h(T_i - t)\Big(\hat{\varphi}_{tin}(y) - \varphi_{tin}(y)\Big) = \bar{o}_p(1)$$

where $\varphi_{tin}(y) = \big(\mathbf{1}_{\{Y_i \leq y\}} - F_{Y|T\Lambda}(y|t, \Lambda_i)\big)\mathbb{E}[W|\Lambda = \Lambda_i]/f_{T|\Lambda}(t|\Lambda_i)$ and a consistent estimator $\hat{\varphi}_{tin}(y) = \big(\mathbf{1}_{\{Y_i \leq y\}} - \hat{F}_{Y|T\Lambda}(y|t, \Lambda_i)\big)\hat{\mathbb{E}}[W|\Lambda = \Lambda_i]/\hat{f}_{T|\Lambda}(t|\Lambda_i)$.

(i) Given the sample, $\hat{F}_{y|t\Lambda_i}$ is monotone increasing in y by construction, so $\big\{f_{ni}(y) \equiv U_i n^{-1/2} h^{d_t/2} K_h(T_i - t) \cdot \big(\hat{\varphi}_{tin}(y) - \varphi_{tin}(y)\big)\big\}$ are manageable. Note $\mathbb{E}f_{ni}(y) = 0$, $\mathbb{E}f_{ni}^2(y) = n^{-1}h^{d_t}K_h^2(T_i - t) \cdot \big(\hat{\varphi}_{tin}(y) - \varphi_{tin}(y)\big)^2$, and $\mathcal{Z}_n(y) = \sum_{i=1}^n f_{ni}(y)$. Assuming $f_{T|\Lambda}(t|\Lambda)$ and W are uniformly bounded away from zero and above, define the envelope $F_{ni} = U_i\sqrt{\frac{h^{d_t}}{n}} K_h(T_i - t)C$.

(ii) $\big|\mathbb{E}\mathcal{Z}_n(y_1)\mathcal{Z}_n(y_2)\big| = \big|\mathbb{E}\sum_{i=1}^n f_{ni}(y_1)f_{ni}(y_2)\big| \leq n^{-1}\sum_{i=1}^n h^{d_t}K_h^2(T_i - t)\big|\hat{\varphi}_{tin}(y_1) - \varphi_{tin}(y_1)\big|\big|\hat{\varphi}_{tin}(y_2) - \varphi_{tin}(y_2)\big| \leq n^{-1}\sum_{i=1}^n h^{d_t}K_h^2(T_i - t) \cdot \big\|\hat{\varphi}_{tin} - \varphi_{tin}\big\|_\infty^2 = O_p(1) \cdot o_p(1) = o_p(1)$.

(iii) and (iv) are the same as the previous calculation for the first dominating term.

(v) $0 \leq \sum_{i=1}^n \mathbb{E}\big[f_{ni}(y_1) - f_{ni}(y_2)\big]^2 \leq n^{-1}\sum_{i=1}^n h^{d_t}K_h^2(T_i - t)\big\|\hat{\varphi}_{tin}(y_1) - \varphi_{tin}(y_1) - \big(\hat{\varphi}_{tin}(y_2) - \varphi_{tin}(y_2)\big)\big\|_\infty^2 \to 0$.

F Proofs of Lemmas for Stochastic Equicontinuity

F.1 Proof of Lemma A.1

Define $Z_{ni}(\upsilon) \equiv n^{-1/2}f(y, t, \Lambda_i)W_i$, indexed by $\upsilon \equiv (y, t, f) \in \Upsilon \equiv \mathcal{Y} \times \mathcal{T} \times \mathcal{F}$. The bracketing CLT will imply $\sum_{i=1}^n \big(Z_{ni}(\upsilon) - \mathbb{E}Z_{ni}(\upsilon)\big)$ is asymptotic stochastic equicontinuous in υ with respect to the pseudo-metric $\rho(\upsilon_1, \upsilon_2) = \max\{|y_1 - y_2|, \|t_1 - t_2\|, \|f_1 - f_2\|_\infty\}$. It suffices to check the conditions for Theorem 2.11.9 in van der Vaart and Wellner (1996):

(i) Since the functions are assumed to be uniformly bounded above and below, $\mathbf{1}_{\{\|Z_{ni}\|_\Upsilon > \eta\}} = 0$ for n large enough. So for any $\eta > 0$, $\sum_{i=1}^n \mathbb{E}\big[\|Z_{ni}\|_\Upsilon \mathbf{1}_{\{\|Z_{ni}\|_\Upsilon > \eta\}}\big] = o_p(1)$.

(ii) It is straightforward to modify Lemma B.2 in Ichimura and Lee (2010) to replace their Lipschitz continuity with Hölder continuity,

$$N(\epsilon_1^{1/2}C_L + \epsilon_2, \mathcal{F}, \|\cdot\|_\infty) \leq N\Big(\epsilon_1, \mathcal{Y} \times \mathcal{T}, |\cdot|\Big) \times \sup_{y \in \mathcal{Y}, t \in \mathcal{T}} N\Big(\epsilon_2, \mathcal{M}, \|\cdot\|_\infty\Big).$$

Since $\mathcal{Y}\times\mathcal{T}$ is a compact set, the result remains.

$$N(\epsilon,\mathcal{F},\|\cdot\|_\infty)\le N\Big(\big(\epsilon/(2C_L)\big)^2,\mathcal{Y}\times\mathcal{T},|\cdot|\Big)\times\sup_{y\in\mathcal{Y},t\in\mathcal{T}}N\Big(\epsilon/2,\mathcal{M},\|\cdot\|_\infty\Big)$$
$$N_{[\,]}(\epsilon,\Upsilon,L_2)\le N\Big(\frac{\epsilon}{2C},\mathcal{Y}\times\mathcal{T},|\cdot|\Big)\times N\Big(\frac{\epsilon}{2C},\mathcal{F},\|\cdot\|_\infty\Big).$$

Therefore, $\int_0^{\delta_n}\sqrt{\log N_{[\,]}(\epsilon,\Gamma,L_2)}d\epsilon\to 0$, $\forall\delta_n\to 0$.

(iii) By the Hölder continuity assumption, for any $\rho(\upsilon_1,\upsilon_2)=o(1)$,

$$\sum_{i=1}^n\mathbb{E}\big(Z_{ni}(\upsilon_1)-Z_{ni}(\upsilon_2)\big)^2=\mathbb{E}\Big(f_1(y_1,t_1,\Lambda)W-f_2(y_2,t_2,\Lambda)W\Big)^2$$
$$=\mathbb{E}\Big[\Big(f_1(y_1,t_1,\Lambda)-f_2(y_1,t_1,\Lambda)+f_2(y_1,t_1,\Lambda)-f_2(y_2,t_2,\Lambda)\Big)^2W^2\Big]=o(1).$$

F.2 Proof of Lemma A.2

For any regressor $V_1=v_1(T,S)$ with $v_1\in\bar{\mathcal{M}}_n$, define the regression estimator

$$\hat{F}_{Y|TXV_1}(y|t,x,u)\equiv\frac{\frac{1}{n}\sum_{i=1}^n\mathbf{1}_{\{Y_i\le y\}}K_h(T_i-t)K_h(X_i-x)K_h(v_1(T_i,S_i)-u)}{\frac{1}{n}\sum_{i=1}^nK_h(T_i-t)K_h(X_i-x)K_h(v_1(T_i,S_i)-u)}\equiv\frac{\hat{g}_1(y,t,x,u)}{\hat{f}_1(t,x,u)}.$$

Denoting $\hat{f}_1\equiv\hat{f}_1(t,x,u)\equiv\hat{f}_{TXV_1}(t,x,u)$ and $f\equiv f_{TXV}(t,x,u)$, linearize $\hat{F}_{Y|TXV_1}(y|t,x,u)=$

$$\frac{\hat{g}_1}{f}+\frac{F_{Y|TXV}(y|t,x,u)}{f}\big(f-\hat{f}_1\big)+\frac{f-\hat{f}_1}{f}\Big(\hat{F}_{Y|TXV_1}(y|t,x,u)-F_{Y|TXV}(y|t,x,u)\Big).\tag{31}$$

The first two terms will dominate the first-order asymptotics of $\hat{F}_{Y|TXV_1}(y|t,x,u)-\hat{F}_{Y|TXV_2}(y|t,x,u)$ and the third term will be collected in a smaller-order term $so2$. That is, $\hat{F}_{Y|TXV_1}(y|t,x,u)-\hat{F}_{Y|TXV_2}(y|t,x,u)=\frac{\hat{g}_1-\hat{g}_2}{f_{TXV}(t,x,u)}+\frac{F_{Y|TXV}(y|t,x,u)}{f_{TXV}(t,x,u)}(\hat{f}_2-\hat{f}_1)+so2$.

By Lemma A.3, uniformly in $v_1,v_2\in\bar{\mathcal{M}}_n$, $\big\|\hat{g}_1-\hat{g}_2\big\|_\infty=$

$$\sup_{y,t,x,u}\big|\mathbb{E}\big[\mathbf{1}_{\{Y\le y\}}K_h(T-t)K_h(X-x)\big(K_h(v_1(T,S)-u)-K_h(v_2(T,S)-u)\big)\big]\big|+O_p(n^{-\kappa_1})$$
$$\le\sup_{y,t,x,u}\mathbb{E}\Big[\Big|\mathbf{1}_{\{Y\le y\}}K_h(T-t)K_h(X-x)\cdot\sum_{l=1}^{d_v}\frac{1}{h_l}k'\Big(\frac{V_{2l}-u_l}{h_l}\Big)\prod_{j=1,j\ne l}^{d_v}\frac{1}{h}k\Big(\frac{V_{2j}-u_j}{h}\Big)\Big|\Big]$$
$$\times\max_{j\in\{1,\dots,d_v\}}\Big\|\frac{v_{1j}-v_{2j}}{h}\Big\|_\infty$$
$$+\sup_{y,t,x,u}\mathbb{E}\Big[\Big|\mathbf{1}_{\{Y\le y\}}K_h(T-t)K_h(X-x)\cdot\frac{1}{2}\Big(\sum_{l=1}^{d_v}\frac{\|k''\|_\infty}{h_l}\prod_{j\ne l}\frac{1}{h_l}k\Big(\frac{V_{2j}-u_j}{h_l}\Big)$$
$$+\sum_{l,m=1,l\ne m}^{d_v}\frac{\|k'\|_\infty}{h_m}\frac{1}{h_l}k'\Big(\frac{V_{2l}-u_l}{h_l}\Big)\prod_{j\ne l,m}\frac{1}{h}k\Big(\frac{V_{2j}-u_j}{h}\Big)\Big)\Big|\Big]\cdot\max_{j\in\{1,\dots,d_v\}}\Big\|\frac{v_{1j}-v_{2j}}{h}\Big\|_\infty^2$$
$$+O_p(n^{-\kappa_1})$$
$$=O_p\Big(\max_{j\in\{1,\dots,d_v\}}\Big\|\frac{v_{1j}-v_{2j}}{h}\Big\|_\infty\Big)+O_p(n^{-\kappa_1})=O_p\big(n^{-(\delta-\eta)_{\min}}+n^{-\kappa_1}\big)\tag{32}$$

given $\|\nabla^2K\|_\infty<\infty$ and $\delta>2\eta$. The mean-value expansion is similar to Theorem 3 in Guerre, Perrigne, and Vuong (2000).

By Lemma A.3 and (32), uniformly in $y \in \mathcal{Y}$, $t \in \mathcal{T}$, $x \in \mathcal{X}$, $u \in \mathcal{V}$, and $v_1, v_2 \in \bar{\mathcal{M}}_n$,

$$\begin{aligned}
&\hat{g}_1 - \hat{g}_2 \\
&= \mathbb{E}\big[F_{Y|TXS}(y|T,X,S)K_h(T-t)K_h(X-x)\big(K_h(v_1(T,S)-u) - K_h(v_2(T,S)-u)\big)\big] + O_p(n^{-\kappa_1}) \\
&= \mathbb{E}\big[F_{Y|TXS}(y|t,X,S)f_{T|XS}(t|X,S)K_h(X-x)\big(K_h(v_1(t,S)-u) - K_h(v_2(t,S)-u)\big)\big] \\
&\quad + O_p(n^{-\kappa_1} + h^r) \\
&\hat{f}_2 - \hat{f}_1 = \mathbb{E}\big[K_h(T-t)K_h(X-x)\big(K_h(v_2(T,S)-u) - K_h(v_1(T,S)-u)\big)\big] + O_p(n^{-\kappa_1}) \\
&= \mathbb{E}\big[f_{T|XS}(t|X,S)K_h(X-x)\big(K_h(v_2(t,S)-u) - K_h(v_1(t,S)-u)\big)\big] + O_p(n^{-\kappa_1} + h^r).
\end{aligned}$$

By Result A.1 and (32), the smaller order term in $\hat{F}_{Y|TXV_1} - \hat{F}_{Y|TXV_2}$ is

$$\begin{aligned}
\big\|so2\big\|_\infty &\le \Big\|\frac{1}{f}(f-\hat{f}_1)\big(\hat{F}_{Y|TXV_1} - F_{Y|TXV}\big)\Big\|_\infty + \Big\|\frac{1}{f}(f-\hat{f}_2)\big(\hat{F}_{Y|TXV_2} - F_{Y|TXV}\big)\Big\|_\infty \\
&= O_p\Big(\Big\|\frac{1}{f}(f-\hat{f}+\hat{f}-\hat{f}_1)\big(\hat{F}_{Y|TXV_1} - \hat{F}_{Y|TXV} + \hat{F}_{Y|TXV} - F_{Y|TXV}\big)\Big\|_\infty\Big) \\
&= O_p\bigg(\Big(\sqrt{\frac{\log n}{nh^{d_2}}} + h^r + \frac{\|v_1 - v_0\|_\infty}{h} + n^{-\kappa_1}\Big)^2\bigg)
\end{aligned}$$

Remark (Continuity assumption with respect to the regressor)
The leading term in $\hat{g}_1$ is

$$\mathbb{E}\Big[\mathbb{E}\big[F_{Y|TS}(y|T,S)K_h(T-t)\big|V_1\big]K_h(V_1-u)\Big] = \mathbb{E}\big[F_{Y|TS}(y|t,S)\big|V_1=u\big]f_{TV_1}(t,u) + O_p(h^r)$$

where $V_1 \equiv v_1(T,S)$ and X is dropped for simplicity. Therefore, the leading term of $\hat{F}_{Y|TV_1}(y|t,u) - \hat{F}_{Y|TV_2}(y|t,u)$ is $\mathbb{E}[A(S)|V_1=u]f_{TV_1}(t,u) - \mathbb{E}[A(S)|V_2=u]f_{TV_2}(t,u)$ where $A(S) \equiv \big(F_{Y|TS}(y|t,S) - F_{Y|TV}(y|t,u)\big)/f_{TV}(t,u)$. Suppose a Lipschitz continuity assumption was further imposed on $\mathbb{E}[A(S)|V_1 = u]f_{V_1}(u)$ with respect to the generated regressor $V_1 = v_1(S)$ for any $v_1 \in \bar{\mathcal{M}}_n$. This assumption of Lipschitz continuity with respect to the regressor is similar to Assumption 4 in MRS12. Then $\hat{F}_{Y|TXV_1}(y|t,x,u) - \hat{F}_{Y|TXV_2}(y|t,x,u)$ is dominated by $O_p(\|v_1 - v_2\|_\infty)$ assuming $\|v_1 - v_2\|_\infty/h = o_p(1)$, instead of $O_p\big(\|v_1 - v_2\|_\infty/h\big)$ assuming $\|v_1 - v_2\|_\infty/h^2 = o_p(1)$. We choose the latter approach to control the remainder terms. The smoothness of the distribution functions with respect to the regressor is unspecified. This is another tradeoff between the smoothness assumption and the estimation accuracy of the generated regressor.

F.3 Proof of Lemma A.3

The proof modifies the proof of Lemma 1 in MRS12. Define $\Delta_i(y,v_1,v_2) \equiv \mathbf{1}_{\{Y_i \le y\}}K_h(T_i - t)\big(K_h(v_1(T_i,S_i) - u) - K_h(v_2(T_i,S_i)-u)\big) - \mathbb{E}\big[\mathbf{1}_{\{Y\le y\}}K_h(T-t)\big(K_h(v_1(T,S)-u) - K_h(v_2(T,S)-u)\big)\big]$. The following observation is useful in the proof (i) $|n^{-1}\sum_{i=1}^n \Delta_i(y,v_1,v_2)| \le Cn^{d_2\eta}\max_j \|v_{1j} - v_{2j}\|_\infty/h$, (ii) $\mathbb{E}\Delta_i(y,v_1,v_2)^2 \le Cn^{d_2\eta}\big(\max_j \|v_{1j} - v_{2j}\|_\infty/h\big)^2$, (iii) $|\Delta_i(y,v_1,v_2)| \le Cn^{d_2\eta}\max_j \|v_{1j} - v_{2j}\|_\infty/h$. The bound (ii) is the key to determine the rate κ_1.

When $\kappa_1 \le (\delta - \eta)$, the results hold from a direct bound. Consider the case $\kappa_1 > (\delta - \eta)$. For $s \ge 0$, let $\bar{\mathcal{M}}^*_{s,n,j}$ be a set of functions chosen such that for each $v_j \in \bar{\mathcal{M}}_{n,j}$, there exists $v_j^* \in \bar{\mathcal{M}}^*_{s,n,j}$ such that $\|v_j - v_j^*\|_\infty \le 2^{-s}n^{-\delta}$. Define $\bar{\mathcal{M}}^*_{s,n} = \bar{\mathcal{M}}^*_{s,n,1} \times ... \times \bar{\mathcal{M}}^*_{s,n,d}$. For $v_1, v_2 \in \bar{\mathcal{M}}_n$, choose $v_1^s, v_2^s \in \bar{\mathcal{M}}^*_{s,n}$ such that $\|v^s_{1,j} - v_{1,j}\|_\infty \le 2^{-s}n^{-\delta}$ and $\|v^s_{2,j} - v_{2,j}\|_\infty \le C2^{-s}n^{-\delta}$ for all $j, s \ge 0$. The functions in $\bar{\mathcal{M}}^*_{s,n,j}$ are the midpoints of a $(2^{-s}n^{-\delta})$-covering of $\bar{\mathcal{M}}_{n,j}$. So the cardinality $\#\bar{\mathcal{M}}^*_{s,n,j}$ is at most $C \cdot \exp\big(\big(2^{-s}n^{-\delta}\big)^{-\beta}n^\xi\big)$.

For any $(y,t,u) \in \mathcal{Y}\times\mathcal{T}\times\mathcal{V}$, consider the chain $\Delta_i(y,v_1,v_2) = \Delta_i(y,v_1^0,v_2^0) - \sum_{s=1}^{G_n}\Delta_i(y,v_1^{s-1},v_1^s) + \sum_{s=1}^{G_n}\Delta_i(y,v_2^{s-1},v_2^s) - \Delta_i(v_1^{G_n},v_1) + \Delta_i(y,v_2^{G_n},v_2)$, where G_n is chosen to be the smallest integer that satisfies $G_n > (1+c_G)(\kappa_1 + d_2\eta - (\delta-\eta))\log n/\log 2$ for a constant $c_G > 0$. So for $l = 1,2$, by (i),

$$T_1 \equiv \Big|\frac{1}{n}\sum_{i=1}^n \Delta_i(y,v_l^{G_n},v_l)\Big| \le C2^{-G_n}n^{d_2\eta-(\delta-\eta)} \le Cn^{-\kappa_1}.$$

For any $a > c_G$, define the constant $c_a = (\sum_{s=1}^\infty 2^{-as})^{-1}$.

$$\begin{aligned}
&Pr\Big(\sup_{v_1\in\bar{\mathcal{M}}_n}\big|\frac{1}{n}\sum_{i=1}^n\sum_{s=1}^{G_n}\Delta_i(y,v_1^{s-1},v_1^s)\big| > n^{-\kappa_1}\Big)\\
&\le \sum_{s=1}^{G_n}\sum_{\bar{\mathcal{M}}^*_{s,n}}\sum_{\bar{\mathcal{M}}^*_{s-1,n}} Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^n\Delta_i(y,v_1^{s-1},v_1^s)\Big| > c_a2^{-as}n^{-\kappa_1}\Big) \qquad (33)\\
&\le \sum_{s=1}^{G_n}\#\bar{\mathcal{M}}^*_{s-1,n}\#\bar{\mathcal{M}}^*_{s,n}Pr\Big(\frac{1}{n}\sum_{i=1}^n\Delta_i(y,v_1^{*,s},v_1^{**,s}) > c_a2^{-as}n^{-\kappa_1}\Big) \qquad (34)\\
&\quad + \sum_{s=1}^{G_n}\#\bar{\mathcal{M}}^*_{s-1,n}\#\bar{\mathcal{M}}^*_{s,n}Pr\Big(\frac{1}{n}\sum_{i=1}^n\Delta_i(y,\tilde{v}_1^{*,s},\tilde{v}_1^{**,s}) < -c_a2^{-as}n^{-\kappa_1}\Big) \equiv T_2+T_3. \qquad (35)
\end{aligned}$$

In (34) and (35), the functions $v_1^{*,s},\tilde{v}_1^{*,s} \in \bar{\mathcal{M}}^*_{s-1,n}$ and $v_1^{**,s},\tilde{v}_1^{**,s} \in \bar{\mathcal{M}}^*_{s,n}$ are chosen such that

$$\begin{aligned}
Pr\Big(\frac{1}{n}\sum_{i=1}^n\Delta_i(y,v_1^{*,s},v_1^{**,s}) > c_a2^{-as}n^{-\kappa_1}\Big) &= \max_{v_1^{s-1},v_1^s} Pr\Big(\frac{1}{n}\sum_{i=1}^n\Delta_i(y,v_1^{s-1},v_1^s) > c_a2^{-as}n^{-\kappa_1}\Big)\\
Pr\Big(\frac{1}{n}\sum_{i=1}^n\Delta_i(y,\tilde{v}_1^{*,s},\tilde{v}_1^{**,s}) < -c_a2^{-as}n^{-\kappa_1}\Big) &= \max_{v_1^{s-1},v_1^s} Pr\Big(\frac{1}{n}\sum_{i=1}^n\Delta_i(y,v_1^{s-1},v_1^s) < -c_a2^{-as}n^{-\kappa_1}\Big).
\end{aligned}$$

To show T_2 and $T_3 \le \exp(-cn^c)$ converging to zero at an exponential rate, we use the Markov inequality and $\mathbb{E}e^X \le 1+|X|\mathbb{E}X^2 \le 1+C\mathbb{E}X^2 \le \exp(C\mathbb{E}X^2)$ by $\mathbb{E}X = 0$ and $|X| \le C$ for some $C > 0$.

$$\begin{aligned}
\mathbb{E}\Big[\exp\big(\gamma_{n,s}n^{-1}\Delta_i(y,v_1^{**,s},v_1^{**,s})\big)\Big] &\le \exp\big(\gamma_{n,s}^2n^{-2}\mathbb{E}[\Delta_i^2(y,v_1^{**,s},v_1^{**,s})]\big) \qquad (36)\\
&\le \exp\big(C\gamma_{n,s}^2n^{-2}n^{d_2\eta-2(\delta-\eta)}2^{-2s}\big)
\end{aligned}$$

by (ii). To satisfy $|X| \le C$,

$$\begin{aligned}
|X| &= \big|\gamma_{n,s}n^{-1}\Delta_i(y,v_1^{**,s},v_1^{**,s})\big| \le C\gamma_{n,s}n^{-1}n^{d_2\eta-(\delta-\eta)}2^{-s}\\
&\le c_\gamma n^{(\delta-\eta)-\kappa_1}2^{-as+s} \le c_\gamma n^{(c_G-a)(\kappa_1-(\delta-\eta))} \le C
\end{aligned}$$

where the first inequality is by (iii) and the second inequality comes from $\gamma_{n,s}$ chosen below. When $a < 1$, $Cn^{(\delta-\eta)-\kappa_1}2^{-as+s} \le Cn^{(\delta-\eta)-\kappa_1}2^{G_n(1-a)}$. The above inequality holds by the chosen G_n. When $a \ge 1$, the above inequality holds for n large enough. Therefore,

$$\begin{aligned}
T_2 &\le C\sum_{s=1}^{G_n}\exp\Big(d_v(1+2^{-\beta})2^{s\beta}n^{\delta\beta+\xi} - \gamma_{n,s}c_a2^{-as}n^{-\kappa_1} + C\gamma_{n,s}^2n^{-1+d_2\eta-2(\delta-\eta)}2^{-2s}\Big)\\
&= C\sum_{s=1}^{G_n}\exp\Big(d_v(1+2^{-\beta})2^{s\beta}n^{\delta\beta+\xi} - c_\gamma 2^{2(1-a)s}n^{1-2\kappa_1-d_2\eta+2(\delta-\eta)}\Big) \le C\sum_{s=1}^{G_n}\exp(-c^sn^c) \le \exp(-cn^c).
\end{aligned}$$

The equality comes from choosing $\gamma_{n,s} = c_\gamma 2^{-as+2s} n^{-\kappa_1+1-d_2\eta+2(\delta-\eta)}$ so that the last two terms in the first line is of the same order. Then κ_1 is chosen such that the second term dominates. And choose a and c_γ such that the sum of the last two terms is negative. Similarly, $T_3 \le \exp(-cn^c)$.

Because $\bar{\mathcal{M}}^*_{0,n}$ can always be chosen such that it contains only a single element and use (i),

$$T_4 = Pr(\sup_{v_1,v_2\in\bar{\mathcal{M}}_n} \big|\frac{1}{n}\sum_{i=1}^n \Delta_i(y, v_1^0, v_2^0)\big| > n^{-\kappa_1}) \le \exp(-cn^c).$$

Therefore,

$$\sup_{u\in\mathcal{V},t\in\mathcal{T},y\in\mathcal{Y}} Pr\Big(\sup_{v_1,v_2\in\bar{\mathcal{M}}_n} \Big|\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i\le y\}} K_h(T_i-t)\Big(K_h(v_1(T_i,S_i)-u) - K_h(v_2(T_i,S_i)-u)\Big)$$
$$-\mathbb{E}\Big[\mathbf{1}_{\{Y\le y\}} K_h(T-t)\Big(K_h(v_1(T,S)-u) - K_h(v_2(T,S)-u)\Big)\Big]\Big| \ge Cn^{-\kappa_1}\Big) \le \exp(-cn^c). \quad (37)$$

(I) (**uniformity in** y) Next we modify the above procedure for (37) to show a stronger result of uniformity in $y \in \mathcal{Y}$,

$$\sup_{u\in\mathcal{V},t\in\mathcal{T}} Pr\Big(\sup_{\substack{v_1,v_2\in\bar{\mathcal{M}}_n\\ y\in\mathcal{Y}}} \Big|\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i\le y\}} K_h(T_i-t)\Big(K_h(v_1(T_i,S_i)-u) - K_h(v_2(T_i,S_i)-u)\Big)$$
$$-\mathbb{E}\Big[\mathbf{1}_{\{Y\le y\}} K_h(T-t)\Big(K_h(v_1(T,S)-u) - K_h(v_2(T,S)-u)\Big)\Big]\Big| \ge Cn^{-\kappa_1}\Big) \le \exp(-cn^c). \quad (38)$$

In (33), denoting $c_a 2^{-as} n^{-\kappa_1} \equiv C$,

$$Pr\Big(\sup_{y\in\mathcal{Y}} \Big|\frac{1}{n}\sum_{i=1}^n \Delta_i(y, v_1^{s-1}, v_1^s)\Big| > C\Big)$$
$$\le Pr\Big(\sup_{y\in\mathcal{Y}} \frac{1}{n}\sum_{i=1}^n \Delta_i(y, v_1^{s-1}, v_1^s) > C\Big) + Pr\Big(\inf_{y\in\mathcal{Y}} \frac{1}{n}\sum_{i=1}^n \Delta_i(y, v_1^{s-1}, v_1^s) < -C\Big).$$

Because $\mathcal{Y}$ is compact and $n^{-1}\sum_{i=1}^n \Delta_i$ is a piecewise constant function that jumps at observed values of Y only, there exists some y^s_{sup} and y^s_{inf} such that $\sup_{y\in\mathcal{Y}} n^{-1}\sum_{i=1}^n \Delta_i(y, v_1^{s-1}, v_1^s) = n^{-1}\sum_{i=1}^n \Delta_i(y^s_{sup}, v_1^{s-1}, v_1^s)$ and $\inf_{y\in\mathcal{Y}} n^{-1}\sum_{i=1}^n \Delta_i(y, v_1^{s-1}, v_1^s) = n^{-1}\sum_{i=1}^n \Delta_i(y^s_{inf}, v_1^{s-1}, v_1^s)$. The functions $v_1^{*,s}, \tilde{v}_1^{*,s} \in \bar{\mathcal{M}}^*_{s-1,n}$ and $v_1^{**,s}, \tilde{v}_1^{**,s} \in \bar{\mathcal{M}}^*_{s,n}$ are chosen such that in (34) and (35),

$$Pr\Big(\frac{1}{n}\sum_{i=1}^n \Delta_i(y^s_{sup}, v_1^{*,s}, v_1^{**,s}) > c_a 2^{-as} n^{-\kappa_1}\Big) = \max_{v_1^{s-1}, v_1^s} Pr\Big(\frac{1}{n}\sum_{i=1}^n \Delta_i(y^s_{sup}, v_1^{s-1}, v_1^s) > c_a 2^{-as} n^{-\kappa_1}\Big)$$
$$Pr\Big(\frac{1}{n}\sum_{i=1}^n \Delta_i(y^s_{inf}, \tilde{v}_1^{*,s}, \tilde{v}_1^{**,s}) < -c_a 2^{-as} n^{-\kappa_1}\Big) = \max_{v_1^{s-1}, v_1^s} Pr\Big(\frac{1}{n}\sum_{i=1}^n \Delta_i(y^s_{inf}, v_1^{s-1}, v_1^s) < -c_a 2^{-as} n^{-\kappa_1}\Big).$$

Then by this modification, showing (38) follows the same procedure as proving (37).

(II) (**uniformity in** $(t,u) \in \mathcal{T}\times\mathcal{V}$) For $C_t > 0$, choose a grid $\mathcal{T}_n \times \mathcal{V}_n$ with $O(n^{C_t})$ points, such that for each $(t,u) \in \mathcal{T}\times\mathcal{V}$, there exists a grid point $(t^*, u^*) = (t^*(t), u^*(u)) \in \mathcal{T}_n \times \mathcal{V}_n$ such that $\|t - t^*\| \le n^{-cC_t}$ and $\|u - u^*\| \le n^{-cC_t}$.

Define $D_i(y,t,u,v_1) \equiv \mathbf{1}_{\{Y_i \le y\}} K_h(T_i - t) K_h(v_1(T_i,S_i) - u)$. Choosing C_t large enough implies

$$\sup_{\substack{u\in\mathcal{V},t\in\mathcal{T}\\ y\in\mathcal{Y},v_1\in\bar{\mathcal{M}}_n}} \Big|\frac{1}{n}\sum_{i=1}^n D_i(y,t^*,u^*,v_1) - D_i(y,t,u,v_1) - \mathbb{E}\big[D_i(y,t^*,u^*,v_1) - D_i(y,t,u,v_1)\big]\Big| \le Cn^{-cC_t} n^{(d_2+1)\eta} \le n^{-\kappa_1}.$$

The triangle inequality and (38) imply the statement in this lemma.

(Ⅲ) Consider the dependent variable to be $A_\alpha(Y)$ where $\{A_\alpha : \alpha \in \mathcal{A}\}$ is a class of uniformly bounded functions, A_α is Lipschitz continuous in α, and $\mathcal{A}$ is a compact set. The result in (37) holds by replacing the dependent variable $\mathbf{1}_{\{Y\le y\}}$ with $A_\alpha(Y)$.
Next we show the uniformity over $\alpha \in \mathcal{A}$. Define $D_i(\alpha,t,u,v_1) \equiv A_\alpha(Y_i) K_h(T_i - t) K_h(v_1(T_i,S_i) - u)$. Following (Ⅱ), similarly choose a grid $\mathcal{A}_n$ with $O(n^{C_t})$ points such that for each $\alpha \in \mathcal{A}$, there exists a grid point $\alpha^* = \alpha^*(\alpha) \in \mathcal{A}_n$ such that $\|\alpha - \alpha^*\| \le n^{cC_t}$. Then by the Lipschitz continuity of A_α,

$$\sup_{\substack{u\in\mathcal{V},t\in\mathcal{T}\\ \alpha\in\mathcal{A},v_1\in\bar{\mathcal{M}}_n}} \Big|\frac{1}{n}\sum_{i=1}^n D_i(\alpha^*,t^*,u^*,v_1) - D_i(\alpha,t^*,u^*,v_1) - \mathbb{E}\big[D_i(\alpha^*,t^*,u^*,v_1) - D_i(\alpha,t^*,u^*,v_1)\big]\Big| \le C|\alpha^* - \alpha| n^{d_2\eta} \le Cn^{-cC_t} n^{d_2\eta} \le n^{-\kappa_1}.$$

We therefore prove

$$\sup_{\substack{u\in\mathcal{V},t\in\mathcal{T}\\ \alpha\in\mathcal{A},v_1,v_2\in\bar{\mathcal{M}}_n}} \Big|\frac{1}{n}\sum_{i=1}^n A_\alpha(Y_i) K_h(T_i - t)\Big(K_h(v_1(T_i,S_i) - u) - K_h(v_2(T_i,S_i) - u)\Big) - \mathbb{E}\Big[A_\alpha(Y) K_h(T - t)\Big(K_h(v_1(T,S) - u) - K_h(v_2(T,S) - u)\Big)\Big]\Big| \le Cn^{-\kappa_1}$$

for n large enough $w.p.a.1$.

F.4 Proof of Lemma A.4

Define $\Delta_i(v_1,v_2) \equiv A(y,t,x,s;W_i,X_i,V_i) K_h(x - X_i)\big(K_h(v_1(t,s) - V_i) - K_h(v_2(t,s) - V_i)\big) - \mathbb{E}_{WXV}\big[A(y,t,x,s;W,X,V) K_h(x - X)\big(K_h(v_1(t,s) - V) - K_h(v_2(t,s) - V)\big)\big]$. First we observe (i) $|n^{-1}\sum_{i=1}^n \Delta_i(v_1,v_2)| \le Cn^{(d_2-d_t)\eta} \max_j \|v_{1j} - v_{2j}\|_\infty / h$, (ii) $\mathbb{E}\Delta_i(v_1,v_2)^2 \le Cn^{(d_2-d_t)\eta}\big(\max_j \|v_{1j} - v_{2j}\|_\infty / h\big)^2$, and (iii) $|\Delta_i(v_1,v_2)| \le Cn^{(d_2-d_t)\eta} \max_j \|v_{1j} - v_{2j}\|_\infty / h$, $w.p.a.1$. Note that the above bounds (i), (ii), and (iii) are implied by those in the proof of Lemma A.3. A weaker bound in (ii) results in a larger bound for κ_{11} compared with κ_1, because κ_1 is determined by (ii) in (36). Implement the same chaining argument by choosing G_n to be the smallest integer that satisfies $G_n > (1 + c_G)(\kappa_1 + (d_2 - d_t)\eta - (\delta - \eta)) \log n / \log 2$ for a constant $c_G > 0$. So that for $l = 1, 2$, by (i), uniformly in $y \in \mathcal{Y}$, $T_1 \equiv \big|n^{-1}\sum_{i=1}^n \Delta_i(v_l^{G_n}, v_l)\big| \le C2^{-G_n} n^{-\delta} \le Cn^{-\kappa_1}$. By the same procedure in the proof of Lemma A.3,

$$\sup_{\substack{x\in\mathcal{X},s\in\mathcal{S}\\ y\in\mathcal{Y},t\in\mathcal{T}}} Pr\Big(\sup_{v_1,v_2\in\bar{\mathcal{M}}_n} \Big|\frac{1}{n}\sum_{i=1}^n A(y,t,x,s;W_i,X_i,V_i) K_h(x - X_i)\big(K_h(v_1(t,s) - V_i) - K_h(v_2(t,s) - V_i)\big) - \mathbb{E}_{WXV}\Big[A(y,t,x,s;W,X,V) K_h(x - X)\big(K_h(v_1(t,s) - V) - K_h(v_2(t,s) - V)\big)\Big]\Big| \ge Cn^{-\kappa_1}\Big) \le \exp(-cn^c).$$

For uniformity in $(y,x,s,t) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{S} \times \mathcal{T}$, for $C > 0$, choose a grid $\mathcal{Y}_n \times \mathcal{X}_n \times \mathcal{S}_n \times \mathcal{T}_n$ with

$O(n^C)$ points, such that for each $(y,x,s,t) \in \mathcal{Y}\times\mathcal{X}\times\mathcal{S}\times\mathcal{T}$. There exists a grid point $(y^*,x^*,s^*,t^*) \in \mathcal{Y}_n\times\mathcal{X}_n\times\mathcal{S}_n\times\mathcal{T}_n$ such that $\|(y,x,s,t)-(y^*,x^*,s^*,t^*)\| \le n^{-cC}$. It suffices to show the following is bounded by $n^{-\kappa_1}$ *w.p.a.*1 uniformly over (y,t,s,v) and $v_1 \in \bar{\mathcal{M}}_n$

$$
\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n A(y,t,x,s,W_i,X_i,V_i)K_h(x-X_i)K_h(v_1(t,s)-V_i) \\
&\qquad\qquad - A(y^*,t^*,x^*,s^*;W_i,X_i,V_i)K_h(x^*-X_i)K_h(v_1(t^*,s^*)-V_i)\Big| \\
&\le \big|\frac{1}{n}\sum_{i=1}^n A(y^*,t^*,x^*,s^*;W_i,X_i,V_i)\big(K_h(x-X_i)K_h(V(t,s)-V_i) \\
&\qquad\qquad - K_h(x^*-X_i)K_h(v_1(t^*,s^*)-V_i)\big)\big| \qquad (39) \\
&+ \big|\frac{1}{n}\sum_{i=1}^n \big(A(y,t,x,s;W_i,X_i,V_i) - A(y^*,t^*,x^*,s^*;W_i,X_i,V_i)\big) \\
&\qquad\qquad \times K_h(v_0(t^*,s^*)-V_i)K_h(x^*-X_i)\big| \qquad (40) \\
&+ \big|\frac{1}{n}\sum_{i=1}^n \big(A(y,t,x,s;W_i,X_i,V_i) - A(y^*,t^*,x^*,s^*;W_i,X_i,V_i)\big) \\
&\qquad\qquad \big(K_h(x-X_i)K_h(v_1(t,s)-V_i) - K_h(x^*-X_i)K_h(V(t^*,s^*)-V_i)\big)\big|.
\end{aligned}
$$

$\sup_{y,t,x,s,v_1}|(39)| \le n^{-\kappa_1}$ for large enough n if C is chosen large enough. $\sup_{y,t,x,s,v_1}|(40)| \le n^{-\kappa_1}$ for large enough n by the smoothness of A. The last term is of smaller order.

F.5 Proof of Lemma A.5

The proof is implied by the proof of Lemma A.3, where $\Delta_i(v_1,v_2) \equiv A(y,t,V_i,W_i)\big(v_1(T_i,S_i) - v_2(T_i,S_i)\big) - \mathbb{E}\big[A(y,t,V,W)\big(v_1(T,S)-v_2(T,S)\big)\big]$ and $A(y,t,V_i,W_i) = \nabla_v F_{Y|TV}(y|t,V_i)W_i$. Note that the following still holds the same as the proof of Lemma A.3: (i) $|n^{-1}\sum_{i=1}^n \Delta_i(v_1,v_2)| \le C\max_j\|v_{1j}-v_{2j}\|_\infty$. (ii) $\mathbb{E}\Delta_i(v_1,v_2)^2 \le C\max_j\|v_{1j}-v_{2j}\|_\infty^2$. (iii) $|\Delta_i(v_1,v_2)| \le C\max_j\|v_{1j}-v_{2j}\|_\infty$. That is, the proof is essentially the same for the case $\eta=0$, $h=1$, and $K_h(x)=x$.