ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

EFFICIENT PROPENSITY SCORE REGRESSION ESTIMATORS OF MULTI-VALUED TREATMENT EFFECTS FOR THE TREATED

Ying-Ying Lee

Number 738
January 2015

Manor Road Building, Manor Road, Oxford OX1 3UQ

Efficient propensity score regression estimators of multi-valued treatment effects for the treated

Ying-Ying Lee *

University of Oxford †

January 5, 2015

Abstract

We study the role of the propensity scores in estimating treatment effects for the treated with a multi-valued treatment. Assume assignment to one of the multiple treatments is random given observed characteristics. Valid causal comparisons for the subpopulation who has been treated a particular treatment level are based on two propensity scores — one for the treated level and one for the counterfactual level. In contrast to the binary treatment case, these two propensity scores do not add up to one. This is the key feature that allows us to distinguish different roles of the propensity scores and to provide new insight in well-known paradoxes in the binary treatment effect and missing data literature: We formally show that knowledge of the propensity score for the *treated level* decreases the semiparametric efficiency bound, *regardless* of knowledge of the propensity score for the counterfactual level. We propose efficient kernel regression estimators that project on a nonparametrically estimated propensity score for the counterfactual level and the true propensity score for the treated level. A surprising result is implied for the binary treatment effect for the treated: when the propensity scores are known, using one estimated propensity score is not efficient. Our efficient estimator regresses on a normalized propensity score that utilizes the information contained in the nonparametrically estimated and the true propensity scores.

Keywords: propensity score, multi-valued treatment, semiparametric efficiency bound, unconfoundedness, generated regressor

*I thank Bryan Gramham for suggesting the idea. I thank Jack Porter and Keisuke Hirano for helpful comments and discussion. I also thank seminar participants in Academia Sinica and 2014 Annual Meeting of the Midwest Econometrics Group. All 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/

1 Introduction

Matching is a widely-used program evaluation estimation method in the economics and statistics literatures. Instead of matching on a rich set of observed characteristics, Rosenbaum and Rubin (1983) propose a method based on the propensity score, defined as the conditional probability of receiving a treatment given the observables. This paper studies the role of the propensity score in estimating treatment effects *for the treated* with a multi-valued treatment. *The treated* is the subpopulation who has received a particular treatment level. The average treatment effect for the treated reveals the change in the average outcome of the treated subpopulation if their treatment is switched from the *treated level* they have received to a *counterfactual level*. In many cases of interest, treatments take on more than two values. For example, participants in active labor market programs often receive different periods or types of training, such as wage subsidy, vocational training classes, or apprenticeships with local employers. Policy makers might be interested in what the average wage for the subsidy recipients would have been if they counterfactually had participated in training classes or served some apprenticeships (Lechner, 2002; Frölich, 2004b; Cattaneo, 2010, for example).

We examine the role of propensity scores in multi-valued treatment effects for the treated from three perspectives: identification, semiparametric efficiency, and nonparametric regression estimation. Our results formally generalize the binary treatment literature and provide new insight in well-known paradoxes. The multi-valued treatment setup enables the investigation of different roles played by *the propensity score for the treated level* and *the propensity score for the counterfactual level*, which are the conditional probabilities of being treated at the treated level and at the counterfactual level, respectively. In the binary treatment case, the propensity score of the treatment group and that of the control group always add up to one. The multi-valued treatment framework provides a general viewpoint to circumvent the inconvenient fact that the two propensity scores of a binary treatment are degenerated to one propensity score. There are three main findings:

(i) (*Identification*) Theorem 1 provides *equivalent* representations of weak unconfoundedness in terms of the propensity scores. By the equivalent representation, we can use the propensity scores to construct the coarsest strata where a causal comparison between two potential outcomes is valid.

(ii) (*Semiparametric efficiency bound*) Knowledge of *the propensity score for the counterfactual level* has no effect on the semiparametric efficiency bounds of both the treatment effects for the treated and for the population. Knowledge of *the propensity score for the treated level* decreases the semiparametric efficiency bound.

(iii) (*Regression estimator*) When the propensity scores are known, the efficient propensity

score regression estimator uses the nonparametrically estimated propensity score for the counterfactual treatment level and the true propensity score for the treated level as generated regressors. The nonparametric propensity score estimator recovers the information of pretreatment variables on the outcome distribution that is not captured by the propensity score.

We consider a multi-valued treatment variable T taking values on a finite discrete set. Following the Roy (1951)-Rubin (1974) model, assume there exists a set of potential outcomes $Y(t)$ corresponding to each treatment level t. We only observe one of the potential outcomes $Y_i = Y_i(t)$ if $T_i = t$ for each individual i. Other potential outcomes $Y_i(t')$ for $t' \neq t$ are unobserved. [1] The object of interest is the *average treatment effect for the treated* (ATT) $\mathbb{E}[Y(t) - Y(t')|T = t']$, the average causal effect of exposing the subpopulation who has received treatment level t' to a counterfactual level t. Assume unconfoundedness: the treatment variable is independent of the potential outcomes given pretreatment variables X. Denote the *propensity score* for treatment level t to be $P_t(X) \equiv Prob(T = t|X)$.

The result (i) implies that adjusting for two propensity scores $(P_t(X), P_{t'}(X))$ or a normalized one $P_t(X)/(P_t(X) + P_{t'}(X))$ removes all biases associated with differences in the pretreatment variables. The identification is first shown by Lechner (2001):

$$\begin{aligned}\mathbb{E}[Y(t)|T = t'] &= \mathbb{E}\big[\mathbb{E}[Y|T = t, P_t(X), P_{t'}(X)]\big|T = t'\big] \\ &= \mathbb{E}\left[\mathbb{E}\left[Y|T = t, \frac{P_t(X)}{P_t(X) + P_{t'}(X)}\right]\middle|T = t'\right]. \qquad (1)\end{aligned}$$

Theorem 1 further generalizes the propensity score methodology in Rosenbaum and Rubin (1983) to a multi-valued treatment and extends Imbens (2000) and Lechner (2001). Our *equivalent* representation is stronger than existing results in the sense that the necessary and sufficient conditions provide the coarsest strata where the causal inference is valid. One of the important motivation of the propensity score method is to reduce the dimension of adjusting for pretreatment variables. So Theorem 1 forms a basis for matching, stratification, or subclassification estimation and can be applied to the extensive literature of Rosenbaum and Rubin (1983), Heckman, Ichimura, and Todd (1998), Dehejia and Wahba (2002), Lechner (2002), Imai and van Dyk (2004), Frölich (2004a), Abadie and Imbens (2012), among others. We focus on nonparametric regression estimation introduced by Heckman, Ichimura, and Todd (1998).

The proposed efficient regression estimators based on the identification (1) consists of three steps: In the first step, the propensity scores $(P_t(X), P_{t'}(X))$ and $P_t(X)/(P_t(X) + P_{t'}(X))$ are estimated nonparametrically as generated regressors. The second step is a non-

[1] The handbook chapter by Heckman, LaLonde, and Smith (1999), Heckman and Vytlacil (2007), and Imbens and Wooldridge (2009) provide comprehensive review and discussions on the program evaluation literature.

parametric kernel regression of the outcome variable given the generated regressors for the subgroup who has received the counterfactual level t. The third step is a sample analog that sums out the propensity scores. We derive the limiting properties using the result of partial means with generated regressors in Lee (2014). Our propensity score regression estimator for a binary case inherits the matching estimator introduced by Heckman, Ichimura, and Todd (1998), where the generated regressor is only one propensity score $\hat{P}_1(X)$,

The result (ii) suggests that we should make use of the true propensity score for the treated level $P_{t'}(X)$ to improve the asymptotic precision. The result (iii) suggests to use the nonparametrically estimated propensity score for the counterfactual level t $\hat{P}_t(X)$, although its true function $P_t(X)$ is known. That is, when the propensity scores are known, our efficient regression estimators project on the generated regressors $(\hat{P}_t(X), P_{t'}(X))$ or the normalized one $\hat{P}_t(X)/(\hat{P}_t(X) + P_{t'}(X))$. The general insight to construct an efficient estimator is to utilize knowledge of $P_{t'}(X)$ and use a nonparametrically estimated $\hat{P}_t(X)$ to recover the information of the pretreatment variables on the outcome that is lost in projection on the propensity scores.

We contribute an efficient regression estimator for $\mathbb{E}[Y(1)|T = 0]$ that projects on the normalized propensity score $\hat{P}_1(X)/(\hat{P}_1(X) + P_0(X))$. It might be surprising that when the propensity score is known, solely regressing on one nonparametrically estimated propensity score $\hat{P}_1(X)$ as in Heckman, Ichimura, and Todd (1998) is not efficient. Intuitively, this is because $\hat{P}_1(X) + \hat{P}_0(X) = 1$. Regressing on $\hat{P}_1(X)$ is equivalent to regressing on $\hat{P}_1(X)/(\hat{P}_1(X) + \hat{P}_0(X))$, which is less efficient than regressing on $\hat{P}_1(X)/(\hat{P}_1(X) + P_0(X))$. Our result responds to Hahn (1998) and Heckman, Ichimura, and Todd (1998) on the question: *Is it better to match on $P_1(X)$ or X if you know $P_1(X)$?* It has long been a paradox in the literature that using the estimated propensity score is more efficient than using the true one (Hahn, 1998; Imai and van Dyk, 2004; Hirano, Imbens, and Ridder, 2003; Graham, 2011; Abadie and Imbens, 2012). The binary treatment literature has discussed intuition or examples to understand the paradox, mostly from a view of GMM or inverse propensity score weighting estimation. We offer a new theoretical explanation from a view of regression estimation: the nonparametrically estimated propensity score recovers the relationship between the pretreatment variables and the outcome.

We contribute to the program evaluation literature by providing an efficient propensity score regression estimator for the ATT. The program evaluation estimators that involve estimated propensity scores as regressors or matching variables are less developed than alternative estimators, such as the propensity score weighting estimators in Hirano, Imbens, and Ridder (2003). Abadie and Imbens (2012) derive the limiting distribution of propensity score matching estimators when the propensity score is estimated parametrically in a first step. There is a recent literature of nonparametric regression with generated regressor. Hahn and

Ridder (2013) show that the propensity score regression estimators of Heckman, Ichimura, and Todd (1998) are efficient when the propensity is not known. Complementary to the theoretical finding in Hahn and Ridder (2013), Mammen, Rothe, and Schienle (2014), and Lee (2014) provide concrete estimators and the limit theory. In this paper, we work on a general multi-valued treatment framework and further consider the efficient estimators when the propensity scores are known.

The multi-valued treatment effects for the population have been studied in Imbens (2000); Lechner (2001); Imai and van Dyk (2004); Frölich (2004b); Cattaneo (2010); among others. Because the treated subpopulation is defined by a treatment level, identification and estimation of the causal effect for the treated requires multiple treatment indicators and propensity scores for the relevant levels. Cattaneo (2010) calculates the semiparametric efficiency bound for the causal effects for the population. Building on Cattaneo (2010), we calculate the bounds for the cases when the propensity scores are unknown and known, for the treated and for the population.

Our discussion focuses on the mean $\mathbb{E}[Y(t)|T = t']$ for ease of exposition, but our results are general for distributional features of $Y(t)$ for the treated with value t'. The identification, semiparametric efficiency bound, and the estimation theory for the quantile treatment effects for the treated are implied. In Section 4.6, by extending the results to the Hadamard-differentiable functionals of the partial mean process, we are able to provide the limiting distribution 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; Donald and Hsu, 2014).

The paper is organized as follows. Section 2 presents identification based on the propensity scores. Section 3 presents the semiparametric efficiency bounds for two cases: unknown and known propensity scores. In Section 4, we introduce the efficient propensity score regression estimators. The Appendix collects the proofs, the notations, and the regularity conditions under which the estimators are root-n consistent and reach the efficiency bounds.

2 Identification

This section discusses identification of multi-valued treatment effects for the treated by an unconfoundedness assumption. Theorem 1 provides equivalent representations of weak unconfoundedness in terms of the propensity scores. The merit of the equivalent representations is to provide both identification of the average causal effects and the coarsest strata for causal comparison. We study direct causal comparison for a multi-valued treatment in Section 2.1. In Section 2.2, we compare our results with the binary treatment effects and the multi-valued

treatment effects for the population.

Let the support of the multi-valued treatment variable T be $\mathcal{T} = \{1, 2, ..., J\}$ a finite discrete set with some fixed positive integer J. Let D_t be the indicator of treatment assignment to level t, i.e., if $T = t$, $D_t = 1$; otherwise, $D_t = 0$. The treatment assignment is mutually exclusive to multiple states in $\mathcal{T}$, so $\sum_{t\in\mathcal{T}} D_t = 1$ and $T = \sum_{t\in\mathcal{T}} t \times D_t$. We only observe one of the potential outcomes $Y_i = \sum_{t\in\mathcal{T}} Y_i(t) \times D_t$ for each individual i. Our dataset consists of $(Y_i, X_i, T_i, D_{1i}, ..., D_{Ji})$ for $i = 1, ..., n$. The key assumption maintained throughout this paper is that treatment assignment is exogenous conditional on pretreatment variables X or unconfoundedness, also known as selection on observables or conditional independence assumption. The symbols $\perp$ and $|$ denote statistical independence and conditioning respectively.

Assumption 1 (Weak unconfoundedness)
Consider any treatment level $t \in \mathcal{T}$. $T \perp Y(t)|X$; equivalently $D_{t'} \perp Y(t)\big|X$, for all $t' \in \mathcal{T}$.

Under Weak unconfoundedness Assumption 1, we can identify the conditional mean of $Y(t)$ for those who are treated at t',

$$\mathbb{E}[Y(t)|T = t', X] = \mathbb{E}[Y(t)|D_{t'} = 1, X] = \mathbb{E}[Y(t)|D_t = 1, X] = \mathbb{E}[Y|D_t = 1, X]. \qquad (2)$$

To derive the identification by adjusting for X in (2), there is no difference in using the treatment variable T or the equivalent expression based on the treatment indicator $D_{t'}$ in Assumption 1. But to adjust for the propensity scores, it turns out to be essential to express Weak unconfoundedness Assumption 1 in terms of $D_{t'}$. The following Theorem 1 shows that we do not need to adjust for the entire set of the propensity scores. Instead, we only need to adjust for the propensity scores for relevant treatment levels, captured by the treatment indicators.

We need a common support assumption, also known as *overlap*.

Assumption 2 (Overlap)
The propensity score $0 < P_t(x) \equiv Prob(T = t|X = x) < 1$, for all $x \in \mathcal{X}$ and $t \in \mathcal{T}$.

It is commonly assumed *Strong unconfoundedness* $T \perp \{Y(t)\}_{t\in\mathcal{T}}|X$ in the treatment effect literature, for example, Rosenbaum and Rubin (1983), Lechner (2001), Hirano, Imbens, and Ridder (2003). The joint independence of the potential outcomes is stronger than what is needed for our identification results. Our results can be easily modified under Strong unconfoundedness. One of the main results of Rosenbaum and Rubin (1983) is that if a treatment assignment is strongly ignorable — the combination of Strong unconfoundedness and overlap assumptions — given X, then it is strongly ignorable given any balancing score. A balancing score is a function of X that is *finer* than the propensity score in the sense that the propensity score can be expressed as a function of a balancing score. We build on and generalize

Rosenbaum and Rubin (1983)'s result to a multi-valued treatment under weakly ignorable assumption. Theorem 1 provides equivalent representations of Weak unconfoundedness Assumption 1, i.e., the necessary and sufficient conditions, based on the propensity scores. We first define the *normalized propensity score* to be the conditional probability of receiving a treatment level given X when the treatment variable is restricted to take values on a subset of $\mathcal{T}$.

Definition (Normalized propensity score)
For any $\mathcal{S} \subseteq \mathcal{T}$ and $t \in \mathcal{S}$, define the *normalized propensity score*

$$P_{t|\mathcal{S}}(X) \equiv \frac{P_t(X)}{\sum_{s\in\mathcal{S}} P_s(X)} = Prob(T = t|X, T \in \mathcal{S}).$$

Note that when $\mathcal{S} = \mathcal{T}$ or $\mathcal{T} = \{1, 2\}$ for a binary treatment, the normalized propensity score equals the propensity score $P_{t|\mathcal{T}}(X) = P_t(X)$.

Theorem 1 (Weak unconfoundedness - Propensity score)
Suppose Assumption 2 holds. Weak unconfoundedness Assumption 1 has the following equivalent representations: Consider any $t \in \mathcal{T}$.

(a) For all $t' \in \mathcal{T}$ and for all measurable functions $g(X)$,

$$D_{t'} \perp Y(t) \;\Big|\; \Big\{P_{t'}(X), g(X)\Big\}.$$

(b) For all $\mathcal{S} \subseteq \mathcal{T}$, $t' \in \mathcal{S}$, and for all measurable functions $g(X)$,

$$D_{t'} \perp Y(t) \;\Big|\; \Big\{P_{t'|\mathcal{S}}(X),\ g(X),\ \sum_{s\in\mathcal{S}} D_s = 1\Big\}.$$

Now we discuss two important applications of Theorem 1. The first is identification of the ATT by adjusting for the propensity scores. Theorem 1(a) implies that for $s \in \{t, t'\}$,

$$D_s \perp Y(t) \;\Big|\; \big\{P_{t'}(X), P_t(X)\big\}.$$

To see this result, for $s = t'$, let $g(X) = P_t(X)$ in Theorem 1(a). And for $s = t$, let $g(X) = P_{t'}(X)$. We therefore obtain identification of

$$\mathbb{E}[Y(t)|T = t', P_t(X), P_{t'}(X)] = \mathbb{E}[Y|T = t, P_t(X), P_{t'}(X)].$$

It means that within the subpopulation with the same value of $(P_t(X), P_{t'}(X))$, the average outcome for units with treatment level t is unbiased for the average value of $Y(t)$ for units with treatment level t'.

Theorem 1(b) has a parallel application: for $s \in \mathcal{S} \equiv \{t, t'\}$,

$$D_s \perp Y(t) \mid \left\{P_{t|\{t,t'\}}(X), D_t + D_{t'} = 1\right\}. \tag{3}$$

This is because $P_{t|\{t,t'\}}(X) + P_{t'|\{t,t'\}}(X) = 1$. We therefore obtain identification of

$$\mathbb{E}\left[Y(t)\middle|T = t', P_{t|\{t,t'\}}(X)\right] = \mathbb{E}\left[Y\middle|T = t, P_{t|\{t,t'\}}(X)\right].$$

It means that within the subpopulation with the same value of the normalized propensity score $P_t(X)/(P_t(X)+P_{t'}(X))$, the average outcome for units with treatment level t is unbiased for the average value of $Y(t)$ for units with treatment level t'.

By Overlap Assumption 2, the average potential outcome $Y(t)$ for the treated with level t' is identified

$$\mathbb{E}[Y(t)|T = t'] = \mathbb{E}\big[\mathbb{E}[Y|T = t, P_t(X), P_{t'}(X)]\big|T = t'\big] \tag{4}$$

$$= \mathbb{E}\big[\mathbb{E}[Y|T = t, P_{t|\{t,t'\}}(X)]\big|T = t'\big]. \tag{5}$$

Lechner (2001) obtains (3) and (5) using the subpopulation who has been treated at level t or t'. So the identification problem is reduced to the binary case in Rosenbaum and Rubin (1983).

The second application of Theorem 1 is the coarsest strata constructed by the propensity scores for causal comparison. Theorem 1(a) implies that

> *$\{P_{t'}(X), P_t(X)\}$ is the coarsest conditioning set such that the treatment assignment to level t, t', or any other level in the complement $\{T \notin \{t, t'\}\}$ is weakly unconfounded.*

To see this, Theorem 1(a) implies $D_s \perp Y(t)|\{g(X), P_t(X), P_{t'}(X)\}$ for $s \in \{t, t'\}$ and for any $g(X)$. Let $g(X)$ be a finer function of $\{P_{t'}(X), P_t(X)\}$ in the sense that $\{P_{t'}(X), P_t(X)\} = h(g(X))$ for some function h. So the conditioning set is $\{P_{t'}(X), P_t(X), g(X)\} = \{g(X)\}$.

Similarly, Theorem 1(b) implies that

> *within the subpopulation who has received treatment level t or t', $\left\{\frac{P_t(X)}{P_t(X)+P_{t'}(X)}\right\}$ defines the coarsest conditioning set such that the treatment assignment to level t or t' is weakly unconfounded.*

The normalized propensity score $P_{t|\{t,t'\}}(X)$ defines a coarser strata than the propensity scores $\{P_t(X), P_{t'}(X)\}$. Both strata are valid for causal comparison, but $\{P_{t|\{t,t'\}}(X)\}$ has the advantage to reduce the dimension of the estimation problem to one. The cost is to lose causal inference on the sample $\{T \notin \{t, t'\}\}$. The causal comparison of $Y(t)$ and $Y(t')$ is

valid within the subpopulation defined by $\{P_{t|\{t,t'\}}(X), T \in \{t,t'\}\}$ in the sense that

$$\mathbb{E}\left[Y(t) - Y(t') \middle| P_{t|\{t,t'\}}(X), T \in \{t,t'\}\right] = \mathbb{E}\left[Y \middle| T = t, P_{t|\{t,t'\}}(X)\right] - \mathbb{E}\left[Y \middle| T = t', P_{t|\{t,t'\}}(X)\right].$$

But we cannot make causal inference on the subpopulation defined by $\{P_{t|\{t,t'\}}(X), T \notin \{t,t'\}\}$. For example, suppose we are interested in the average effect of switching from treatment level 1 to 2 for those who are receiving any other treatment levels, $\mathbb{E}[Y(2) - Y(1)|T \notin \{1,2\}]$. In this case, adjusting for $P_{1|\{1,2\}}(X)$ is not sufficient. The identification requires $P_1(X)$ and $P_2(X)$ for $\mathbb{E}[Y(2) - Y(1)|T \notin \{1,2\}] = \mathbb{E}[\mathbb{E}[Y|T = 2, P_1(X), P_2(X)] - \mathbb{E}[Y|T = 1, P_1(X), P_2(X)]|T \notin \{1,2\}]$.

2.1 Causal comparison

The subclassification or stratification method involves direct comparison within each strata. In practice, it can be difficult to define the strata based on high-dimensional pretreatment variables which motivates the propensity score methodology developed by Rosenbaum and Rubin (1983). The following Corollary implied by Theorem 1 is important to provide the coarsest strata to reduce dimension.

Corollary 1
Consider any $t \in \mathcal{T}$. Theorem 1 implies for all subsets of treatment levels $\mathcal{S} \subseteq \mathcal{T}$ and for all $t' \in \mathcal{S}$,

(a) $D_{t'} \perp Y(t) \left| \{P_s(X)\}_{s\in\mathcal{S}}\right.$. The conditioning set defined by the value of $\{P_s(X)\}_{s\in\mathcal{S}}$ is the coarsest strata such that the treatment assignment to a particular level in $\mathcal{S}$ or to any other level in the complement $\{T \notin \mathcal{S}\}$ is weakly unconfounded.

(b) $D_{t'} \perp Y(t) \left| \left\{\{P_{s|\mathcal{S}}(X)\}_{s\in\mathcal{S}^-}, \sum_{s\in\mathcal{S}} D_s = 1\right\}\right.$. Define a subset $\mathcal{S}^- \subset \mathcal{S}$ to contain all the elements in $\mathcal{S}$ but dropping the first one. Since $\sum_{s\in\mathcal{S}} P_{s|\mathcal{S}}(X) = 1$, the conditioning set $\{\{P_{s|\mathcal{S}}(X)\}_{s\in\mathcal{S}}\} = \{\{P_{s|\mathcal{S}}(X)\}_{s\in\mathcal{S}^-}\}$. Then within the subpopulation who has received treatment in $\mathcal{S}$, $\left\{\{P_{s|\mathcal{S}}(X)\}_{s\in\mathcal{S}^-}\right\}$ is the coarsest strata such that the treatment assignment to a particular level in $\mathcal{S}$ is weakly unconfounded.

An intuition of Corollary 1 is that when more treatment levels are concerned, we need a finer conditioning set by including more propensity scores. An extreme example is when $\mathcal{S} = \mathcal{T}$,

$$T \perp Y(t) \left| \{P_s(X)\}_{s\in\mathcal{T}}\right..$$

This is in line with Imbens (2000) and Lechner (2001) who point out that there is in general no scalar function of the covariates such that the treatment variable is independent of the set

of potential outcomes under Strong unconfoundedness. This is also related to our previous discussion on expressing Weak unconfoundedness Assumption 1 in terms of the treatment indicator $D_{t'}$ instead of the multi-valued treatment variable T. Using $D_{t'}$ helps us focus on the relevant treatment levels of interest and reduce the dimension of the conditioning set. Corollary 1 implies the binary case in Rosenbaum and Rubin (1983): $D_t \perp Y(t)|P_t(X)$.

We illustrate the roles of the propensity score in direct causal comparison and in estimation by the following example. Suppose we are interested in the causal effect of switching the treatment level from 1 to 2 for the subpopulation with treatment level 3, $\mathbb{E}[Y(2)-Y(1)|T=3]$. The set of treatment levels of interest is $\mathcal{S} \equiv \{1,2,3\} \subset \mathcal{T}$. Corollary 1 implies that the subset defined by the value of $(P_{2|\mathcal{S}}(X), P_{3|\mathcal{S}}(X))$ is the coarsest strata where we can make valid causal comparison for the subpopulation receiving treatment level 3: $\mathbb{E}[Y(2)-Y(1)|T=3, P_{2|\mathcal{S}}(X), P_{3|\mathcal{S}}(X)] = \mathbb{E}[Y|T=2, P_{2|\mathcal{S}}(X), P_{3|\mathcal{S}}(X)] - \mathbb{E}[Y|T=1, P_{2|\mathcal{S}}(X), P_{3|\mathcal{S}}(X)]$. Alternatively, we can use $\mathbb{E}[Y(2)-Y(1)|T=3, P_1(X), P_2(X), P_3(X)]$ in Corollary 1(a). But, for estimation, we only need to use one normalized propensity score or two propensity scores as in (4) and (5) for $\mathbb{E}[Y(2)|T=3]$ and $\mathbb{E}[Y(1)|T=3]$ separately. This is the insight made by Imbens (2000) that for estimation, it is not necessary to divide the population into subpopulations where causal comparisons are valid. For causal comparison, we need all the propensity scores of the treatment levels of interest.

2.2 The binary case

Because the multi-valued treatment variable takes on more than two values, Weak unconfoundedness Assumption 1 implies the following weaker condition.

Assumption 3 (Weaker uncondoundedness)
For any $t \in \mathcal{T}$, $D_t \perp Y(t)|X$.

When the treatment is binary, Weak unconfoundedness is equivalent to Weaker unconfoundedness Assumption 3, also known as missing at random assumption in the missing data literature.

Weaker unconfoundedness only suffices to identify the conditional mean of $Y(t)$ for those who are *not treated* at level t, i.e., $\mathbb{E}[Y(t)|T \neq t, X] = \mathbb{E}[Y(t)|D_t = 0, X] = \mathbb{E}[Y(t)|D_t = 1, X] = \mathbb{E}[Y|D_t = 1, X]$. No information on any particular treated subpopulation $\{T = t'\}$ can be recovered, in contrast to equation (2). But Weaker unconfoundedness Assumption 3 and Overlap Assumption 2 suffice to identify the population average treatment effect (ATE) $\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y|D_t = 1, X]]$. For a multi-valued treatment, we can view the subpopulation not receiving t' $\{T \neq t\}$ as the control group in the binary treatment case. This explains why for the ATE of a multi-valued treatment, we only need Weaker unconfoundedness Assump-

tion 3 as in Imbens (2000) and Cattaneo (2010). [2] But for the ATT, we need a stronger *Weak unconfoundedness* Assumption 1 to identify the treated group with value t', rather than everyone else with treatment value different from t'. This simple observation provides intuition for identification based on the propensity scores.

We discuss different roles played by the two propensity scores $(P_t(X), P_{t'}(X))$ by the auxiliary sample and primary sample defined in Chen, Hong, and Tarozzi (2008),

(i) The *auxiliary sample* is the subpopulation with treatment level t, $\{T = t\}$ whose $Y(t)$ is observed.

(ii) The *primary sample* is from the subpopulation of interest, where we aim to infer the causal effects and $Y(t)$ might not be observed.

For the ATT, the primary sample is the subpopulation with treatment level t', $\{T = t'\}$. The identification of the ATT uses the observations with treatment level t as the *auxiliary* sample to recover the average $Y(t)$ for the *primary* sample.

The general insight is that we need the propensity scores of both the auxiliary sample and the primary sample to achieve identification. For the ATE, the primary sample is the entire population. So $\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y|T = t, X]] = \mathbb{E}[\mathbb{E}[Y|T = t, P_t(X)]]$. The treatment group $\{T = t\}$ and the control group $\{T \neq t\}$ are exactly the same as the binary treatment case where two propensity scores add up to unity. So the propensity score of the control group is redundant given the propensity score of the auxiliary sample. Therefore, we only need to adjust for one propensity score for the following conditional causal objects.

$$\mathbb{E}[Y(t)|T \neq t] = \mathbb{E}\left[\mathbb{E}[Y|T = t, X]\ \frac{1 - P_t(X))}{Prob(T \neq t)}\right] = \mathbb{E}\left[\mathbb{E}\big[Y|T = t, P_t(X)]\ \frac{1 - P_t(X)}{Prob(T \neq t)}\right].$$

For $\mathcal{T} = \{0, 1\}$,

$$\mathbb{E}[Y(0)|T = 1] = \mathbb{E}\left[\mathbb{E}\big[Y|T = 0, X]\ \frac{1 - P_1(X)}{Prob(T = 0)}\right] = \mathbb{E}\left[\mathbb{E}\big[Y|T = 0, P_1(X)]\ \frac{1 - P_1(X)}{Prob(T = 0)}\right].$$

This is in line with our discussion on Weaker unconfoundedness. We should view the propensity score of the auxiliary sample $P_t(X)$ playing two roles in estimating $\mathbb{E}[Y(t)|T \neq t]$ and the ATT $\mathbb{E}[Y(1) - Y(0)|T = 1]$: one is for the auxiliary sample and one is for the primary sample. But for the ATE $\mathbb{E}[Y(t)]$, $P_t(X)$ is solely for the auxiliary sample and to reduce dimension. Hahn (1998) makes this point for a binary treatment from an efficiency point of view. The above discussion offers a view of identification that is hidden for a binary treatment.

[2] Imbens (2000) and Cattaneo (2010) refer to Assumption 3 as "Weak unconfoundedness." In the model of multi-valued treatments, we point out that it is "weaker" than the conventional Weak Unconfoundedness Assumption 1.

3 Semiparametric efficiency bounds

Consider a general treated subpopulation $\{T \in \mathcal{S}\}$ defined by a set of treatment levels of interest $\mathcal{S} \subseteq \mathcal{T}$. We calculate the semiparametric efficiency bound for estimating $\mathbb{E}[Y(t)|T \in \mathcal{S}]$. The advantage of working on the general treated subpopulation is that it encompasses both ATT and ATE. For the ATT $\mathbb{E}[Y(t)|T = t']$, the primary sample is the treated subpopulation with level t' by $\mathcal{S} = \{t'\}$. For the ATE $\mathbb{E}[Y(t)]$, the primary sample is the entire population by $\mathcal{S} = \mathcal{T}$. Another interesting example is the subpopulation defined by a set of levels $\mathcal{S} = \{1, 2, 3\} \subset \mathcal{T} = \{1, 2, 3, 4\}$. Researchers might be interested in the ATT defined by $\mathbb{E}[Y(1) - Y(2)|T \in \{1, 2, 3\}]$. Our results suggest that when the primary sample is the entire population, e.g., $\mathbb{E}[Y(t)]$ for the ATE, the semiparametric efficiency bound is the same whether the propensity scores are known or not. When $\mathcal{S}$ is a strict subset of $\mathcal{T}$ such that $Prob(T \in \mathcal{S}|X) = \sum_{t' \in \mathcal{S}} P_{t'}(X) < 1$, knowledge of the propensity score reduces the semiparametric efficiency bound of $\mathbb{E}[Y(t)|T \in \mathcal{S}]$. Another contribution is that we formally show the propensity score for the counterfactual level $P_t(X)$ has no effect on the efficiency bound of $\mathbb{E}[Y(t)|T \in \mathcal{S}]$ for any $\mathcal{S} \subseteq \mathcal{T}$, i.e., for both the ATE and ATT. This result is obscured in the binary treatment setup, where the propensity scores add up to one.

The calculation of the semiparametric efficiency bound is standard in the literature. We follow the setup of Cattaneo (2010) who calculates the semiparametric efficiency bound of the multi-valued treatment effects for the population. The parameter of interest is the distributional feature of the potential outcome $Y(t)$ for the subpopulation $\{T \in \mathcal{S}\}$, denoted by β_t. We define β_t via a generic function $m : \mathcal{Y} \times \mathcal{B} \mapsto \mathbb{R}^{d_m}$, where the parameter space $\mathcal{B} \subset \mathbb{R}^{d_\beta}$ and $d_m \geq d_\beta$. We assume that the moment condition $\mathbb{E}[m(Y(t); \beta_t)|T \in \mathcal{S}] = 0$ uniquely defines the parameter β_t. For example, the parameter is the mean $\beta_t = \mathbb{E}[Y(t)|T \in \mathcal{S}]$ by setting $m(y; \beta) = y - \beta$. Under Assumptions 1, 2 and using equation (2), β_t satisfies

$$\mathbb{E}\left[\mathbb{E}\big[m(Y; \beta_t)\big|T = t, X\big] \, \frac{Prob(T \in \mathcal{S}|X)}{Prob(T \in \mathcal{S})}\right] = 0.$$

We focus our discussion on the treated subpopulation with a single level t', i.e., $\mathcal{S} = \{t'\}$. For the mean when $m(y; \beta) = y - \beta$, the parameter $\beta_t = \mathbb{E}[Y(t)|T = t'] = \int \mathbb{E}[Y|X = x, T = t] dF_{X|T}(x|t')$. When $m(y; \beta) = \mathbf{1}_{\{y \leq \beta\}} - \tau$, the parameter β_t is the τth quantile of the conditional distribution $F_{Y(t)|T}(\cdot|t')$ that satisfies $\int F_{Y|TX}(\beta_t|t, x) dF_{X|T}(x|t') = \tau$.

Before we present the semiparametric efficiency bound in Theorem 2 below, we introduce the main component of the *efficient influence function*

$$\psi_t(Y, T, X) \equiv \frac{D_t}{P_t(X)}\Big(m(Y; \beta_t) - e_t(X; \beta_t)\Big)\frac{\sum_{t' \in \mathcal{S}} P_{t'}(X)}{p_{\mathcal{S}}} + e_t(X; \beta_t)\frac{\sum_{t' \in \mathcal{S}} D_{t'}}{p_{\mathcal{S}}} \qquad (6)$$

where $e_t(x;\beta_t) \equiv \mathbb{E}\big[m(Y;\beta_t)\big|T=t, X=x\big]$ and $p_{\mathcal{S}} \equiv Prob(T \in \mathcal{S})$. For the population parameter $\mathcal{S} = \mathcal{T}$, Theorem 2 coincides with the result in Cattaneo (2010) by $\sum_{t' \in \mathcal{S}} D_{t'} = 1$, $p_{\mathcal{S}} = 1$, and $\sum_{t' \in \mathcal{S}} P_{t'}(X) = 1$. Our calculation in the proof shows that the efficient influence function for the population parameter is reduced from (6) by the following term,

$$\left(\frac{\sum_{t' \in \mathcal{S}} D_{t'}}{\sum_{t' \in \mathcal{S}} P_{t'}(X)} - 1 \right) e_t(X;\beta_t) \frac{\sum_{t' \in \mathcal{S}} P_{t'}(X)}{p_{\mathcal{S}}}. \tag{7}$$

For the population treatment effects, this term (7) is zero. For the treatment effects for the treated $\mathcal{S} \subset \mathcal{T}$, the propensity scores for the treated levels $\{P_{t'}(X)\}_{t' \in \mathcal{S}}$ contribute this term (7) in the efficient influence function (6). It comes from the treated observations with a treatment level in $\mathcal{S}$, i.e., the primary sample. We show below in Theorem 3 that this term (7) disappears when the propensity scores are known.

Now we formally present the theorems for the semiparametric efficiency bounds. We consider the object of interest to be a $J \times 1$ vector $\underline{\beta} \equiv (\beta_1, ..., \beta_J)^\top$. Define $\psi(Y,T,X)$ to be a $J \times 1$ vector whose t-th component is $\psi_t(Y,T,X)$ defined in (6). The following assumption ensures the existence of the bound.

Assumption 4
For all $t \in \mathcal{T}$:

(i) The propensity score $P_t(x)$ is bounded away from zero and one.

(ii) $\mathbb{E}\big[m(Y(t);\beta_t)^2\big|T \in \mathcal{S}\big] < \infty$ and $\mathbb{E}\big[m(Y(t);\beta)\big|T \in \mathcal{S}\big]$ is differentiable in $\beta \in \mathcal{B}$ at β_t.

(iii) Define the gradient matrix

$$\Gamma_* \equiv \begin{bmatrix} \Gamma_1 & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \Gamma_2 & ... & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & ... & \Gamma_J \end{bmatrix}, \text{where } \Gamma_t \equiv \frac{\partial}{\partial \beta^\top} \mathbb{E}\Big[\mathbb{E}[m(Y;\beta)|T=t,X]\Big|T \in \mathcal{S}\Big]\Big|_{\beta=\beta_t}$$

and $\mathbf{0}$ is a $d_m \times d_\beta$ matrix of zeros. The rank of Γ_* is $d_\beta J$.

Theorem 2 (Efficiency - Unknown propensity score)
Suppose Assumptions 1 and 4 hold. Then the efficient influence function of $\underline{\beta}$ is given by $\Psi = -\big(\Gamma_*^\top V_*^{-1} \Gamma_*\big)^{-1} \Gamma_*^\top V_*^{-1} \psi$ where $V_* = var[\psi]$. The semiparametric efficiency bound for any regular estimator of $\underline{\beta}$ is given by $V^* = \big(\Gamma_*^\top V_*^{-1} \Gamma_*\big)^{-1}$.

Theorem 2 encompasses the binary treatment literature: ATT in Hahn (1998) and Hirano, Imbens, and Ridder (2003), the quantile treatment effects in Firpo (2007), distributional effects in Firpo and Pinto (2011), and the missing data model in Chen, Hong, and Tarozzi (2008).

Theorem 3 (Efficiency - Known propensity score)
Suppose the propensity scores $\{P_t(X)\}_{t\in\mathcal{T}}$ *are known. Suppose Assumptions 1 and 4 holds. Then the efficient influence function of* $\underline{\beta}$ *is given by* $\Psi^{PS} = -\left(\Gamma_*' V_*^{-1}\Gamma_*\right)^{-1}\Gamma_*' V_*^{-1}\psi^{PS}$ *where* $\psi^{PS}(Y,T,X)$ *is defined by its t-th component*

$$\psi_t^{PS}(Y,T,X) \equiv \left(\frac{D_t}{P_t(X)}m(Y;\beta_t) \; - \left(\frac{D_t}{P_t(X)} - 1\right)e_t(X;\beta_t)\right)\frac{\sum_{t'\in\mathcal{S}} P_{t'}(X)}{p_{\mathcal{S}}}. \tag{8}$$

The semiparametric efficiency bound of β_t *with known propensity scores is smaller than the bound when the propensity scores are unknown. Furthermore, the result still holds when* $P_{t'}(X)$ *is known for only* $t' \in \mathcal{S} \subset \mathcal{T}$.

The semiparametric efficiency bound of the distributional feature of $Y(t)$ *for the population satisfying* $\mathbb{E}[m(Y(t);\beta_t)] = \mathbb{E}[\mathbb{E}[m(Y;\beta_t)|T=t,X]] = 0$ *is the same regardless the propensity scores are known.*

Theorem 3 suggests that even when the propensity scores are known for only the treated levels in $\mathcal{S} \subset \mathcal{T}$ and $P_t(X)$ is not known (i.e., $t \notin \mathcal{S}$), knowledge of these propensity scores for the treated levels reduces the semiparametric efficiency bound. The important lesson is that the asymptotic efficiency is improved by more information on the propensity scores for the treated level $\{P_{t'}(X)\}_{t'\in\mathcal{S}}$, but not affected by the propensity score for the counterfactual level $P_t(X)$. More specifically, the difference between the efficient influence functions for the unknown and known propensity scores cases is the term (7), i.e., ψ_t in (6) can be decomposed to ψ_t^{PS} in (8) and the influence from $\{P_{t'}(X)\}_{t'\in\mathcal{S}}$ in (7). The second part associated with $e_t(X;\beta_t)$ in (8) resembles (7) and comes from the auxiliary sample $\{T=t\}$. That is to say the influence from $P_t(X)$ remains whether $P_t(X)$ is known.

Frölich (2004b) discusses some heuristic intuition using the normalized propensity score and the result in the binary treatment literature. He notes that the variance bound depends on which and how many propensity scores are known. Our Theorem 3 provides a theoretical justification. The multi-valued treatment setup distinguishes different propensity scores, which are degenerated to one propensity score in a binary treatment case. The main reason behind this result is that the propensity score for the treated level defines the parameter of interest β_t when $\mathcal{S}$ is a strict subset of $\mathcal{T}$. But $P_t(X)$ does not play a role in the definition of ATE, as pointed out by Chen, Hong, and Tarozzi (2008).

Theorem 3 coincides with the results from the binary treatment effect literature that knowledge of the propensity score does not affect the efficiency bound of the ATE, but decreases that of the ATT. That includes the ATE and ATT for a binary treatment in Hahn (1998), Hirano, Imbens, and Ridder (2003), and the "verify-out-of-sample" for missing data in Chen, Hong, and Tarozzi (2008). The statement in Chen, Hong, and Tarozzi (2008), "*more information on the propensity score will not affect the asymptotic efficiency bounds*

for the verify-in-sample case," does not apply to a multi-valued treatment. We illustrate by an example of estimating $\mathbb{E}[Y(1)|T \in \{1,2,3\}]$, where the primary sample is $\mathcal{S} \equiv \{1,2,3\} \subset \mathcal{T} = \{1,2,3,4\}$. The auxiliary sample $\{T = 1\}$ is a subset of the primary sample, so this is the verify-in-sample case in Chen, Hong, and Tarozzi (2008). But our results suggest that knowledge of the propensity scores decreases the efficiency bound. This is true whether the auxiliary sample $\{T = t\}$ is independent or out of the primary sample $\{T \in \mathcal{S}\}$. The key is if the propensity score enters the definition of the object of interest $\mathbb{E}[Y(t)|T \in \mathcal{S}]$.

The semiparametric efficiency bound gives insight to construct an efficient regression estimator: when the propensity scores are known, we should make use of the true propensity score for the treated subpopulation $P_{t'}(X)$. We study the limiting property of our efficient propensity score regression estimators in the next section.

4 Efficient regression estimators

We estimate the cumulative distribution function of the potential outcome $Y(t)$ for the treated subpopulation with level t', for any $t, t' \in \mathcal{T}$. The estimand is based on the identification in (4) and (5)

$$\begin{aligned} F_{Y(t)|T}(y|t') \equiv \mathbb{E}\big[\mathbf{1}_{\{Y(t)\le y\}}\big|T = t'\big] &= \mathbb{E}\big[\mathbb{E}[\mathbf{1}_{\{Y\le y\}}|T = t, P_t(X), P_{t'}(X)]W\big] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{Y\le y\}}\middle|T = t, \frac{P_t(X)}{P_t(X)+P_{t'}(X)}\right]W\right] \end{aligned}$$

where the weight $W = D_{t'}/p_{t'}$ or $W = P_{t'}(X)/p_{t'}$. For the notation in Section 3, $m(Y;\beta) = \mathbf{1}_{\{Y\le y\}} - \beta$ for $y \in \mathcal{Y}$. The common distributional features, such as mean and quantiles, can be extended straightforwardly. For example, by changing the dependent variable from $\mathbf{1}_{\{Y\le y\}}$ to Y, we estimate the ATT $\mathbb{E}[Y(t)|T = t']$. We provide limit theory for estimating various distributional features of $F_{Y(t)|T}(y|t')$ in Section 4.6.

The regression estimators use the propensity scores as generated regressors. Denote the generated regressors to be $V = v_0(X)$, a vector of measurable functions v_0 of X. We use and extend the limit theory for the partial mean with generated regressors in Lee (2014). We construct efficient estimators for two cases — when the propensity scores are unknown and when they are known. We first present the limit theory using the propensity scores $V = v_0(X) = (P_t(X), P_{t'}(X))$. We show the limit theory for the estimators regressing on X in Section 4.3. Then we present the estimators regressing on the normalized propensity score $V = v_0(X) = P_t(X)/(P_t(X) + P_{t'}(X))$ in Section 4.4. These results imply efficient estimators for the binary treatment effect for the treated in Section 4.5. Denote the estimand $F_{Y(t)|T}(y|t') \equiv \beta_{t|t'}(y) \equiv \beta(y)$ that suppresses the dependence on t, t', for ease of exposition. All the results presented in this section are for any $t, t' \in \mathcal{T}$ and we will omit this statement

without loss of clarity.

4.1 Unkonwn propensity scores

The estimation procedure of the propensity score regression estimator follows three steps. We use a product kernel $K_h(x) \equiv h^{-d_x}\Pi_{l=1}^{d_x} k\big(\frac{x_l}{h}\big)$, where h is the bandwidth assumed the same for all the elements of the vector x for simplicity, and k is the r-order kernel function satisfying Assumption 6 in the Appendix.

Step 1. (Generated regressor) Estimate the propensity scores by a kernel regression,
$\hat{P}_t(x) = \sum_{i=1}^n D_{ti} K_{h_1}(X_i - x) / \sum_{i=1}^n K_{h_1}(X_i - x)$.

Step 2. (Regression) Estimate the conditional cdf by a kernel regression with the generated regressor $\hat{V}_i = \hat{v}(X_i) \equiv \big(\hat{P}_t(X_i), \hat{P}_{t'}(X_i)\big)^\top$ from Step 1.
$\hat{F}_{Y|T\hat{V}}(y|t,v) = \sum_{i=1}^n \mathbf{1}_{\{Y_i \le y\}} D_{ti} K_h(\hat{V}_i - v) / \sum_{i=1}^n D_{ti} K_h(\hat{V}_i - v)$.

Step 3. (Partial mean) Consider two estimators:

$$\hat{\beta}(y) = \frac{1}{n}\sum_{i=1}^n \hat{F}_{Y|T\hat{V}}(y|t,\hat{V}_i)\ \frac{D_{t'i}}{\hat{p}_{t'}} \quad \text{where} \quad \hat{p}_{t'} = \frac{1}{n}\sum_{i=1}^n D_{t'i}$$

$$\hat{\beta}^p(y) = \frac{1}{n}\sum_{i=1}^n \hat{F}_{Y|T\hat{V}}(y|t,\hat{V}_i)\ \frac{\hat{P}_{t'}(X_i)}{\hat{p}_{t'}}.$$

Note that the weight $D_{t'}/p_{t'}$ only uses the observations from the treated group with level t'. The weight $P_{t'}(X)/p_{t'}$ uses all the observations in the sample but contains sample variation of estimating $P_{t'}(X)$. We show that both weights lead to efficient estimators.

The estimator will include two fixed trimming functions in Step 2 and Step 3: In Step 2, the boundary of pretreatment variables X is trimmed. In Step 3, the density of the conditioning variables are bounded away from zero. The trimming functions ensure uniform convergence of the first- and second-step estimators over the range of integration that suffices for deriving the properties of the third-step estimator. In fairness, the choice of fixed trimming function can affect the interpretation of the estimands considered. That is, we estimate the ATT for the subpopulation whose pretreatment variables do not take extreme values and the common support Assumption 4(i) holds. Heckman, Ichimura, and Todd (1998) also estimate the region of common support. The fixed trimming choice allows us to focus on the technical issues associated with estimating the generated regressor. We suppress the two fixed trimming functions for notational ease.

Theorem 4 (Unknown propensity scores)
Suppose Assumptions 1, 4, and the Assumptions in the Appendix hold. Then uniformly in

$y \in \mathcal{Y}$,

$$\sqrt{n}\big(\hat{\beta}(y) - \beta(y)\big) = \sqrt{n}\big(\hat{\beta}^p(y) - \beta(y)\big) + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_t(Y_i, T_i, X_i; y) + o_p(1)$$

where

$$\psi_t(Y_i, T_i, X_i; y) \equiv \frac{D_{ti}}{P_t(X_i)} \frac{P_{t'}(X_i)}{p_{t'}} \big(\mathbf{1}_{\{Y_i \le y\}} - F_{Y|TX}(y|t, X_i)\big) + \big(F_{Y|TX}(y|t, X_i) - \beta(y)\big) \frac{D_{t'i}}{p_{t'}}$$

the same efficient influence function (6) derived in Theorem 2.

The asymptotic linear representation in Theorem 4 accounts for the sampling variation of estimating the propensity scores that are used in the generated regressors and the weight. The propensity score can be viewed as an index function of the pretreatment variables X. Lee (2014) shows that the estimation error associated with estimating the propensity scores as generated regressors is characterized by the *index bias* $F_{Y|TV}(y|t, v_0(X)) - F_{Y|TX}(y|t, X)$. The index bias is the impact of X on the outcome distribution that is not captured by the propensity scores $v_0(X) = (P_t(X), P_{t'}(X))$. The index bias is the key feature of the nonparametric regression with generated regressors that has been shown in the literature, for example, Hahn and Ridder (2013), Escanciano, Jacho-Chávez, and Lewbel (2014), Mammen, Rothe, and Schienle (2014), Lee (2014) among others. This is the key for the nonparametric propensity score regression estimator to reach the semiparametric efficiency bound. The important message is that

> *(Index bias)* The nonparametric estimation of the propensity scores recovers the information of the pretreatment variables on the outcome distribution that is not captured by the propensity scores.

More specifically, we show in the proof that the estimation error of the generated regressor $\hat{P}_{t'}(X_i)$ contributes

$$-\Big(\frac{D_{t'i}}{P_{t'}(X_i)} - 1\Big)\big(F_{Y|TV}(y|t, v_0(X_i)) - F_{Y|TX}(y|t, X_i)\big)\frac{P_{t'}(X_i)}{p_{t'}} \tag{9}$$

to the final estimators $\hat{\beta}(y)$ and $\hat{\beta}^p(y)$. Observe that the second part $F_{Y|TX}(y|t, X_i)\big(D_{t'i}/P_{t'}(X_i) - 1\big)P_{t'}(X_i)/p_{t'}$ is (7), the part in the efficient influence function from the treated group $\{T = t'\}$. The nonparametrically estimated propensity scores pick up the additional term in the efficient influence function in (7) . This is in line with the inverse propensity score weighting estimator and the doubly robust estimator in Cattaneo (2010). The first part $F_{Y|TV}\big(D_{t'i}/P_{t'}(X_i) - 1\big)P_{t'}(X_i)/p_{t'}$ cancels out the sampling variation of the estimation in Step 2 and Step 3. The estimation error of the generated regressor $\hat{P}_t(X_i)$ contributes the

term

$$\Big(\frac{D_{ti}}{P_t(X_i)} - 1\Big)\big(F_{Y|TV}(y|t, v_0(X_i)) - F_{Y|TX}(y|t, X_i)\big)\frac{P_{t'}(X_i)}{p_{t'}} \tag{10}$$

to the final estimators. Similarly, the second part associated with $F_{Y|TX}(y|t, X_i)$ recovers the part of the efficient influence function from using the observations from the auxiliary sample $\{T = t\}$ in the inner expectation, i.e., the second part associated with e_t in (8).

The generated regressor $V = (P_t(X), P_{t'}(X))$ plays two roles — the *regressor* that determines the regression function $F_{Y|TV}$ and the *argument* of the regression function that is averaged out in the third step. The estimation error of $\hat{V}$ from its role as the *regressor* is the key of our results. More detail is in Lee (2014) and we discuss some intuition in the following. The third step integrates over the weight $P_{t'}(X)$ and the *argument* $V = (P_t(X), P_{t'}(X))$ of the regression function. By changing the order of the integrations in the second step and third step, the kernel function effectively makes the weight and the argument evaluated at the estimated $\hat{V}$. We then use a Taylor expansion to derive the first order influence of the estimation error $\hat{V} - V$.

4.2 Known propensity scores

When the propensity scores are known, we propose an efficient regression estimator using a nonparametrically estimated $P_t(X)$ and the true known $P_{t'}(X)$ as the generated regressors. We modify Step 2 and Step 3 in the above estimation procedure to the following

Step 2′. (Regression) Estimate the conditional cdf by a kernel regression using the regressors $(T, \hat{V})$ where $\hat{V} = (\hat{P}_t(X), P_{t'}(X))^\top$ from Step 1.

Step 3′. (Partial mean)

$$\tilde{\beta}(y) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}_{Y|T\hat{V}}(y|t, \hat{P}_t(X_i), P_{t'}(X_i))\ \frac{P_{t'}(X_i)}{\tilde{p}_{t'}} \quad \text{where} \quad \tilde{p}_t = \frac{1}{n}\sum_{i=1}^{n} P_{t'}(X_i).$$

The estimation for the treated subpopulation is more precise by using all observations that share the same value of the propensity score $P_{t'}(X_i)$, comparing with using only the treated observations by $D_{t'i}$.

Theorem 5 (Known propensity scores)
Suppose Assumptions 1, 4, and the Assumptions in the Appendix hold. Then uniformly in $y \in \mathcal{Y}$,

$$\sqrt{n}\big(\tilde{\beta}(y) - \beta(y)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_t^{PS}(Y_i, T_i, X_i; y) + o_p(1)$$

where

$$\psi_t^{PS}(Y_i, T_i, X_i; y) \equiv \frac{D_{ti}}{P_t(X_i)} \frac{P_{t'}(X_i)}{p_{t'}} \big(\mathbf{1}_{\{Y_i \leq y\}} - F_{Y|TX}(y|t, X_i)\big) + \big(F_{Y|TX}(y|t, X_i) - \beta(y)\big) \frac{P_{t'}(X_i)}{p_{t'}}$$

the same efficient influence function (8) derived in Theorem 3.

It has been a widely-studied puzzle in the literature that using the estimated propensity score is more efficient than using the true one (Hahn, 1998; Imai and van Dyk, 2004; Hirano, Imbens, and Ridder, 2003; Graham, 2011; Abadie and Imbens, 2012). Our insight in the index bias provides a new theoretical explanation. Theorem 3 suggests that knowledge of $P_t(X)$ does not affect the semiparametric efficiency bound. This gives us the freedom not to use the true propensity score. From the above discussion on the index bias, the nonparametric estimation of the propensity score captures the information of X on the outcome distribution that is reduced in the propensity scores. This information is lost if regressing on the true propensity score. The sampling variation of $\hat{P}_t(X)$ picks up the term associated with the auxiliary sample in the efficient influence function. Therefore, we should use a nonparametrically estimated $P_t(X)$ even though it is known.

Imai and van Dyk (2004) discuss this issue for subclassification by an application with randomized treatment assignment. They claim that an estimated propensity score accounts for the sample-specific relationship of the treatment and the covariates, which is lost in the true propensity score. Abadie and Imbens (2012) find that when the propensity score is parametrically specified, for the ATE, matching on the estimated propensity score is more efficient than matching on the true one. When the number of matches increases with the sample size, a subclassification or matching estimator is essentially like a nonparametric regression estimator. We provide a new theoretical explanation that the nonparametrically estimated propensity scores recover the relationship of the covariates and the outcome.

Theorem 5 parallels previous findings in the inverse propensity score weighting estimators or GMM framework in Hirano, Imbens, and Ridder (2003), Cattaneo (2010), and Graham (2011). In contrast, we study this paradox from a view of projection onto the propensity score. Hirano, Imbens, and Ridder (2003) interpret the efficiency loss of using the true propensity score by the empirical likelihood estimation. Nonparametrically estimating the propensity score captures the information content of a conditional moment restriction of the propensity score ($\mathbb{E}[D_t - P_t(X)|X] = 0$) by a sequence of unconditional moment restrictions. Graham (2011) calculates the efficiency bound incorporating the conditional moment of the propensity score as the auxiliary moment. While a parametric estimate of the propensity score only satisfies a finite number of the moment conditions, using the true propensity score makes no use of any information contained in the auxiliary moment. So the efficiency is improved in the same way as adding moment restrictions in a GMM framework.

Remark (Propensity score weighting estimation)
Our estimator shares the same spirit with the inverse propensity score weighting estimator. When the propensity scores are known, the efficient estimator in Hirano, Imbens, and Ridder (2003) is inversely weighted by the nonparametrically estimated propensity score $\hat{P}_t(X)$ and then reweighed by the true propensity score $P_{t'}(X)$ to adjust for the treated group. And the conditional outcome distribution given $T = t$, $F_{Y|T}(y|t) = F_{Y(t)|T}(y|t)$, is efficiently estimated by $n^{-1}\sum_{i=1}^n \mathbf{1}_{\{Y_i \le y\}} \frac{D_{ti}}{\hat{P}_t(X_i)} \frac{P_t(X_i)}{n^{-1}\sum_{j=1}^n P_t(X_j)}$. In contrast to the case when the propensity score is unknown, $F_{Y|T}(y|t) = F_{Y(t)|T}(y|t)$ is efficiently estimated by $\sum_{i=1}^n \mathbf{1}_{\{Y_i \le y\}} D_{ti} \big/ \sum_{i=1}^n D_{ti}$.

For a multi-valued treatment, Ao, Calonico, and Lee (2014) provide efficient inverse propensity score weighting estimators and doubly robust estimators by extending the idea in Cattaneo (2010). When the propensity scores are known, we expect the corresponding efficient estimators to be constructed similarly as Hirano, Imbens, and Ridder (2003). This is left for future research.

4.3 Regression on observed characteristics

Consider regression estimators adjusting for pretreatment variables X:

$$\hat{\beta}_x(y) = \frac{1}{n}\sum_{i=1}^n \hat{F}_{Y|TX}(y|t, X_i)\ \frac{D_{t'i}}{\hat{p}_{t'}}$$
$$\hat{\beta}^p_x(y) = \frac{1}{n}\sum_{i=1}^n \hat{F}_{Y|TX}(y|t, X_i)\ \frac{\hat{P}_{t'}(X_i)}{\hat{p}_{t'}}$$
$$\tilde{\beta}_x(y) = \frac{1}{n}\sum_{i=1}^n \hat{F}_{Y|TX}(y|t, X_i)\ \frac{P_{t'}(X_i)}{\tilde{p}_{t'}}$$

The following Theorem 6 shows that $\hat{\beta}_x(y)$ and $\hat{\beta}^p_x(y)$ are semiparametrically efficient when the propensity scores are unknown. And $\tilde{\beta}_x(y)$ is semiparametrically efficient when the propensity score are known. Hahn (1998) proposes series estimators for $\hat{\beta}_x(y)$ and $\tilde{\beta}_x(y)$ for a binary treatment and shows that they reach the semiparametric efficiency bounds.

Theorem 6 (Regression on X)
Suppose Assumption 5 in the Appendix holds for $V = v_0(X) = X$. Suppose Assumptions 6 and 7 hold. For $\hat{\beta}^p_x(y)$, assume $P_{t'}(X) \in \mathcal{C}^{\alpha_v}_M(\mathcal{X})$ with $\alpha_v > d_x/2$, $g < \min\{1/(d_x + 2\alpha_v), 1/d_x - \eta\}$, and $r_1 > 1/(2g)$. Then uniformly in $y \in \mathcal{Y}$,

(a) when the propensity scores are unknown,

$$\sqrt{n}\big(\hat{\beta}_x(y) - \beta(y)\big) = \sqrt{n}\big(\hat{\beta}^p_x(y) - \beta(y)\big) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_t(Y_i, T_i, X_i; y) + o_p(1);$$

(b) when the propensity scores are known,

$$\sqrt{n}\big(\tilde{\beta}_x(y) - \beta(y)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_t^{PS}(Y_i, T_i, X_i; y) + o_p(1).$$

Theorem 6(b) implies that the estimator using the true propensity scores $V = v_0(X) = (P_t(X), P_{t'}(X))$ $\tilde{\beta}_v(y) = n^{-1}\sum_{i=1}^n \hat{F}_{Y|TX}(y|t, V_i)\ P_{t'}(X_i)/\tilde{p}_{t'}$ has the influence function equal to

$$\psi_t^{PS}(Y_i, T_i, V_i; y) \equiv \frac{D_{ti}}{P_t(X_i)}\frac{P_{t'}(X_i)}{p_{t'}}\big(\mathbf{1}_{\{Y_i \le y\}} - F_{Y|TV}(y|t, V_i)\big) + \big(F_{Y|TV}(y|t, V_i) - \beta(y)\big)\frac{P_{t'}(X_i)}{p_{t'}}.$$

Because the whole set of observables X provides finer conditioning variables than its index $v_0(X)$, $\tilde{\beta}_{Vt}$ regressing on the true propensity scores is less efficient than $\tilde{\beta}_{Xt}$ regressing on X (Hahn, 1998; Heckman, Ichimura, and Todd, 1998; Lee, 2014). Lemma D.1 in Lee (2014) formally shows that $\mathbb{E}[\psi_t^{PS}(Y, T, V; y)^2] \ge \mathbb{E}[\psi_t^{PS}(Y, T, X; y)^2]$. When the index bias is not zero, i.e., there exists x such that $F_{Y|TV}(y|t, v_0(x)) \ne F_{Y|TX}(y|t, x)$, the inequality between the corresponding asymptotic variances is strict. The difference between $\psi_t^{PS}(Y_i, T_i, X_i; y)$ and $\psi_t^{PS}(Y_i, T_i, V_i; y)$ is exactly the same as the influence of estimating the generated regressor $\hat{P}_t(X)$ in (10). This again shows that $\hat{P}_t(X)$ recovers the information lost in regressing on the true $P_t(X)$.

4.4 Normalized propensity score

Consider the estimator regressing on the normalized propensity score $P_{t|\{t,t'\}}(X) = P_t(x)/(P_t(x)+P_{t'}(x)) \equiv b_t(X)$ for notational ease. We first consider the case when we do not know the true propensity scores. The estimator is modified from the previous procedure.

Step 1b. (Generated regressor) Estimate the normalized propensity scores by a kernel regression,

$$\hat{b}_t(x) = \frac{\hat{P}_t(x)}{\hat{P}_t(x) + \hat{P}_{t'}(x)} = \frac{\sum_{i=1}^n D_{ti}K_{h_1}(X_i - x)}{\sum_{i=1}^n \big(D_{ti} + D_{t'i}\big)K_{h_1}(X_i - x)}.$$

Step 2b. (Regression) Estimate the conditional cdf by a kernel regression with the generated regressor $\hat{V}_i = \hat{b}_t(X_i)$ from Step 1.
$\hat{F}_{Y|T\hat{V}}(y|t, v) = \sum_{i=1}^n \mathbf{1}_{\{Y_i \le y\}} D_{ti}K_h(\hat{b}_t(X_i) - v)\big/\sum_{i=1}^n D_{ti}K_h(\hat{b}_t(X_i) - v)$.

Step 3b. (Partial mean)

$$\hat{\beta}_b(y) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}_{Y|T\hat{V}}(y|t, \hat{b}_t(X_i))\ \frac{D_{t'i}}{\hat{p}_{t'}}.$$

Theorem 7 (Unknown normalized propensity scores)
Suppose Assumptions 1, 4, and the Assumptions in the Appendix hold. Then uniformly in $y \in \mathcal{Y}$,

$$\sqrt{n}\big(\hat{\beta}_b(y) - \beta(y)\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_t(Y_i, T_i, X_i; y) + o_p(1)$$

that has the same efficient influence function in Theorem 4.

When the propensity scores are known, we use the known propensity score for the treated level $P_{t'}(X)$ in estimating the normalized propensity score. So the above Step 1b is modified to

Step 1b′. (Generated regressor) Using the known $P_{t'}(X)$, estimate the normalized propensity scores by a kernel regression,

$$\tilde{b}_t(x) = \frac{\hat{P}_t(x)}{\hat{P}_t(x) + P_{t'}(x)} = \frac{\sum_{i=1}^{n} D_{ti} K_{h_1}(X_i - x)}{\sum_{i=1}^{n} \big(D_{ti} + P_{t'}(x)\big) K_{h_1}(X_i - x)}.$$

Step 2b and Step 3b are the same. Denote the corresponding estimator using $\hat{V} = \tilde{b}_t(X)$ as a generated regressor to be

$$\tilde{\beta}_b(y) = \frac{1}{n} \sum_{i=1}^{n} \hat{F}_{Y|T\hat{V}}(y|t, \tilde{b}_t(X_i)) \ \frac{P_{t'}(X_i)}{\tilde{p}_{t'}}$$

Theorem 8 (Known normalized propensity scores)
Under Assumptions 1, 4, and the Assumptions in the Appendix, for each $t \in \mathcal{T}$, *uniformly in* $y \in \mathcal{Y}$

$$\sqrt{n}\big(\tilde{\beta}_b(y) - \beta(y)\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_t^{PS}(Y_i, T_i, X_i; y) + o_p(1)$$

that has the same efficient influence function in Theorem 5.

4.5 The binary case

Theorem 7 implies that when the propensity score is unknown, the efficient estimator for $\mathbb{E}[Y(1)|T = 0]$ for a binary treatment is

$$\hat{\beta}_b(y) = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{E}}\big[Y \big| T = 1, \hat{P}_1(X) = \hat{P}_1(X_i)\big] \ \frac{D_{0i}}{\hat{p}_0}.$$

When the propensity score is known, Theorem 8 suggests that an efficient regression estimator for the binary case

$$\tilde{\beta}_b(y) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}\big[Y\big|T=1, \tilde{b}_1(X) = \tilde{b}_1(X_i)\big]\ \frac{P_0(X_i)}{\tilde{p}_0}$$
$$\text{where}\ \ \tilde{b}_1(X) = \frac{\hat{P}_1(X)}{\hat{P}_1(X)+P_0(X)}\ \ \text{and}\ \ \ \tilde{p}_0 = \frac{1}{n}\sum_{i=1}^{n} P_0(X_i).$$

The efficient estimator $\tilde{\beta}_b(y)$ utilizes the knowledge of the propensity score for the treated level $P_0(X)$. And estimating the propensity score for the counterfactual level $\hat{P}_1(X)$ recovers the information of the covariates that is lost from regressing on the propensity scores.

When the propensity score is known, it might be intuitive to construct an estimator based on Heckman, Ichimura, and Todd (1998)

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}[Y|T=1, \hat{P}_1(X) = \hat{P}_1(X_i)]\ \frac{P_0(X_i)}{\tilde{p}_0}.$$

We label this estimator as the HIT estimator. Theorem 9 below implies that the HIT estimator is not efficient when the propensity score is known. This is in contrast to the efficient estimator for the population $\hat{\mathbb{E}}[Y(1)] = n^{-1}\sum_{i=1}^{n}\hat{\mathbb{E}}[Y|T=1,\hat{P}_1(X)=\hat{P}_1(X_i)]$ in Hahn and Ridder (2013); Mammen, Rothe, and Schienle (2014); Lee (2014). This is also different from the estimator regressing on the covariates $\tilde{\beta}_x(y) = n^{-1}\sum_{i=1}^{n}\hat{\mathbb{E}}[Y|T=1,X=X_i]P_0(X_i)/\tilde{p}_0$ that is efficient by Theorem 6 and Hahn (1998). From Theorem 4 and the discussion in the last paragraph of Section 4.1, the generated regressor $\hat{P}_1(X)$ has an influence term through the weight $P_0(X)/p_0 = (1-P_1(X))/p_0$. This situation does not appear in the multi-valued case. We formalize the above discussion in Theorem 9 below. Let the HIT estimator of $F_{Y(t)|T}(y|t')$ be

$$\check{\beta}(y) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}\big[\mathbf{1}_{\{Y\le y\}}\big|T=1, \hat{P}_1(X) = \hat{P}_1(X_i)\big]\ \frac{P_0(X_i)}{\tilde{p}_0}.$$

Theorem 9 (HIT estimator)
Suppose Assumptions 1, 4, and the Assumptions in the Appendix hold. Then uniformly in $y \in \mathcal{Y}$,

$$\sqrt{n}\big(\check{\beta}(y) - \beta(y)\big) = \frac{1}{n}\sum_{i=1}^{n}\psi_t^{PS}(Y_i,T_i,X_i;y)$$
$$+ \big(D_{1i} - P_1(X_i)\big)\big(F_{Y|TV}(y|1,V(X_i)) - F_{Y|TX}(y|1,X_i)\big)\frac{1}{p_0} + o_p(1).$$

The second line of the above influence function of $\check{\beta}(y)$ can be expressed as $-\big(D_{0i} - P_0(X_i)\big)\big(F_{Y|TV}(y|1, V(X_i)) - F_{Y|TX}(y|1, X_i)\big)/p_0$ that equals the estimation error associated with estimating $\hat{P}_0(X)$ as a generated regressor in (9) discussed in Section 4.1. Intuitively, this is because $\hat{P}_0(X)+\hat{P}_1(X) = 1$, regressing on $\hat{P}_1(X)$ is effectively regressing on $(\hat{P}_0(X), \hat{P}_1(X))$ or $\hat{P}_1(X)/(\hat{P}_0(X) + \hat{P}_1(X))$.

4.6 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 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 in empirical process theory. These Hadamard-differentiable functionals can be highly nonlinear functionals of the cdf, but admit a linear functional derivative. Hadamard-differentiablility is a high-level assumption that could impose restrictions or smoothness on the distribution functions of potential outcomes. Additional assumptions might be needed for different policy functionals. For instance, Bhattacharya (2007) gives regularity conditions for Hadamard-differentiability of Lorenz and Gini functionals.

In this section, we denote the estimand $\beta_{t|t'}(y) = F_{Y(t)|T}(y|t')$. The corresponding Theorems 4 to 9 in the previous sections provide the uniform asymptotic linear representation for the corresponding estimator $\hat{\beta}_{t|t'}(y)$.

Theorem 10
Consider any $t, t' \in \mathcal{T}$. Assume the conditions in the asymptotic theorem for $\hat{\beta}_{t|t'}(y)$ hold such that uniformly in $y \in \mathcal{Y}$

$$\sqrt{n}\big(\hat{\beta}_{t|t'}(y) - \beta_{t|t'}(y)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_{t|t'i}(y) + o_p(1).$$

Then

1. *(Weak Convergence)*

$$\sqrt{n}\big(\hat{\beta}_{t|t'}(\cdot) - \beta_{t|t'}(\cdot)\big) \Rightarrow \mathbb{G}_{t|t'}(\cdot)$$

   where $\mathbb{G}_{t|t'}(y)$ is a Gaussian process indexed by $y \in \mathcal{Y}$ in $l^\infty(\mathcal{Y})$, with mean zero and covariance kernel defined by the limit of the second moment of $\psi_{t|t'i}$.

2. *(Functional Delta Method) Consider the parameter β as an element of a parameter space $D_\beta \subset l^\infty(\mathcal{Y})$ with D_β containing the true value $\beta_{t|t'}$. Suppose a functional $\Gamma(\beta)$*

mapping D_β to $l^\infty(\mathcal{W})$ is Hadamard differentiable in β at $\beta_{t|t'}$ with derivative Γ'_β. [3]

$$\left|\sqrt{n}\big(\Gamma(\hat{\beta}_{t|t'})(w) - \Gamma(\beta_{t|t'})(w)\big) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma'_\beta(\psi_{t|t'i})(w)\right| = o_p(1)$$

$$\sqrt{n}\big(\Gamma(\hat{\beta}_{t|t'})(w) - \Gamma(\beta_{t|t'})(w)\big) \Rightarrow \Gamma'_\beta(\mathbb{G}_{t|t'})(w) \equiv G(w)$$

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 $\Gamma'_\beta(\psi_{t|t'i})$.

5 Conclusion

We examine the role of the propensity scores in estimating the multi-valued treatment effects for the treated from three perspectives: (i) (*Identification*) The equivalent representations of weak unconfoundedness generalize the propensity score methodology in Rosenbaum and Rubin (1983), Imbens (2000), and Lechner (2001). (ii) (*Semiparametric efficiency bound*) We calculate the semiparametric efficiency bounds for the cases when the propensity scores are unknown and known, for the population and for the treated. Knowledge of the propensity score for the counterfactual level has no role in the semiparametric efficiency bound. But knowledge of the propensity score for the treated level improves asymptotic precision. (iii) (*Regression estimator*) We propose efficient regression estimators that project on two propensity scores — one for the counterfactual level and one for the treated level. The nonparametrically estimated propensity score recovers the impact of the pretreatment variables on the outcome that is not captured by the propensity score. By the insight of the efficiency bound, we should utilize the true propensity score for the treated value. These findings provide new explanation to some paradoxes on the propensity scores in the binary treatment effect and missing data literature.

This paper provides a theoretical foundation for propensity scores matching methodology. Our insights on the multi-valued treatment can be extended to the generalized propensity scores for a continuous treatment in Hirano and Imbens (2004) and Lee (2014). For practical application of the efficient regression estimators, the bandwidth choice and small-sample performance are left for future research. For example, Busso, DiNardo, and McCrary (2014) is an extensive small-sample study on various estimators.

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

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

Appendix

A Proofs in Section 2 Identification

Proof of Theorem 1

(a) For any $t, t' \in \mathcal{T}$,

$$\begin{aligned}
&\mathbb{E}[D_{t'}|Y(t), P_{t'}(X), g(X)] = \mathbb{E}\big[\mathbb{E}[D_{t'}|Y(t), P_{t'}(X), g(X), X]\big|Y(t), P_{t'}(X), g(X)\big] \\
&= \mathbb{E}\big[\mathbb{E}[D_{t'}|Y(t), X]\big|Y(t), P_{t'}(X), g(X)\big] = \mathbb{E}\big[\mathbb{E}[D_{t'}|X]\big|Y(t), P_{t'}(X), g(X)\big] \\
&= \mathbb{E}\big[P_{t'}(X)\big|Y(t), P_{t'}(X), g(X)\big] = P_{t'}(X)
\end{aligned}$$

where the third equality is by Assumption 1. By the law of iterated expectations, $\mathbb{E}[D_{t'}|P_{t'}(X), g(X)] = P_{t'}(X)$. So $\mathbb{E}[D_{t'}|Y(t), P_{t'}(X), g(X)] = \mathbb{E}[D_{t'}|P_{t'}(X), g(X)]$ for any $t, t' \in \mathcal{T}$ and $g(X)$.

The reverse direction follows by letting $g(X) = X$.

(b) We first prove another equivalent representation to Weak unconfoundedness Assumption 1:

For any $t \in \mathcal{T}$, for all $\mathcal{S} \subseteq \mathcal{T}$, $T \perp Y(t)\big|\big\{X, \sum_{s\in\mathcal{S}} D_s = 1\big\}$ or equivalently $D_{t'} \perp Y(t)\big|\big\{X, \sum_{s\in\mathcal{S}} D_s = 1\big\}$ for $t' \in \mathcal{S}$.

We abuse the notation f for joint probability density functions.

$$\begin{aligned}
\mathbb{E}\Big[D_{t'}\Big|Y(t), X, \sum_{s\in\mathcal{S}} D_s = 1\Big] &= \frac{f\big(D_{t'} = 1, Y(t), X\big)}{f\big(Y(t), X, \sum_{s\in\mathcal{S}} D_s = 1\big)} = \frac{Prob\big(D_{t'} = 1\big|X\big)}{Prob\big(\sum_{s\in\mathcal{S}} D_s = 1\big|X\big)} \\
&= \mathbb{E}\Big[D_{t'}\Big|X, \sum_{s\in\mathcal{S}} D_s = 1\Big].
\end{aligned}$$

The reverse direction follows by letting $\mathcal{S} = \mathcal{T}$.

Using the above result, (b) is proved following the same procedure as the proof in (a).

□

Proof of Corollary 1

Theorem 1 implies for any $t' \in \mathcal{S}$,

$$Prob\big(D_{t'} = 1\big|Y(t), \{P_s(X)\}_{s\in\mathcal{S}}, g(X)\big) = Prob\big(D_{t'} = 1\big|\{P_s(X)\}_{s\in\mathcal{S}}, g(X)\big).$$

Setting $g(X) = 1$ gives the first result in (a).

Following the same argument in Rosenbaum and Rubin (1983), the set of the propensity scores $\{P_s(X)\}_{s\in\mathcal{S}}$ has a balancing property: within strata with the same value of $\{P_s(X)\}_{s\in\mathcal{S}}$, the probability of being assigned to a level s in $\mathcal{S}$ or some other levels not in $\mathcal{S}$ does not depend on the value of X. And $\{P_s(X)\}_{s\in\mathcal{S}}$ is the "coarsest" balancing score and X is the "finest." [4] The second result in (a) follows by setting $g(X)$ to be a balancing

[4] A balancing score $g(X)$ for D_t satisfies $\mathbb{E}[D_t|X, g(X)] = \mathbb{E}[D_t|g(X)]$. The left-hand-side is $P_t(X)$ and the

score of S. Then the conditioning set $\{\{P_s(X)\}_{s\in\mathcal{S}}, g(X)\} = \{g(X)\}$.

The proof of the result (b) follows the same procedure. □

B Proofs in Section 3 Semiparametric Efficiency Bounds

B.1 Proof of Theorem 2

We follow the procedure in the proof of Theorem 1 in Cattaneo (2010), so we skip the repetition and note the difference. Consider a regular parametric submodel of the joint distribution the observed data (Y,T,X) indexed by θ. Define the score to be $S(y,t,x;\theta_0) = S_y(y,t,x) + S_p(t,x) + S_x(x)$, where

$$S_y(y,T,x) \equiv \sum_{j\in\mathcal{T}} D_j s_j(y,x), \quad s_j(y,x) \equiv \frac{\partial}{\partial\theta}\log f_{Y(j)|X}(y|x;\theta)\big|_{\theta_0},$$
$$S_p(T,x) \equiv \sum_{j\in\mathcal{T}} D_j \frac{\dot{P}_j(x)}{P_j(x)}, \quad \dot{P}_j(x) \equiv \frac{\partial}{\partial\theta}P_j(x;\theta)\big|_{\theta_0},$$
$$S_x(x) \equiv \frac{\partial}{\partial\theta}\log f(x;\theta)\big|_{\theta_0}.$$

The tangent space is characterized by $\mathcal{H}_y + \mathcal{H}_p + \mathcal{H}_x$ in Cattaneo (2010), where $\mathcal{H}_y \equiv \{S_y(Y,T,X) : s_j(Y,X) \in L_0^2(F_{Y(j)|X}(Y|X)), \forall j \in \mathcal{T}\}$, $\mathcal{H}_p \equiv \{S_p(T,X) \in L_0^2(F_{T|X})\}$, and $\mathcal{H}_x \equiv \{S_x(X) : S_x(X) \in L_0^2(F_x)\}$. So $\mathbb{E}[S_p(T,X)|X] = \sum_{j\in\mathcal{T}} \dot{P}_j(X) = 0$ and $\mathbb{E}[S_p^2(T,X)|X] = \sum_{j\in\mathcal{T}} \dot{P}_j^2(X)/P_j(X) < \infty$. Define the main component of the efficient influence function

$$EIF \equiv \frac{D_t}{P_t(X)} m(Y;\beta_t) \frac{\sum_{t'\in\mathcal{S}} P_{t'}(X)}{p_{\mathcal{S}}}$$
$$- \sum_{t'\in\mathcal{S}} \frac{P_{t'}(X)}{p_{\mathcal{S}}} e_t(X;\beta_t)\Big(\frac{D_t}{P_t(X)} - 1\Big) + \sum_{t'\in\mathcal{S}} \frac{P_{t'}(X)}{p_{\mathcal{S}}} e_t(X;\beta_t)\Big(\frac{D_{t'}}{P_{t'}(X)} - 1\Big)$$

belonging to the tangent space.

We first obtain the following equality, for any function $A(Y,X)$,

$$\mathbb{E}\Big[D_t \mathbb{E}\big[A(Y,X)\big|X,D_t\big]\Big|X\Big] = \mathbb{E}\big[A(Y,X)\big|T=t,X\big]\ P_t(X). \tag{11}$$

Let $\underline{m}(\underline{\beta}) = \big[m(Y(t);\beta_1)^\top, ..., m(Y(t);\beta_J)^\top\big]$. So $\mathbb{E}_{\theta_0}[\underline{m}(\underline{\beta}(\theta_0))|T\in\mathcal{S}] = 0$. For any $d_\beta J \times d_m J$ positive semidefinite matrix A, $\frac{\partial}{\partial\theta} A\mathbb{E}_\theta[\underline{m}(\underline{\beta}(\theta))|T\in\mathcal{S}]\Big|_{\theta=\theta_0} = A\frac{\partial}{\partial\underline{\beta}}\mathbb{E}[\underline{m}(\underline{\beta})|T\in\mathcal{S}]\frac{\partial}{\partial\theta}\underline{\beta}(\theta) + A\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\underline{m}(\underline{\beta})|T\in\mathcal{S}]\Big|_{\theta=\theta_0} = 0$. So $\frac{\partial}{\partial\theta}\underline{\beta}(\theta) = -\big(A\Gamma_{*|t'}\big)^{-1} A\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\underline{m}(\underline{\beta})|T\in\mathcal{S}]\Big|_{\theta=\theta_0}$.

right-hand-side is $\mathbb{E}[P_t(X)|g(X)]$. Therefore, a function $b(X)$ is a *balancing score* for D_t if and only if there exists a function h such that $P_t(X) = h(g(X))$.

The t-th element of $\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\underline{m}(\underline{\beta})|T\in\mathcal{S}]$ is

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[\mathbb{E}_\theta[m(Y(t);\beta_t)|X]\frac{\sum_{t'\in\mathcal{S}}P_{t'}(X;\theta)}{p_{\mathcal{S}}(\theta)}\right]\Bigg|_{\theta=\theta_0}\\
&=\sum_{t'\in\mathcal{S}}\int m(y;\beta_t)f_{Y(t)|X}(y|x;\theta)f(x;\theta)P_{t'}(x;\theta)\Big(s_t(y,x)+\frac{\dot{P}_{t'}(x)}{P_{t'}(x)}+S_x(x)\Big)dydx\frac{1}{p_{\mathcal{S}}}\Bigg|_{\theta=\theta_0}\\
&=\mathbb{E}\left[\frac{D_t}{P_t(X)}m(Y;\beta_t)\frac{\sum_{t'\in\mathcal{S}}P_{t'}(X)}{p_{\mathcal{S}}}S(Y,T,X;\theta_0)\right]\\
&+\sum_{t'\in\mathcal{S}}\mathbb{E}\left[\frac{P_{t'}(X)}{p_{\mathcal{S}}}\mathbb{E}\Big[m(Y;\beta_t)\Big|T=t,X\Big]\Big(\frac{\dot{P}_{t'}(X)}{P_{t'}(X)}-\frac{\dot{P}_t(X)}{P_t(X)}\Big)\right].
\end{aligned}\tag{12}$$

The first term comes from (11) by setting $A(Y,X)=m(Y;\beta_t)s_t(Y,X)$. The second term needs some calculation.

$$\begin{aligned}
&\mathbb{E}\left[\frac{D_t-P_t(X)}{P_t(X)}S(Y,T,X;\theta_0)\Big|X\right]=\mathbb{E}\left[\frac{D_t-P_t(X)}{P_t(X)}\Big(\sum_{j\in\mathcal{T}}D_js_j(Y,X)+D_j\frac{\dot{P}_j(X)}{P_j(X)}+S_x(X)\Big)\Big|X\right]\\
&=\mathbb{E}\left[\frac{D_t}{P_t(X)}\Big(s_t(Y,X)+\frac{\dot{P}_t(X)}{P_t(X)}\Big)\Big|X\right]-\mathbb{E}\left[\sum_{j\in\mathcal{T}}\Big(D_js_j(Y,X)+D_j\frac{\dot{P}_j(X)}{P_j(X)}\Big)\Big|X\right]\\
&=\mathbb{E}[s_t(Y,X)|T=t,X]+\frac{\dot{P}_t(X)}{P_t(X)}-\sum_{j\in\mathcal{T}}\Big(P_j(X)\mathbb{E}[s_j(Y,X)|T=j,X]+\dot{P}_j(X)\Big)=\frac{\dot{P}_t(X)}{P_t(X)}
\end{aligned}\tag{13}$$

using (11), $\mathbb{E}[s_j(Y,X)|T=j,X]=0\ \forall j\in\mathcal{T}$, and $\mathbb{E}[S_p(T,X)|X]=\sum_{j\in\mathcal{T}}\dot{P}_j(X)=0$. Therefore, by the law of iterated expectations,

$$\begin{aligned}
&\mathbb{E}\left[\frac{P_{t'}(X)}{p_{\mathcal{S}}}\mathbb{E}\Big[m(Y,\beta_t)\Big|T=t,X\Big]\frac{D_t-P_t(X)}{P_t(X)}S(Y,T,X;\theta_0)\right]\\
&=\mathbb{E}\left[\frac{P_{t'}(X)}{p_{\mathcal{S}}}e_t(X;\beta_t)\frac{\dot{P}_t(X)}{P_t(X)}\right]
\end{aligned}\tag{14}$$

$$\begin{aligned}
&\sum_{t'\in\mathcal{S}}\mathbb{E}\left[\frac{P_{t'}(X)}{p_{\mathcal{S}}}\mathbb{E}\Big[m(Y,\beta_t)\Big|T=t,X\Big]\frac{D_{t'}-P_{t'}(X)}{P_{t'}(X)}S(Y,T,X;\theta_0)\right]\\
&=\sum_{t'\in\mathcal{S}}\mathbb{E}\left[\frac{P_{t'}(X)}{p_{\mathcal{S}}}e_t(X;\beta_t)\frac{\dot{P}_{t'}(X)}{P_{t'}(X)}\right].
\end{aligned}\tag{15}$$

We discuss some intuition from the above equations and (12). The first equation (14) comes from the fact that only the observations with treatment level t are used in the inner regression to identify the potential outcome $Y(t)$, $\mathbb{E}[Y(t)|X]=\mathbb{E}[Y|T=t,X]$. The second equation (15) is from the outer expectation that uses only the observations from the treated subpopulation $\{T\in\mathcal{S}\}$.

It follows

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta\Big[\mathbb{E}_\theta\big[m(Y,\beta_t)\big|T=t,X\big]\frac{\sum_{t'\in\mathcal{S}}P_{t'}(X;\theta)}{p_{\mathcal{S}}(\theta)}\Big]\Big|_{\theta=\theta_0}=\mathbb{E}\big[\psi(Y,T,X)S(Y,T,X;\theta_0)\big].$$

So the parameter is path wise differentiable $\frac{\partial}{\partial\theta}\underline{\beta}(\theta)=-\big(A\Gamma_{*|t'}\big)^{-1}A\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\underline{m}(\underline{\beta})|T\in\mathcal{S}]\Big|_{\theta=\theta_0}$
$=-\big(A\Gamma_*\big)^{-1}A\mathbb{E}[\psi S]=\mathbb{E}\big[\big(-\big(A\Gamma_*\big)^{-1}A\psi\big)S\big]$. □

B.2 Proof of Theorem 3

The tangent space is $\mathcal{H}_y+\mathcal{H}_x$ when the propensity score is known. The corresponding score is $S(y,t,x;\theta_0)=S_y(y,t,x)+S_x(x)$. Define $EIF_{PS}\equiv\frac{D_t}{P_t(X)}\frac{\sum_{t'\in\mathcal{S}}P_{t'}(X)}{p_{\mathcal{S}}}\big(m(Y;\beta_t)-e_t(X;\beta_t)\big)+\frac{\sum_{t'\in\mathcal{S}}P_{t'}(X)}{p_{\mathcal{S}}}e_t(X;\beta_t)\in\mathcal{H}_y+\mathcal{H}_x$. It suffices to show $\mathbb{E}\big[(S_y(Y,T,X)+S_x(X))EIF_{PS}\big]=0$.

In the proof of Theorem 2, we obtain

$$\mathbb{E}\Big[\big(S_y(Y,T,X)+S_p(T,X)+S_x(X)\big)\big(EIF_{PS}+\sum_{t'\in\mathcal{S}}\frac{D_{t'}-P_{t'}(X)}{P_{t'}(X)}\frac{P_{t'}(X)}{p_{\mathcal{S}}}e_t(X;\beta_t)\big)\Big]=0$$

when the propensity score is unknown $\dot{P}_j(X)\neq 0$. It remains to show

(i) $\mathbb{E}\big[(S_y(Y,T,X)+S_x(X))\sum_{t'\in\mathcal{S}}\frac{D_{t'}-P_{t'}(X)}{P_{t'}(X)}\frac{P_{t'}(X)}{p_{\mathcal{S}}}e_t(X;\beta_t)\big]=0$ by the calculation in (13).

(ii)

$$\mathbb{E}\Big[S_p(T,X)EIF_{PS}\Big]=\mathbb{E}\left[\dot{P}_t(X)\frac{D_t}{P_t^2(X)}\frac{\sum_{t'\in\mathcal{S}}P_{t'}(X)}{p_{\mathcal{S}}}\big(m(Y;\beta_t)-e_t(X;\beta_t)\big)\right]=0$$

using (11) and $\sum_{j\in\mathcal{T}}\dot{P}_j(X)=0$.

(iii)

$$\begin{aligned}&\sum_{t'\in\mathcal{S}}\mathbb{E}\Big[S_p(T,X)\Big(\frac{D_{t'}(X)}{P_{t'}(X)}-1\Big)\frac{P_{t'}(X)}{p_{\mathcal{S}}}e_t(X;\beta_t)\Big]\\&=\sum_{t'\in\mathcal{S}}\mathbb{E}\Big[\frac{D_{t'}}{P_{t'}(X)}\frac{\dot{P}_{t'}(X)}{p_{\mathcal{S}}}e_t(X;\beta_t)\Big]=\mathbb{E}\Big[\frac{\sum_{t'\in\mathcal{S}}\dot{P}_{t'}(X)}{p_{\mathcal{S}}}e_t(X;\beta_t)\Big]\end{aligned}$$

is zero for two cases:

1. $\dot{P}_{t'}(X)=0$, i.e., the propensity score is known. So knowledge of the propensity score will affect the efficiency bound for estimating the treatment effects on the treated.
   When $\dot{P}_{t'}(X)=0$ only for $\mathcal{S}\subset\mathcal{T}$, we work on the same tangent space $\mathcal{H}_y+\mathcal{H}_p+\mathcal{H}_x$ with the score $S(y,t,x;\theta_0)=S_y(y,t,x)+S_p(t,x)+S_x(x)$ as in the proof of Theorem 2. In this case, EIF_{PS} is the efficient influence function by satisfying (i) and (iii).
2. $\mathcal{S}=\mathcal{T}$, i.e., the treatment effects for the population. We can also see from the following calculation: $\mathbb{E}\Big[S_p(T,X)\Big(\frac{D_t}{P_t(X)}m(Y;\beta_t)-\frac{D_t-P_t(X)}{P_t(X)}e_t(X;\beta_t)\Big)\Big]=0$ by $\sum_{j\in\mathcal{T}}\dot{P}_j(X)=0$, regardless $\dot{P}_j(X)$ equals zero.

To show the efficiency bound is reduced, we calculate the covariance

$$\mathbb{E}\Big[EIF_{PS}\Big(\frac{D_{t'}(X)}{P_{t'}(X)}-1\Big)\frac{P_{t'}(X)}{p_{\mathcal{S}}}e_t(X;\beta_t)\Big]=0. \qquad \square$$

C Proofs in Section 4 Estimation

The proofs in this section follow closely to the proofs of Theorem 1 and Theorem 2 in Lee (2014).

NOTATION. Let $(Z_1, Z_1, ..., Z_n)$ be an independent and identically distributed (*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 $\mathbb{E}\phi = \int \phi dP$ for the empirical process at ϕ. 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. Let C denote a generic constant.

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.

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 $\|g\|_\alpha = \max_{q.\le\underline{\alpha}} \sup_s |D^q g(s)| + \max_{q.\le\underline{\alpha}} \sup_{s\neq s'} |D^q g(s) - D^q g(s')|/\|s - s'\|^{\alpha-\underline{\alpha}}$ where $\max_{q.\le\underline{\alpha}}$ denotes the maximum over $(q_1, ..., q_d)$ such that $q. \le \underline{\alpha}$ 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 $\|g\|_\alpha \le M$.

Assumption 5 (Smoothness)

(i) *The data $\{Y_i, T_i, X_i\}$, $i = 1, ..., n$, is i.i.d.. The random vector $V = v_0(X)$ is a vector of measurable functions of X.*

(ii) *The support of V, $\mathcal{V}$, is a compact and convex subset of $\mathbb{R}^{d_v}$. (T, V) has a probability density function $f_{TV}(t, v)$, which is bounded away from zero and is Δ-order continuously differentiable with respect to v, 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 $F_{Y|TV}(y|t, v)$ is Δ-order continuously differentiable with respect to v, with uniformly bounded derivatives.*

(iv) *For each fixed $y \in \mathcal{Y}$ and $t \in \mathcal{T}$, $F_{Y|TV}(y|t, \cdot) \in \mathcal{C}_{\mathcal{M}}^\alpha(\mathcal{V})$ with $\alpha > d_v/2$.*

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

Assumption 6 (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 $r \ge 2$. (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.[5]

Assumption 7 (Nonparametric tuning parameters)
The bandwidth h satisfies (i) $h \to 0$, (ii) $nh^{2r} \to 0$, and (iii) $nh^{d_v+2\alpha}/\log(n) \to \infty$, as $n \to \infty$. The smoothness parameters in Assumptions 5 and 6 satisfy $\Delta \geq \alpha + r$, $\Delta_k \geq \alpha$, and $\alpha > d_v/2$.

Assumption 8 (Generated regressor)
The j-th component of v_0 satisfies $v_{0j}(X) \in \mathcal{C}_M^{\alpha_v}(\mathcal{X})$ with $\alpha_v > d_x/2$, for all $j = 1, ..., d_v$. Let the estimator $\|\hat{v}_j - v_{0j}\|_\infty = o_p(n^{-\delta})$, for all $j = 1, ..., d_v$. Let the second-step bandwidth $h \sim n^{-\eta}$ and the first-step bandwidth $h_1 \sim n^{-g}$ satisfying $0 < \eta < 1/(2d_v)$ and $0 < g < 1/(d_x + 2\alpha_v)$. [6]

The accuracy of the first-step generated regressor estimation satisfies $\max\{\eta + 1/4, \eta(1 + d_v/2)/(1 - d_x/(2\alpha_v))\} < \delta < (1 - gd_x)/2$. We choose a bias-reducing kernel with order $r_1 > 1/(2g) - d_x/2$. [7] *When the weight is estimated by $\hat{W} = \hat{P}_{t'}(X)/\hat{p}_{t'}$, further let $g < (1 - \eta d_v)/d_x$.*

We use a fixed trimming function that chooses a compact, interior subsupport of X such that the estimators $\hat{v}(X)$ and $\hat{F}_{Y|TV}(y|t, V)$ satisfies the uniform convergence rate in Result A.1 in Lee (2014). Therefore, the trimmed estimator consistently estimates $F_{Y|TV}(y|t, V)$ for the subpopulation whose observables X do not take extreme values. Then the third step uses this subsample with the second trimming function. That is, we work on a compact subsupport where the density functions are bounded away from zero, as in Assumption 5 (ii). So we can use the uniform linear representation

$$\hat{P}_t(x) - P_t(x) = \frac{1}{n}\sum_{i=1}^{n} \big(D_{ti} - P_t(x)\big) K_{h_1}(X_i - x)/f(x) + R_n^v(x) = O_p\big((nh_1^{d_x})^{-1/2}\big) \quad (16)$$

where $\|R_n^v\|_\infty = O_p\big(\big(\log n/(nh_1^{d_x})\big)\big)$ by choose the order of the kernel r_1. We might estimate the propensity score by other nonparametric estimator that admits a uniform linear representation.

C.1 Proofs in Sections 4.1 and 4.2

Proof of Theorem 4

We heuristically discuss the proof in the following. Define an infeasible estimator as if we knew the true W, $\hat{\beta}^0 = n^{-1}\sum_{i=1}^n \hat{F}_{Y|T\hat{V}}(y|t, \hat{V}_i)W_i$. We suppress the dependence on y in the notations for the estimand $\beta(y)$ and estimators for simplicity. Corollary 4 in Lee (2014)

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

[6] This condition comes from $\sup_x \|\frac{d^q}{dx^q}\hat{v}(x) - \frac{d^q}{dx^q}v_0(x)\| = o_p(1)$ for all $q \leq \alpha_v$. This ensures $Prob(\hat{v}_j \in \mathcal{C}_M^{\alpha_v}) \to 1$ as $n \to \infty$.

[7] It is often assumed undersmoothing for nonparametric estimation. An example of series estimation is in Hirano, Imbens, and Ridder (2003).

implies when the weight is not estimated,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\hat{\beta}^0-\beta\big)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_{ti}}{P_t(X_i)}\frac{P_{t'}(X_i)}{p_{t'}}\big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|TX}(y|t,X_i)\big)$$
$$+F_{Y|TX}(y|t,X_i)\frac{D_{t'i}}{p_{t'}}-\beta+F_{Y|TV}(y|t,V_i)\Big(W_i-\frac{D_{t'i}}{p_{t'}}\Big)+o_p(1). \quad (17)$$

Now consider the case when the weight is estimated. The estimation error from $\hat{p}_{t'}=n^{-1}\sum_{i=1}^{n}B_i$ contributes additional influence function

$$\beta-\beta\frac{B_i}{p_{t'}}. \quad (18)$$

For $W=P_{t'}(X)/p_{t'}$, an additional influence function comes from estimating $P_{t'}(X)$:

$$F_{Y|TV}(y|t,V_i)\Big(\frac{D_{t'i}}{p_{t'}}-W_i\Big). \quad (19)$$

Combining (17), (18), and (19), we derive Theorem 4.

The following is the detail of the proof. Denote $F_i=F_{Y|TV}(y|t,V_i)$ and $\hat{F}_i=\hat{F}_{Y|T\hat{V}}(y|t,\hat{V}_i)$. We decompose

$$\frac{1}{n}\sum_{i=1}^{n}\hat{F}_i\hat{W}_i-\mathbb{E}[F_iW_i]$$
$$=\frac{1}{n}\sum_{i=1}^{n}\hat{F}_iW_i-\mathbb{E}[F_iW_i]+\frac{1}{n}\sum_{i=1}^{n}F_i\big(\hat{W}_i-W_i\big)+\frac{1}{n}\sum_{i=1}^{n}\big(\hat{F}_i-F_i\big)\big(\hat{W}_i-W_i\big). \quad (20)$$

The first term is derived in Lee (2014) in the above (17). The third term is $\big((nh^{d_v})^{-1/2}+h^r\big)\|\hat{W}-W\|_\infty$ by the Assumptions $o_p(n^{-1/2})$. When $\hat{W}=\hat{P}_{t'}(X)/\hat{p}_{t'}$, we need $g<(1-\eta d_v)/d_x$.

WEIGHT. For estimating the weight $W=A/B$, linearize $\hat{W}_i-W_i=\hat{A}_i/\hat{B}-A_i/B=(\hat{A}_i-A_i)/B-(\hat{B}-B)A/B^2+O(|\hat{B}-B|^2+|\hat{B}-B|\|\hat{A}_i-A_i\|_\infty)$. The second term in (20) is

$$\frac{1}{n}\sum_{i=1}^{n}F_{Y|TV}(y|t,V_i)\big(\hat{W}_i-W_i\big)=\frac{1}{n}\sum_{i=1}^{n}F_{Y|TV}(y|t,V_i)\frac{\hat{A}_i-A_i}{B}-\frac{1}{n}\sum_{i=1}^{n}F_{Y|TV}(y|t,V_i)W_i\frac{\hat{B}-B}{B}$$
$$+O_p\big(|\hat{B}-B|^2+|\hat{B}-B|\|\hat{A}_i-A_i\|_\infty\big)$$

where the last term is $o_p(n^{-1/2})$ by $|\hat{B}-B| = |\hat{p}_t - p_t| = O_p(n^{-1/2})$. For the second term,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} F_{Y|TV}(y|t,V_i)W_i\frac{\hat{B}-B}{B} =& \ \mathbb{E}\left[F_{Y|TV}(y|t,V_i)\frac{W_i}{B}\right](\hat{B}-B) \\
&+\left(\frac{1}{n}\sum_{i=1}^{n} F_{Y|TV}(y|t,V_i)\frac{W_i}{B} - \mathbb{E}[F_{Y|TV}(y|t,V_i)]\frac{W_i}{B}\right)(\hat{B}-B) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{D_{t'i}}{p_{t'}}\beta - \beta + o_p(n^{-1/2}) \qquad (21)
\end{aligned}$$

when $\hat{B} = \hat{p}_{t'} = n^{-1}\sum_{i=1}^{n} D_{t'i}$.

For the first term, we use the stochastic equicontinuity argument in Theorem A.1 in Lee (2014). [8]

$$\frac{1}{n}\sum_{i=1}^{n} F_{Y|TV}(y|t,V_i)\frac{\hat{A}_i - A_i}{B} = \mathbb{E}\left[F_{Y|TV}(y|t,v_0(X))\frac{\hat{A}(X)-A(X)}{B}\right] + o_p(n^{-1/2}).$$

When $A_i = A(X_i) = P_{t'}(X_i)$, by (16),

$$\begin{aligned}
&\mathbb{E}\left[F_{Y|TV}(y|t,v_0(X))\frac{\hat{A}(X)-A(X)}{B}\right] \\
&= \mathbb{E}\left[F_{Y|TV}(y|t,v_0(X))\frac{1}{B}\frac{1}{n}\sum_{i=1}^{n}\big(D_{t'i} - P_{t'}(X)\big)\frac{K_{h_1}(X_i - X)}{f(X)}\right] + O_p\big(\|R_n^v\|_\infty\big) \\
&= \frac{1}{n}\sum_{i=1}^{n} F_{Y|TV}(y|t,V_i)\frac{1}{p_{t'}}\big(D_{t'i} - P_{t'}(X_i)\big) + O_p\big(h_1^{r_1} + \|R_n^v\|_\infty\big) \qquad (22)
\end{aligned}$$

where the last term is made $o_p(n^{-1/2})$.

Generated regressor. The key is that $Prob(T=t|V=v_0(X_i)) = P_t(X_i)$ and hence $\nabla_v P(T=t|V=v))\big|_{v=v_0(X_i)} = (1,0)^\top$. This determines the influence of estimating $P_t(X)$ as a generated regressor. Also $\mathbb{E}[W|X=X_i] = P_{t'}(X_i)/p_{t'} = \mathbb{E}[W|V=v_0(X_i)]$ and $\nabla_v \mathbb{E}\big[W\big|V=v\big]\big|_{v=v_0(X_i)} = \big(0, 1/p_{t'}\big)^\top$. This determines the influence of estimating $P_{t'}(X)$

[8] In Theorem A.1 in Lee (2014), let $f(y,X) = F_{Y|TV}(y|t,v_0(X))P_{t'}(X)$. For any fixed $y \in \mathcal{Y}$, $f(y,X) \in \mathcal{C}_M^\alpha(X)$ with $\alpha > d_x/2$. Then

$$\begin{aligned}
\sup_{y\in\mathcal{Y}}\Bigg|&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(F_{Y|TV}(y|t,v_0(X_i))\hat{P}_{t'}(X_i) - F_{Y|TV}(y|t,v_0(X_i))P_{t'}(X_i)\Big) \\
&- \sqrt{n}\,\mathbb{E}\Big[\Big(F_{Y|TV}(y|t,v_0(X))\hat{P}_{t'}(X) - F_{Y|TV}(y|t,v_0(X))P_{t'}(X)\Big)\Big]\Bigg| = o_p(1).
\end{aligned}$$

By $0 < g < 1/(d_x + 2\alpha_v)$ in Assumption 7, $\sup_{x\in\mathcal{X}}\big|\frac{\partial^\alpha}{\partial x^\alpha}\hat{P}_{t'}(x) - \frac{\partial^\alpha}{\partial x^\alpha}P_{t'}(x)\big| = O_p\big((nh_1^{d_x+2\alpha_v})^{-1/2} + h_1^{r_1}\big) = o_p(1)$. So $Prob\big(\hat{P}_{t'}(X) \in \mathcal{C}_M^\alpha(X)\big) \to 1$ as $n \to \infty$.

as generated regressor. Corollary 4 in Lee (2014) implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\hat{F}_{Y|T\hat{V}}(y|t,\hat{V}_i)W_i-\beta\Big)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(F_{Y|TV}(y|t,V_i)W_i-\mathbb{E}\big[F_{Y|TV}(y|t,V)W\big]$$
$$+\frac{D_{ti}}{P_t(X_i)}\Big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|TV}(y|t,V_i)\Big)\mathbb{E}\big[W\big|V=V_i\big]\Bigg)+\Delta_{ARG}+\Delta_{REG}+R_n$$

where $R_n=O_p\big(n^{-\kappa_1}+n^{-\kappa_2}+n^{-r\eta}\big)$, $0<\kappa_1<(1-d_v\eta)/2+(\delta-\eta)-\delta d_x/(2\alpha_v)$, and $\kappa_2<\min\big\{1-d_v\eta,2(\delta-\eta)\big\}$. The estimation errors associated with the generated regressor can be decomposed to two parts: for the argument in the known regression function

$$\Delta_{ARG}=\mathbb{E}\big[(\hat{v}(X)-v_0(X))'\nabla_V F_{Y|TV}(y|t,v)\big|_{v=v_0(X)}W\big]$$
$$=\frac{1}{n}\sum_{i=1}^{n}\big(D_{ti}-P_t(X_i),D_{t'i}-P_{t'}(X_i)\big)^{\top}\nabla_v F_{Y|TV}(y|t,v)\Big|_{v=v_0(X_i)}\mathbb{E}[W|X_i]+O_p(\|R_n^v\|_\infty+h_1^{r_1})$$

for the regressor that determines the regression function

$$\Delta_{REG}$$
$$=\frac{1}{n}\sum_{i=1}^{n}\big(D_{ti}-P_t(X_i),D_{t'i}-P_{t'}(X_i)\big)\Bigg\{-\nabla_v F_{Y|TV}(y|t,v)\Big|_{v=v_0(X_i)}\mathbb{E}\big[W\big|V=v_0(X_i)\big]$$
$$+\Big(F_{Y|TV}(y|t,v_0(X_i))-F_{Y|TX}(y|t,X_i)\Big)\Big(-\nabla_v\mathbb{E}\big[W\big|V=v\big]\Big|_{v=v_0(X_i)}\tag{23}$$
$$+\frac{\nabla_v Prob(T=t|V=v))\big|_{v=v_0(X_i)}}{P(T=t|V=v_0(X_i))}\mathbb{E}\big[W\big|V=v_0(X_i)\big]\Big)\Bigg\}\ \frac{P_t(X_i)}{Prob(T=t|V=v_0(X_i))}\tag{24}$$
$$+O_p(\|R_n^v\|_\infty+h_1^{r_1})$$
$$=\frac{1}{n}\sum_{i=1}^{n}\big(D_{ti}-P_t(X_i),D_{t'i}-P_{t'}(X_i)\big)\Bigg\{-\nabla_v F_{Y|TV}(y|t,v)\Big|_{v=v_0(X_i)}\frac{P_{t'}(X_i)}{p_{t'}}$$
$$+\Big(F_{Y|TV}(y|t,v_0(X_i))-F_{Y|TX}(y|t,X_i)\Big)\Big(\frac{P_{t'}(X_i)}{P_t(X_i)p_{t'}},-\frac{1}{p_{t'}}\Big)^{\top}\Bigg\}+O_p(\|R_n^v\|_\infty+h_1^{r_1}).\tag{25}$$

The role of the PS $\hat{P}_{t'}(X_i)$ is to recover the causal effects for the treated subpopulation. From the term $\nabla_v\mathbb{E}\big[W\big|V=v\big]$ in (23), regressing on the PS $\hat{P}_{t'}(X_i)$ contributes the term $\big(F_{Y|TX}(y|t,X_i)-F_{Y|TV}(y|t,v_0(X_i))\big)\big(D_{t'i}-P_{t'}(X_i)\big)/p_{t'}$. This term recovers (15) in the influence function . The term $\nabla_v Prob(T=t|V=v)$ in (24) comes from the partial mean that fixes the treatment level at t for the potential outcome. This term recovers (14) in the influence function. □

Proof of Theorem 5

From (25) in the Proof of Theorem 4, Δ_{REG} for $\hat{V}_i = (\hat{P}_t(X_i), P_{t'}(X_i))^\top$ is

$$\frac{1}{n}\sum_{i=1}^{n}\big(D_{ti} - P_t(X_i)\big)\Big\{ -\nabla_v F_{Y|TV}(y|t,v)\Big|_{v=v_0(X_i)}\frac{P_{t'}(X_i)}{p_{t'}} \\ + \Big(F_{Y|TV}(y|t,v_0(X_i)) - F_{Y|TX}(y|t,X_i)\Big)\frac{P_{t'}(X_i)}{P_t(X_i)p_{t'}}\Big\} + O_p(\|R_n^v\|_\infty + h_1^{r_1}). \quad (26)$$

The term associated with $\nabla_v \mathbb{E}\big[W\big|V=v\big]$ in (23) is dropped because the true $P_{t'}(X)$ is used. Corollary 4 in Lee (2014) implies that

$$\sqrt{n}\big(\hat{\beta}^0 - \beta\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_{ti}}{P_t(X_i)}\frac{P_{t'}(X_i)}{p_{t'}}\big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|TX}(y|t,X_i)\big) \\ + F_{Y|TX}(y|t,X_i)\frac{P_{t'}(X_i)}{p_{t'}} - \beta + F_{Y|TV}(y|t,V_i)\Big(W_i - \frac{P_{t'}(X_i)}{p_{t'}}\Big) + o_p(1).$$

By (21), the estimation error from $p_{t'}$ contributes an additional term to the influence function $\beta - \beta P_{t'}(X_i)/p_{t'}$. The result is derived. □

C.2 Proof of Theorem 6

(a) By the proof of Theorem 4, the estimation error from the weight contributes (21) and (22). Theorem 1 in Lee (2014) implies that $\sqrt{n}\big(\hat{\beta}_X - \beta\big)$ and $\sqrt{n}\big(\hat{\beta}_X^p - \beta\big)$ have the following linear representation

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}F_{Y|TX}(y|t,X_i)W_i - \beta + \frac{D_{ti}}{P_t(X_i)}\frac{P_{t'}(X_i)}{p_{t'}}\big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|TX}(y|t,X_i)\big) \\ + \beta - \beta\frac{D_{t'i}}{p_{t'}} + F_{Y|TX}(y|t,X_i)\Big(\frac{D_{t'i}}{p_{t'}} - W_i\Big) + o_p(1).$$

(b)

$$\sqrt{n}\big(\tilde{\beta}_x - \beta\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}F_{Y|TX}(y|t,X_i)\frac{P_{t'}(X_i)}{p_{t'}} - \beta \\ + \frac{D_{ti}}{P_t(X_i)}\frac{P_{t'}(X_i)}{p_{t'}}\big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|TX}(y|t,X_i)\big) + \beta - \beta\frac{P_{t'}(X_i)}{p_{t'}} + o_p(1).$$

□

C.3 Proofs in Sections 4.4 and 4.5

Proof of Theorem 7

Let $P_{\mathcal{S}}(x) = P_t(x) + P_{t'}(x)$. We first derive the uniform linear representation of the generated

regressor,

$$\hat{b}_t(x) - b_t(x) = \frac{\sum_{i=1}^n D_{ti} K_{h_1}(X_i - x)}{\sum_{i=1}^n (D_{ti} + D_{t'i}) K_{h_1}(X_i - x)} - \frac{P_t(x) f(x)}{P_{\mathcal{S}}(x) f(x)}$$
$$= \frac{n^{-1} \sum_{i=1}^n D_{ti} K_{h_1}(X_i - x)}{P_{\mathcal{S}}(x) f(x)} - \frac{b_t(x)}{P_{\mathcal{S}}(x) f(x)} \frac{1}{n} \sum_{i=1}^n (D_{ti} + D_{t'i}) K_{h_1}(X_i - x) + R_n^v(x)$$
$$= \frac{1}{n} \sum_{i=1}^n \frac{D_{ti} K_{h_1}(X_i - x) b_{t'}(x)}{P_{\mathcal{S}}(x) f(x)} - \frac{D_{t'} K_{h_1}(X_i - x) b_t(x)}{P_{\mathcal{S}}(x) f(x)} + R_n^v(x)$$

where $\|R_n^v\|_\infty = O_p(\log n/(n h_1^{d_x}))$. By the result of Corollary 4 in Lee (2014), the estimation error associated with $\hat{b}_t(X)$ is $\Delta_{ARG} + \Delta_{REG}$. Observe that $b_t(X) + b_{t'}(X) = 1$ and $\mathbb{E}[W | b_t(X), T \in \{t, t'\}] = (1 - b_t(X))/b_{t'}$. So $\frac{\partial}{\partial b} \mathbb{E}[W | b_t(X) = b, T \in \{t, t'\}] = -1/b_{t'}$. And $Prob(T = t | V = v, T \in \{t, t'\}) = b_t(x)$. So $\nabla_v Prob(T = t | V = v, T \in \{t, t'\})|_{v = b_t(x)} = 1$. In (23) and (24), $1/b_{t'} + (1 - b_t(X))/(b_{t'} b_t(x)) = 1/(b_t(x) b_{t'})$. So $\Delta_{ARG} + \Delta_{REG}$

$$= \mathbb{E}\left[\big(\hat{b}_t(X) - b_t(X)\big) \big(F_{Y|TV}(y|t, V(X)) - F_{Y|TX}(y|t, X)\big) \frac{1}{b_t(X) b_{t'}} \middle| T \in \{t, t'\} \right]$$
$$= \frac{1}{n} \sum_{i=1}^n \Big(\frac{D_{ti} b_{t'}(X_i)}{P_{\mathcal{S}}(X_i)} - \frac{D_{t'} b_t(X_i)}{P_{\mathcal{S}}(X_i)} \Big) \big(F_{Y|TV}(y|t, V(X_i)) - F_{Y|TX}(y|t, X_i)\big) \frac{1}{b_t(X_i) b_{t'}} \frac{P_{\mathcal{S}}(X_i)}{p_{\mathcal{S}}}$$
$$+ O_p(\|R_n^v\|_\infty + h_1^{r_1})$$
$$= \frac{1}{n} \sum_{i=1}^n \Big(D_{ti} \frac{P_{t'}(X_i)}{P_t(X_i)} - D_{t'} \Big) \big(F_{Y|TV}(y|t, V(X_i)) - F_{Y|TX}(y|t, X_i)\big) \frac{1}{p_{t'}} + O_p(\|R_n^v\|_\infty + h_1^{r_1})$$

that is the same as (25) in the proof of Theorem 4. □

Proof of Theorem 8

Let $P_{\mathcal{S}}(x) = P_t(x) + P_{t'}(x)$. We first derive the uniform linear representation of the generated regressor,

$$\tilde{b}_t(x) - b_t(x) = \frac{\sum_{i=1}^n D_{ti} K_{h_1}(X_i - x)}{\sum_{i=1}^n (D_{ti} + P_{t'}(x)) K_{h_1}(X_i - x)} - \frac{P_t(x) f(x)}{P_{\mathcal{S}}(x) f(x)}$$
$$= \frac{n^{-1} \sum_{i=1}^n D_{ti} K_{h_1}(X_i - x)}{P_{\mathcal{S}}(x) f(x)} - \frac{b_t(x)}{P_{\mathcal{S}}(x) f(x)} \frac{1}{n} \sum_{i=1}^n (D_{ti} + P_{t'}(x)) K_{h_1}(X_i - x) + R_n^v(x)$$
$$= \frac{n^{-1} \sum_{i=1}^n D_{ti} K_{h_1}(X_i - x) b_{t'}(x)}{P_{\mathcal{S}}(x) f(x)} - \frac{b_t(x) P_{t'}(x)}{P_{\mathcal{S}}(x) f(x)} \frac{1}{n} \sum_{i=1}^n K_{h_1}(X_i - x) + R_n^v(x)$$
$$= \frac{1}{n} \sum_{i=1}^n \frac{\big(D_{ti} - P_t(x)\big) K_{h_1}(X_i - x) b_{t'}(x)}{P_{\mathcal{S}}(x) f(x)} + R_n^v(x)$$

where $\|R_n^v\|_\infty = O_p(\log n/(n h_1^{d_x}))$. The third and forth equalities use $1 - b_t(x) = b_{t'}(x)$ and $b_t(x) P_{t'}(x) = b_{t'}(x) P_t(x)$.

Following the proof of Theorem 7, $\Delta_{ARG} + \Delta_{REG}$

$$
\begin{aligned}
&= \mathbb{E}\left[\big(\tilde{b}_t(X) - b_t(X)\big)\big(F_{Y|TV}(y|t,V(X)) - F_{Y|TX}(y|t,X)\big)\frac{1}{b_t(X)b_{t'}}\middle| T \in \{t,t'\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\big(D_{ti} - P_t(X_i)\big)b_{t'}(X_i)}{P_{\mathcal{S}}(X_i)}\big(F_{Y|TV}(y|t,V(X_i)) - F_{Y|TX}(y|t,X_i)\big)\frac{1}{b_t(X_i)b_{t'}}\frac{P_{\mathcal{S}}(X_i)}{p_{\mathcal{S}}} \\
&+ O_p(\|R_n^v\|_\infty + h_1^{r_1}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(D_{ti} - P_t(X_i)\big)\big(F_{Y|TV}(y|t,V(X_i)) - F_{Y|TX}(y|t,X_i)\big)\frac{P_{t'}(X_i)}{P_t(X_i)p_{t'}} + O_p(\|R_n^v\|_\infty + h_1^{r_1})
\end{aligned}
$$

that is the same as (26) in the proof of Theorem 5. □

Proof of Theorem 9

The key is the term $\nabla_v \mathbb{E}[W|V=v]$ in (23). When $V = P_1(X)$, $\nabla_v \mathbb{E}[W|V=v] = \nabla_v P_0(X)/p_0 = \nabla_v(1 - P_1(X))/p_0 = -1/p_0$. Using the uniform linear representation (16), $\Delta_{ARG} + \Delta_{REG}$

$$
\begin{aligned}
&= \mathbb{E}\left[\big(\hat{P}_1(X) - P_1(X)\big)\big(F_{Y|TV}(y|1,V(X)) - F_{Y|TX}(y|1,X)\big)\left(\frac{1}{p_0} + \frac{P_0(X)}{P_1(X)p_0}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(D_{1i} - P_1(X_i)\big)\big(F_{Y|TV}(y|1,V(X_i)) - F_{Y|TX}(y|1,X_i)\big)\frac{1}{P_1(X_i)p_0} + O_p(\|R_n^v\|_\infty + h_1^{r_1}).
\end{aligned}
$$

Comparing with (26) and following the proof of Theorem 5, we obtain $\check{\beta} - \beta = n^{-1}\sum_{i=1}^{n}\psi_t^{PS} + \big(D_{1i} - P_1(X_i)\big)\big(F_{Y|TV}(y|t,V(X_i)) - F_{Y|TX}(y|t,X_i)\big)\frac{1-P_0(X_i)}{P_1(X_i)p_0} + o_p(n^{-1/2})$. A simple algebra shows that the covariance of the second term and ψ_t^{PS} is zero. □

C.4 Proof of Theorem 10

Define the class of measurable functions $\mathcal{H} = \{\big(\mathcal{Y}\times\mathcal{T}\times\mathcal{X}\big) \to \psi(Y,T,X;y) : y \in \mathcal{Y}\}$. By Lemma A2 in Donald and Hsu (2014) and the Assumptions in the Appendix, $\mathcal{H}$ is P-Donsker. The weak convergence is implied by Donsker's Theorem in Section 2.8.2 in van der Vaart and Wellner (1996). We use the Functional Delta Method in Section 3.9 in van der Vaart and Wellner (1996). □

References

Abadie, A. and G. W. Imbens (2012). Matching on the estimated propensity score. Working paper.

Ao, W., S. Calonico, and Y.-Y. Lee (2014). Evaluating the effects of lengths of participation in the workforce investment act adult program via decomposition analysis. Working paper.

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

Busso, M., J. DiNardo, and J. McCrary (2014). New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators. *The Review of Economics and Statistics, forthcoming* (forthcoming).

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.

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

Dehejia, R. H. and S. Wahba (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economics and statistics 84*(1), 151–161.

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.

Frölich, M. (2004a, February). Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators. *The Review of Economics and Statistics 86*(1), 77–90.

Frölich, M. (2004b, 04). Programme Evaluation with Multiple Treatments. *Journal of Economic Surveys, Wiley Blackwell 18*(2), 181–224.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica 79*(2), 437–452.

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., R. J. LaLonde, and J. A. Smith (1999). The economics and econometrics of active labor market programs. *Handbook of labor economics 3*, 1865–2097.

Heckman, J. J. and E. J. Vytlacil (2007, January). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6 of *Handbook of Econometrics*, Chapter 70-71. Elsevier.

Hirano, K. and G. W. Imbens (2004). The propensity score with continuous treatments. In A. Gelman and X.-L. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pp. 73–84. New York: Wiley.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Imai, K. and D. van Dyk (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association 99*(467), 854–866.

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

Imbens, G. W. and J. M. Wooldridge (2009, September). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In M. Lechner and F. Pfeiffer (Eds.), *Econometric Evaluation of Labour Market Policies*, Volume 13 of *ZEW Economic Studies*, pp. 43–58. Physica-Verlag HD.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics 84*(2), 205–220.

Lee, Y.-Y. (2014). Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models. Working paper.

Mammen, E., C. Rothe, and M. Schienle (2014). Semiparametric estimation with generated covariates. working paper.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

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

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

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.