ISOTONIC PROPENSITY SCORE MATCHING

MENGSHAN XU AND TAISUKE OTSU

Abstract. We propose a one-to-many matching estimator of the average treatment effect based on propensity scores estimated by isotonic regression. The method relies on the monotonicity assumption on the propensity score function, which can be justified in many applications in economics. We show that the nature of the isotonic estimator can help us to fix many problems of existing matching methods, including efficiency, choice of the number of matches, choice of tuning parameters, robustness to propensity score misspecification, and bootstrap validity. As a by-product, a uniformly consistent isotonic estimator is developed for our proposed matching method.

1. Introduction

In both randomized experiments and observational studies, matching estimators are widely used to estimate treatment effects. This paper proposes a novel one-to-many propensity score matching method of the average treatment effect (ATE), where the propensity score is assumed to be monotone increasing in the exogenous covariate and is estimated by the isotonic regression. Our matching scheme is exact, i.e., for an outcome Y, treatment W, covariate X, and sample of size N, the matched set for the i-th unit is defined as

$$\mathcal{J}(i) = \{j = 1, \ldots, N : W_j = 1 - W_i \text{ and } \tilde{p}(X_j) = \tilde{p}(X_i)\},$$

where $\tilde{p}(\cdot)$ is a uniformly consistent isotonic estimator developed in Section 2.2. For multi-dimensional covariates X, we employ a monotone index model and consider the matched set:

$$\mathcal{J}(i) = \left\{j = 1, \ldots, N : W_j = 1 - W_i \text{ and } \tilde{p}_{\tilde{\alpha}}(X_j'\tilde{\alpha}) = \tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha})\right\},$$

where $\tilde{p}_{\tilde{\alpha}}(\cdot)$ is a uniformly consistent monotone single-index estimator developed in Section 3.

Surprisingly, the isotonic regression is particularly suitable to be the first stage nonparametric estimator in a two-stage semiparametric estimation of ATE. It renders features of both matching and weighting estimators to the second-stage ATE estimator and fixes at least five problems faced by the existing matching methods in the causal inference literature.

First, it is well known that the existing matching estimators of ATE with a fixed number of matches are inefficient (Abadie and Imbens, 2006) since they do not balance bias and variance accumulated in the second stage estimation. In comparison, our isotonic matching estimator is more efficient. In the univariate case, our method can achieve the semiparametric efficiency bound; in the multivariate case, where the efficiency bound becomes more complicated, we show that our proposed estimator performs better than ones that are based on a fixed number of

We are grateful to Markus Frölich, Daniel Gutknecht, Yoshi Rai, Christoph Rothe, Carsten Trenkler, and seminar participants at Mannheim for helpful comments and discussions.

matches and have propensity scores estimated by popular parametric models, such as probit and logit, which are widely adopted in applied work.

Second, although the performance of fixed-number matching estimators can be improved by increasing the number of matches with the sample size, the efficiency gain is somewhat artificial (Imbens, 2004) since the optimal number of matches and data-dependent ways of choosing it remain open questions. The isotonic estimator provides an elegant solution: It gives a piece-wise monotone increasing estimator, which partitions observations into different groups. Within these groups, the treated and untreated observations have the same estimated propensity scores, so they can be naturally matched to each other without choosing the number of matches, weights, and relevant distance measures (For our method, the distance is zero under any measure). In contrast, these choice problems are unavoidable in traditional methods for both covariates matching and propensity score matching, no matter whether they are based on the inverse variance matrix (e.g., Abadie and Imbens, 2006) or (empirical) density (e.g., Imbens, 2004) of covariates. Surprisingly, the set of the matching counterparts adaptively selected by isotonic estimator automatically becomes the optimal choice in the second stage, in that it achieves the semiparametric efficiency bound of ATE (Hahn, 1998) for the univariate case.

Third, compared to other semiparametric matching methods, where the first stage propensity score is estimated with kernel or series-based techniques, our method is tuning-parameter-free in a twofold sense. It is not only free from the choice of the optimal number of matches, as mentioned in the second point above, but also free from the choice of tuning parameters of conventional nonparametric methods, such as series length or bandwidth. In general, choosing tuning parameters of a first-stage nonparametric estimator remains a difficult open question in the semiparametric estimation literature. The traditional cross-validation fails in certain cases (e.g., in the case of ATE, where the second stage is not a minimum distance estimator) since the optimal first-stage estimator of the nuisance function does not imply the optimality of the second stage semiparametric estimation (Bickel and Ritov, 2003). The doubly-robust estimator (or debiased estimator) (Robins and Rotnitzky, 1995; Robins and Ritov, 1997; Scharfstein, Rotnitzky and Robins, 1999; Chernozhukov *et al.*, 2018; among others) partly solves this problem by introducing an additional unknown nuisance function, but estimating it (if nonparametrically) unavoidably adds new tuning parameters to the choice problem. At the cost of a monotonicity assumption imposed on the nuisance function, our proposed estimator avoids this choice problem while still maintaining other desirable properties of a decent semiparametric estimator.

Fourth, compared to popular parametric models of propensity scores, such as probit and logit, our proposed method contains a nonparametric first stage, so it is more robust to model misspecification. We acknowledge that combined with a single index structure, probit and logit models can also approximate many different data generating processes. But our method will always be more robust than them since both probit and logistic functions are monotone increasing themselves. In other words, the isotonic regression can well estimate all the data generating processes that can be well approximated by Probit or Logit model, but not vice versa. In addition, this robustness is achieved without costing the efficiency of the second-stage matching estimator.

Fifth, it is well known that the nonparametric bootstrap of the fixed-number matching estimator is invalid in the presence of continuous covariates (Abadie and Imbens, 2008). In the past decade, much work has tried to solve this problem by proposing cleverly structured wild bootstrap procedures. Otsu and Rai (2017) proposed a consistent wild bootstrap for covariates matching, and their approach was extended by Adusumilli (2020) and Bodory et al. (2016) to propensity score matching estimators. In our paper, we show that all these intricate bootstraps are no longer necessary in the case of monotone increasing propensity scores since the nonparametric bootstrap inference is asymptotically valid for our isotonic matching estimator.

Our method relies on the monotonicity assumption on propensity scores. Monotonicity is a natural shape restriction that can be justified in many applications in social science, economic studies, and medical research. Well-known examples in economics include the demand function, which is usually monotone decreasing in prices, and the supply or utility functions, which are often monotone increasing in quantities. Furthermore, many functions derived from cumulative distribution functions (CDF) inherit the monotonicity from the latter. For example, in a threshold crossing binary choice model

$$Y = \begin{cases} 1 & \text{if } X'\beta_0 > \varepsilon \\ 0 & \text{if } X'\beta_0 \leq \varepsilon \end{cases}, \tag{1}$$

the conditional expectation of Y on X can be written as $\mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X) = F_\varepsilon(X'\beta_0)$, where $F_\varepsilon(\cdot)$ is the CDF of an independent noise ε. If we assume $\varepsilon \sim N(0, 1)$, (1) becomes a probit model; if we assume $\varepsilon \sim \mathit{Logistic}(0, \frac{\pi^2}{3})$, it becomes a logit model. If we do not impose any distributional assumptions on ε, we can express (1) with a semiparametric model $Y = F_\varepsilon(X'\beta_0) + \nu$, with a nonparametric link function $F_\varepsilon(\cdot)$. It is monotone increasing by the nature of any CDF. See Cosslett (1983, 1987, 2007), Matzkin (1992), and Klein and Spady (1993) for more discussions of the model (1).

One of the main challenges of developing the asymptotic properties of the proposed estimator is the inconsistency of the isotonic estimator at its boundaries, sometimes called the "spiking" problem in the literature. If the dependent variable is binary, there is a non-trivial probability for a non-shrinking group of left-end estimates to be exactly zero even under the strict overlap condition, regardless of the sample size; the right-end estimates have the same issue. As a result, the matched sets for observations at two ends are empty, and we cannot construct a valid sample analog of ATE. Furthermore, observations near two ends are matched according to inconsistently estimated propensity scores, which are biased towards zero or one, resulting in a detrimental effect on the ATE estimator similarly to the one caused by limited overlaps (Khan and Tamer, 2010; Rothe, 2017). Although truncating those observations, whose propensity scores (either estimated parametrically or nonparametrically) are closer to 0 and 1, is widely implemented in applied work, this strategy has two caveats if one works with isotonic estimator. The first problem is the size of truncation: If too little was truncated, it might be insufficient to correct the boundary problem. A safe choice of truncation in the literature for different problems involving isotonic estimators is to truncate the first and last α_N-th quantile, with $\alpha_N \sim N^{-1/3}$ (or up to a logarithmic factor, see Wright, 1981; Durot, Kulikov and Lopuhaä, 2013; and Babbi and Kumar, 2021). However, this

truncation scheme is too much for our purpose. In fact, for any α_N such that $\alpha_N N^{1/2} \to \infty$, the truncated ATE estimator might be no longer $\sqrt{N}$-consistent.[1] Second, as discussed in Appendix A.6.1, one of the key conditions for $\sqrt{N}$-consistency and efficient estimation of ATE is (41) below, but whether this condition still holds after truncation is unclear. To solve these two problems, we extend the everywhere-consistent isotonic estimator of Meyer (2006) to a uniformly consistent one, which is by design to suit our two-stage semiparametric matching estimator. The proposed estimation procedure does not involve any truncation, the above-mentioned favorable properties of isotonic estimator remain intact, and the full set of data is utilized in both the first stage estimation of the propensity score and the second stage estimation of ATE.

Our proposed method builds on the large literature of causal inference for covariate and propensity score matching estimators, e.g., Rosenbaum and Rubin (1983, 1984), Rosenbaum (1989), Heckman, Ichimura and Todd (1997, 1998), Heckman, Ichimura, Smith and Todd (1998), Dehejia and Wahba (1999), Abadie and Imbens (2006, 2008, 2011, 2016), Imbens (2004), Frölich (2004), Frölich, Huber and Wiesenfarth (2017), Otsu and Rai (2017), Adusumilli (2020), Bodory, Camponovo, Huber and Lechner (2016), among others. The propensity score matching estimators studied in the literature mainly apply parametrically estimated propensity scores, such as probit and logit. Our proposed method, in contrast, uses a special type of nonparametric estimator, the isotonic estimator, to estimate the propensity score.

The isotonic estimator has a long history. The earlier work includes Ayer *et al.* (1955), Grenander (1956), Rao (1969, 1970), and Barlow and Brunk (1972), among others. The isotonic estimator of a regression function can be formulated as a least square estimation with monotonicity constraint. Suppose that the conditional expectation $\mathbb{E}[Y|X] = p_0(X)$ is monotone increasing, for an iid random sample $\{Y_i, X_i\}_{i=1}^N$, the isotonic estimator is the minimizer of the sum of squared errors, $\underset{p\in M}{min} \sum_{i=1}^N \{Y_i - p(X_i)\}^2$, where $\mathcal{M}$ is the class of monotone increasing functions. The minimizer can be calculated with the pool adjacent violators algorithm (Barlow and Brunk, 1972), or equivalently by solving the greatest convex minorant of the cumulative sum diagram $\{(0,0), (i, \sum_{j=1}^i Y_j), i = 1, \ldots, N\}$, where the corresponding $\{X_i\}_{i=1}^N$ are ordered sequence. See Groeneboom and Jongbloed (2014) for a comprehensive discussion of different aspects of isotonic regression.

Our work is linked to the vast literature on semiparametric estimation, e.g., Chamberlain (1987), Robinson (1988), Newey (1990, 1994), van der Vaart (1991), Andrews (1994), Hahn (1998), Ai and Chen (2003), Bickel and Ritov (2003), Chen, Linton, and Van Keilegom (2003), Chen and Santos (2018), Rothe and Firpo (2018), Chernozhukov *et al.* (2016, 2018), among others. In most of the works cited above, nonparametric methods involving smoothing parameters were applied at the initial stage, while our work uses the isotonic estimation that is non-smooth and tuning-parameter-free.

[1]This problem is not universal for every semiparametric estimator. For example, for a partially linear model $Y = X\beta + \psi(Z) + \varepsilon$, we can truncate more than its $N^{-1/2}$-th quantile, and the estimator of β maintains $\sqrt{N}$-consistency. In fact, one can get $\sqrt{N}$-rate even if β is estimated from an arbitrary sub-sample with a size proportional to N since different X's are linked to the same β. However, for ATE, in general, the truncated parts directly constitute estimation bias.

There are some authors working on concrete semiparametric models with plug-in isotonic estimators. Huang (2002) studied the properties of the monotone partially linear model, and his work was extended by Cheng (2009) and Yu (2014) to the monotone additive model. Balabdaoui, Durot, and Jankowski (2019) studied the monotone single index model with the monotone least square method, and Groeneboom and Hendrickx (2018), Balabdaoui, Groeneboom and Hendrickx (2019), and Balabdaoui and Groeneboom (2021) (these three papers are called BGH hereafter) developed a score-type approach for the monotone single index model and show the single index parameter can be estimated at $\sqrt{N}$-rate. Building on previous works, Xu (2021) studied a general framework of semiparametric Z-estimation with plug-in isotonic estimator, monotone single index estimator, or monotone additive estimator, and applied the generic result to inverse probability weighting (IPW) estimators of ATE. For the augmented IPW (AIPW) model, Qin *et al.* (2019) and Yuan, Yin and Tan (2021) applied the monotone single index model to estimate the propensity score, then plugged the estimated propensity scores with other estimates of potential outcomes into a doubly robust moment function. Their asymptotic results rely on the consistent estimations of both propensity scores and potential outcomes.

The rest of the paper is organized as follows. After introducing the setting and notations, Section 2 shows the implementation and asymptotic properties of the proposed isotonic matching estimator with a univariate covariate. These results are extended to the case of multivariate covariates, where the propensity score is modeled by a semiparametric single index model with an unknown monotone increasing link function. In Section 4, we establish the validity of nonparametric bootstrap. Monte Carlo simulation studies are presented in Section 5. All proofs are presented in Appendix.

2. Main results

2.1. **Setup and isotonic propensity score.** Suppose we observe the triple (Y, W, X) drawn randomly from the population, where $W \in \{0, 1\}$ is a binary treatment variable, $Y = W \cdot Y(1) + (1 - W) \cdot Y(0)$ is an outcome variable with potential outcomes $Y(1)$ and $Y(0)$ for $W = 1$ and 0, respectively, and X is a scalar covariate with continuous support $\mathcal{X} = [x_L, x_U] \subset \mathbb{R}$. In this section, we tentatively assume X is scalar, and discuss extensions for multivariate X in Section 3. Without loss of generality, let $\{Y_i, W_i, X_i\}_{i=1}^N$ be an iid sample of (Y, W, X) ordered by X, i.e., $X_1 < X_2 < \cdots < X_N$. Since X is assumed to be continuous, we rule out the case with ties (i.e., there is no pair $i \neq i'$ such that $X_i = X_{i'}$).

In this section, we consider estimation of the average treatment effect (ATE) $\tau = \mathbb{E}[Y(1) - Y(0)]$ by matching the propensity score $p(x) = \mathbb{P}(W = 1|X = x) = \mathbb{E}[W|X = x]$, where $p(\cdot)$ is an unknown monotone increasing function. In particular, we estimate $p(\cdot)$ by the isotonic estimator

$$\hat{p}(\cdot) = \underset{p \in \{\text{all monotone increasing functions on } X\}}{arg\,min} \sum_{i=1}^{N} \{W_i - p(X_i)\}^2. \tag{2}$$

It is known that $\hat{p}(\cdot)$ is a monotone increasing piecewise constant function with jump points $\{n_k\}_{k=1}^K$ for $K \leq N$. Based on these jump points, we split the support of X by $K + 1$ disjoint groups, and let N_k be the number of observations belonging to the k-th group. To avoid ambiguity

caused by splitting a flat piece into several sub-pieces with the same estimated value, we impose

$$\hat{p}(X_{n_1}) < \hat{p}(X_{n_2}) < \cdots < \hat{p}(X_{n_K}), \tag{3}$$

to ensure uniqueness of this partition (i.e., if $\hat{p}(X_{n_k}) = \hat{p}(X_{n_{k+1}})$, we simply combine the groups k and $k+1$). Note that $\{n_k\}_{k=1}^K$ are the first indices of these K partitions, $n_k + N_k = n_{k+1}$, and $\sum_{k=1}^K N_k = N$. The isotonic estimator $\hat{p}(\cdot)$ is characterized as follows.

Assumption 1. *[Sampling]* $\{Y_i, W_i, X_i\}_{i=1}^N$ *is an iid sample of* $(Y, W, X) \in \mathbb{R} \times \{0,1\} \times \mathcal{X}$, *where* $\mathcal{X} = [x_L, x_U] \in \mathbb{R}$. X *is continuously distributed and the sample is indexed according to* $X_1 < X_2 < \cdots < X_N$.

Proposition 1. *Under Assumption 1,* $\hat{p}(\cdot)$ *satisfying* (3) *partition the sample into* K *disjoint groups in the sense that for any* $k = 1, \ldots, K$ *and* $i = n_k, \ldots, n_k + N_k - 1$,

$$\hat{p}(X_i) = \frac{1}{N_k} \sum_{j=n_k}^{n_k+N_k-1} W_j. \tag{4}$$

To define our propensity score matching estimator based on $\hat{p}(\cdot)$, let $N_{k,1}$ and $N_{k,0}$ denote the numbers of treated and controlled observations within group k, i.e., $N_{k,1} = \sum_{i=n_k}^{n_k+N_k-1} W_i$ and $N_{k,0} = N_k - N_{k,1}$. Our one-to-many matching method is implemented within each of these $K+1$ groups, and each treated (controlled) observation in group k will be matched with its $N_{k,0}$ ($N_{k,1}$) counterparts, which belong the same group and have the same value of the estimated propensity score. The following results directly follow from Proposition (1).

Proposition 2. *Suppose Assumption 1 holds true.*

(i): *[Isotonic estimator] For any integer* $k = 1, \ldots, K$, *the isotonic estimator* $\hat{p}(\cdot)$ *for* $p(\cdot)$ *is represented as*

$$\hat{p}(x) = \frac{N_{k,1}}{N_k}, \tag{5}$$

for each $x \in \{X_i\}_{i=n_k}^{n_k+N_k-1}$.

(ii): *[Existence of matching counterparts] For any* $i = 1, \ldots, N$ *with* $0 < \hat{p}(X_i) < 1$, *the set of its matching counterparts* $\{j : W_j = 1 - W_i, \hat{p}(X_j) = \hat{p}(X_i)\}$ *is non-empty.*

Before we proceed, we need to solve the problem of potential lack of matching counterpart for those i's with $\hat{p}(X_i) = 0$ or 1. Under the strict overlaps (in Assumption 3 below), the problem is essentially associated with the inconsistency of the isotonic estimator at the boundary. In the next subsection, we propose a modified isotonic estimator that is uniformly consistent on $\mathcal{X}$.

2.2. **Uniformly consistent isotonic estimator.** Like other nonparametric estimators, the isotonic estimator is well known for being inconsistent at the boundary. If we apply the isotonic estimator to the binary dependent variable W, there is non-trivial probability of $\hat{p}(X_i) = 0$ or 1 even if the true propensity score $p(x)$ is bounded away from zero and one for all $x \in \mathcal{X}$. Then if $\hat{p}(X_1) = 0$, (5) implies $N_{1,1} = 0$, i.e., no matching counterpart for the treated units.

To fix this problem, we propose a modified isotonic estimator that is uniformly consistent on its support at $N^{-1/3}$ rate and is easy to implement. For the sample $\{W_i, X_i\}_{i=1}^N$ with $X_1 < \cdots <$

X_N, we transform $\{W_i\}_{i=1}^N$ into $\{\tilde{W}_i\}_{i=1}^N$ by averaging its first and last $\lfloor N^{2/3} \rfloor$ observations:

$$\tilde{W}_i = \begin{cases} \frac{1}{\lfloor N^{2/3} \rfloor} \sum_{i=1}^{\lfloor N^{2/3} \rfloor} W_i & \text{for } i \leq \lfloor N^{2/3} \rfloor \\ W_i & \text{for } \lfloor N^{2/3} \rfloor < i \leq N - \lfloor N^{2/3} \rfloor \\ \frac{1}{\lfloor N^{2/3} \rfloor} \sum_{i=N-\lfloor N^{2/3} \rfloor+1}^{N} W_i & \text{for } i > N - \lfloor N^{2/3} \rfloor \end{cases}. \tag{6}$$

Our uniformly consistent isotonic (hereafter, UC-isotonic) estimator is obtained by implementing the standard isotonic regression of $\tilde{W}$ on X:

$$\tilde{p}(x) = \begin{cases} \underset{s \leq i}{max} \underset{t \geq i}{min} \sum_{j=s}^{t} \tilde{W}_j/(t-s+1) & \text{for } x = X_i \\ \tilde{p}(X_i) & \text{for } X_{i-1} < x \leq X_i \\ \tilde{p}(X_N) & \text{for } x > X_N \end{cases}. \tag{7}$$

A similar modified estimator was proposed by Meyer (2006), where she averaged the first and last $\lceil \log(N) \rceil$ dependent variables instead of the first and last $\lfloor N^{2/3} \rfloor$ ones. The choices are different because she focuses on the consistency of isotonic estimator itself, while we are interested in the performance of the second-stage matching estimator. To achieve a $N^{-1/2}$ rate at the second stage, we need the isotonic estimator to be uniform consistent at a rate faster than $N^{-1/4}$, which won't be achieved under Meyer's choice. Meyer (2006) gives a theorem on the consistency of her modified estimator at the boundary, but she did not give a proof, nor did she discuss asymptotic properties of her estimator.

In this paper, we formally establish the uniform convergence rate of the modified isotonic estimator $\tilde{p}(\cdot)$. To this end, we impose the following assumption.

Assumption 2. *[Monotonicity and continuity] (i) $p(x) = \mathbb{E}[W|X = x]$ is a monotone increasing function of $x \in \mathcal{X}$, (ii) $p(x)$ is continuously differentiable with its first derivative $p^{(1)}(x) > 0$ for all $x \in \mathcal{X}$, and (iii) X has a continuous density $f(x)$ satisfying that for some positive constants $\overline{f}$ and $\underline{f}$, it holds $\underline{f} < f(x) < \overline{f}$ all $x \in \mathcal{X}$.*

Assumption 2 (i) is our main assumption, monotonicity of $p(\cdot)$. Assumption 2 (ii) contains smoothness conditions for $p(\cdot)$. Assumption 2 (iii) imposes an upper and lower bound for the density of X.

To avoid unnecessarily repeatedly defined notations, we let the same set of notations, K, $N_{k,1}$, N_k, and n_k, denote the number of groups, the number of treated observation in group k, the number of members in group k, and the index of the first element of group k, under the grouping scheme given by the *UC-isotonic* estimator $\tilde{p}(\cdot)$ and calculated with the *original* treatment variable $\{W_i\}_{i=1}^N$. We obtain an analogous result to Proposition 2 for the UC-isotonic estimator.

Proposition 3. *Under Assumptions 1 and 2, it holds*

(i): *$N_1 \geq \lfloor N^{2/3} \rfloor$ and $N_K \geq \lfloor N^{2/3} \rfloor$.*
(ii): *$\tilde{p}(x) = \frac{N_{k,1}}{N_k}$ for each $k = 1, \ldots, K$ and $x \in \{X_i\}_{i=n_k}^{n_k+N_k-1}$.*
(iii): *$N_1 = O_p(N^{2/3})$ and $N_K = O_p(N^{2/3})$.*

Part (i) of this proposition says that all the averaged W_i's at the beginning and end of the data are absorbed in the first and last groups. Part (ii) provides an analogous representation of the UC-isotonic estimator $\tilde{p}(\cdot)$ as $\hat{p}(\cdot)$. While Part (i) gives a lower bound of the lengths of the first and last partitions, Part (iii) gives (stochastic) upper bounds of them. Based on this proposition, the uniform convergence rate of the UC-isotonic estimator is obtained as follows.

Theorem 1. *Under Assumptions 1 and 2, it holds*

$$\sup_{x\in\mathcal{X}}|\tilde{p}(x)-p(x)|= O_p\left(\frac{\log N}{N}\right)^{1/3}.$$

Finally, to guarantee existence of matching counterparts by $\tilde{p}(\cdot)$, we impose the strict overlap condition.

Assumption 3. *[Strict overlaps] There exist positive constants $\underline{p}$ and $\bar{p}$ such that $0<\underline{p}\le p(x)\le \bar{p}<1$ for all $x\in\mathcal{X}$.*

Assumption 3 is standard in the treatment effect literature. It is necessary for the identification and $\sqrt{N}$-consistent estimation of the ATE. Combining Proposition 3 and Theorem 1 with Assumption 3, existence of the matching counterparts by $\tilde{p}(\cdot)$ is obtained as follows.

Corollary 1. *Suppose Assumptions 1-3 hold true. For each $i=1,\dots,N$, the set of its matching counterparts $\{j=1,\dots,N: W_j=1-W_i, \tilde{p}(x_j)=\tilde{p}(x_i)\}$ is non-empty with probability approaching one.*

2.3. **Isotonic propensity score matching.** Based on the UC-isotonic estimator $\tilde{p}(\cdot)$, the isotonic propensity score matching estimator for the ATE τ can be implemented as follows.

(1) Transform the sample $\{Y_i,W_i,X_i\}_{i=1}^N$ indexed by $X_1<\cdots<X_N$ into $\{Y_i,\tilde{W}_i,X_i\}_{i=1}^N$ by (6).
(2) Compute the UC-isotonic estimator $\tilde{p}(\cdot)$ by (7).
(3) For each $i=1,\dots,N$, compute the matching counterparts

$$\mathcal{J}(i)=\{j=1,\dots,N: W_j=1-W_i \text{ and } \tilde{p}(X_j)=\tilde{p}(X_i)\}. \tag{8}$$

(4) Calculate the matching estimator for the ATE τ by

$$\hat{\tau}=\frac{1}{N}\sum_{i=1}^{N}(2W_i-1)\left(Y_i-\frac{1}{M_i}\sum_{j\in\mathcal{J}(i)}Y_j\right), \tag{9}$$

where $M_i=|\mathcal{J}(i)|$ is the number of matches for i.

We proceed with the following assumptions.

Assumption 4. *[Potential outcomes] (i) $\mathbb{E}[Y(0)^2]<\infty$ and $\mathbb{E}[Y(1)^2]<\infty$, (ii) $\mathbb{E}[Y(0)|X=x]$ and $\mathbb{E}[Y(1)|X=x]$ are continuously differentiable for all $x\in\mathcal{X}$, and (iii) $Y(1),Y(0)\perp W|X$ almost surely.*

Assumption 4 (i)-(ii) regulate the tail behaviors of the (conditional) potential outcomes, which are necessary for $\sqrt{N}$-consistent estimation. Assumption 4 (iii) is the standard unconfoundedness assumption.

Under these assumptions, we have the following key equivalence result

Theorem 2. *Under Assumptions 1-4, the matching estimator $\hat{\tau}$ for the ATE τ using the UC-isotonic estimator $\tilde{p}(\cdot)$ is equivalent to the corresponding inverse probability weighting (IPW) estimator.*[2]

Remark on Theorem 2. Imbens (2004) pointed out that with $M \to \infty$ and $M/N \to 0$, the matching estimator is essentially like a regression estimator. In comparison, we find out that with propensity scores estimated by the UC-isotonic estimator, the (propensity score) matching estimator is numerically equivalent to the weighting estimator in each finite sample. This equivalence is essentially associated with the fact that the isotonic estimator can be regarded as a type of partitioning estimator (e.g., Györfi *et al.*, 2002; Cattaneo and Farrell, 2013), if the conditional means are estimated by the sample mean within each partition. For a two-stage matching or weighting estimator of ATE, the same set of partitions serves both the first- and second-stage nonparametric estimations. Usually, these two stages are not associated with each other since they have distinct objects, the propensity scores and the potential outcomes. Obviously, an ATE estimator based on propensity scores estimated with partitioning estimator should also exhibit a similar equivalence between matching and weighting estimator. However, one has to choose the number of partitions and their volume sizes, and they have to satisfy certain undersmoothing conditions to ensure a desirable asymptotic property of the second stage ATE estimator. In contrast, our proposed isotonic matching estimator automatically and optimally chooses these tuning parameters.

Moreover, as mentioned in the introduction, the equivalence result in Theorem 2 relies crucially on the implementation of the UC-isotonic estimator (7), which guarantees that both the matching and IPW estimators at the second stage are well-defined. □

Our main result, consistency and asymptotic normality of the isotonic propensity score matching estimator, is obtained as follows.

Theorem 3. *Under Assumptions 1-4, it holds $\hat{\tau} \xrightarrow{p} \tau$ and*

$$\sqrt{N}(\hat{\tau} - \tau) \xrightarrow{d} N(0, \Omega),$$

where $\Omega = \mathbb{V}(\mathbb{E}[Y(1) - Y(0)]|X) + \mathbb{E}[\mathbb{V}(Y(1)|X)/p(X)] + \mathbb{E}[\mathbb{V}(Y(0)|X)/(1 - p(X))]$.

We note that the asymptotic variance Ω is the semiparametric efficiency bound for τ. Also it is interesting to note that our estimator $\hat{\tau}$ is free from tuning constants, such as bandwidths or series lengths. Although we may conduct inference based on an estimator of Ω, we suggest a bootstrap inference method, which will be discussed in Section 4.

3. Multivariate covariates

Certainly, researchers are more interested in models with multivariate covariates X. One way to balance the robustness and the curse of dimensionality is to estimate the propensity score

[2]At almost the same time, Liu and Qin (2022) derive a similar equivalence result for the average treatment effect on treated (ATT) in an independent work.

with the monotone single index model:

$$W = p_0(X'\alpha_0) + \varepsilon, \qquad \mathbb{E}[\varepsilon|X] = 0, \tag{10}$$

where $p_0(\cdot)$ is a monotone increasing link function of its index $X'\alpha_0$ and $X \in \mathbb{R}^k$. For identification, α_0 is a k-dimensional vector normalized with $||\alpha_0||= 1$.[3]

For a binary dependent variable, this model can be derived from (1), and $p_0(\cdot)$ is by nature monotone increasing. It was studied by Cosslett (1983, 1987, 2007), Han (1987), Matzkin (1992), Sherman (1993), Klein and Spady (1993), among others. In the case where $p_0(\cdot)$ is estimated with isotonic regression, Balabdaoui, Durot, and Jankowski (2019) studied (10) with the monotone least square method, and Groeneboom and Hendrickx (2018), Balabdaoui, Groeneboom, and Hendrickx (2019), and Balabdaoui and Groeneboom (2021) (BGH) estimated α_0 and $p_0(\cdot)$ by solving a score-type sample moment condition:

$$\mathbb{E}[X\{W - p_0(X'\alpha_0)\}] = 0. \tag{11}$$

To estimate p_0 and α_0, we can apply the method of BGH. For a fixed α, define

$$\hat{p}_\alpha = \arg\min_{p\in\mathcal{M}} \frac{1}{N}\sum_{i=1}^{N}\{W_i - p(X_i'\alpha)\}^2, \tag{12}$$

where $\mathcal{M}$ is the set of monotone increasing functions defined on $\mathbb{R}$. Note that $\hat{p}_\alpha(u)$ can be solved with isotonic regression of W_i on the data points $\{X_i'\alpha\}_{i=1}^N$. Then α_0 can be estimated by minimizing the squared sum of a score function. For example, the simple score estimator in Balabdaoui and Groeneboom (2021) is given by solving

$$\hat{\alpha} = \arg\min_{\alpha} ||\frac{1}{N}\sum_{i=1}^{N} X_i'\{W_i - \hat{p}_\alpha(X_i'\alpha)\}||^2. \tag{13}$$

BGH showed that under certain assumptions, $\hat{\alpha}$ is a $\sqrt{N}$-consistent estimator for α_0, and $\mathbb{E}[\hat{p}_{\hat{\alpha}}(X'\hat{\alpha}) - p_0(X'\alpha_0)] = O_P((\log N)N^{-2/3})$. We apply their method to estimate the propensity score with multi-dimensional control variables X.

In this section, $\tilde{\tau}$ denotes the ATE estimator based on the multi-dimensional covariates X. Similarly to Section 2.2, to solve the boundary problem of isotonic estimator to ensure that each observation has a non-empty matched set, we develop a uniformly consistent monotone single-index (hereafter, UC-iso-index) estimator, denoted $\tilde{p}_{\tilde{\alpha}}$. The matching procedure can be implemented as follows.

(1) Compute $\hat{\alpha}$ by (12) and (13).
(2) Let $\tilde{\alpha} = \hat{\alpha}$, and transform the sample $\{Y_i, W_i, X_i\}_{i=1}^N$ indexed by $X_1'\tilde{\alpha} < \cdots < X_N'\tilde{\alpha}$ into $\{Y_i, \tilde{W}_i, X_i\}_{i=1}^N$ with (6).

[3]In the estimation, the constraint $||\alpha_0||= 1$ can be dealt with reparametrization or the augmented Lagrange method by Balabdaoui and Groeneboom (2021). In this section, we study our model without discussing those technical details. See BGH for more details.

(3) Compute the UC-iso-index estimator $\tilde{p}_{\tilde{\alpha}}$ by

$$\tilde{p}_{\tilde{\alpha}} = \arg\min_{p \in \mathcal{M}} \frac{1}{N} \sum_{i=1}^{N} \{\tilde{W}_i - p(X_i'\tilde{\alpha})\}^2.$$

(4) For each $i = 1, \ldots, N$, compute the matching counterparts

$$\mathcal{J}(i) = \left\{ j = 1, \ldots, N : W_j = 1 - W_i \text{ and } \tilde{p}_{\tilde{\alpha}}(X_j'\tilde{\alpha}) = \tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha}) \right\}. \tag{14}$$

(5) Calculate the matching estimator for the ATE τ by

$$\tilde{\tau} = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1) \left(Y_i - \frac{1}{M_i} \sum_{j \in \mathcal{J}(i)} Y_j \right), \tag{15}$$

where $M_i = |\mathcal{J}(i)|$ is the number of matches for i.

We modify Assumptions 1-4 in Section 2 as follows.

Assumption 1'. *[Sampling]* $\{Y_i, W_i, X_i\}_{i=1}^N$ *is an iid sample of* $(Y, W, X) \in \mathbb{R} \times \{0, 1\} \times \mathcal{X}$*, where the space* $\mathcal{X}$ *is a convex subset of* $\mathbb{R}^k$ *with nonempty interior. There exists* $R > 0$ *such that* $\mathcal{X} \subset \mathcal{B}(0, R) := \{x : ||x|| \leq R\}$.

Given α, we define the true link function of (12):

$$p_\alpha(u) = \mathbb{E}[W | X'\alpha = u].$$

Obviously, $p_{\alpha_0} = p_0$. Let a_0 and b_0 be the minimum and the maximum of the interval $I_{\alpha_0} = \{x'\alpha_0 : x \in \mathcal{X}\}$, respectively.

Assumption 2'. *[Monotonicity and continuity] (i) There exists* $\delta_0 > 0$ *such that for each* $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$*, the function* $u \mapsto \mathbb{E}[W | X'\alpha = u]$ *is monotone increasing in* u *and differentiable in* α*; (ii)* $p_0(\cdot)$ *is continuously differentiable with its first derivative* $p^{(1)}(u) > 0$ *on* $u \in (a_0 - \delta_0 R, b_0 + \delta_0 R)$*, and (iii)* X *has a continuous density* $f(x)$ *satisfying that for some positive constants* $\overline{f}$ *and* $\underline{f}$*, it holds* $\underline{f} < f(x) < \overline{f}$ *all* $x \in \mathcal{X}$.

Assumption 3'. *[Strict overlaps] There exist positive constants* $\underline{p}$ *and* $\bar{p}$ *such that* $0 < \underline{p} \leq p_0(x'\alpha_0) \leq \bar{p} < 1$ *for all* $x \in \mathcal{X}$.

Assumption 4'. *[Potential outcomes] (i)* $\mathbb{E}[Y(0)^2] < \infty$ *and* $\mathbb{E}[Y(1)^2] < \infty$*, (ii)* $u \mapsto \mathbb{E}[Y(1)|X = x]$ *are continuously differentiable for all* $x \in \mathcal{X}$ *and* $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$*, and (iii)* $Y(1), Y(0) \perp W | X$ *almost surely.*

The following assumptions are adapted from BGH, which ensure that the score estimator (12) and (13) have desirable properties.

Assumption 5. *For all* $\alpha \neq \alpha_0$ *such that* $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$*, the random variable* $Cov[(\alpha - \alpha_0)'X, p_0(X'\alpha_0) | X'\alpha]$ *is not equal to 0 almost surely.*

Assumption 6. *[Potential outcomes] Let* $p_0^{(1)}(u)$ *denote the first derivative of* $p_0(u)$*. The matrix* $\mathbb{E}[p_0^{(1)}(X'\alpha_0) Cov(X | X'\alpha_0)]$ *has rank* $k - 1$.

Based on Assumptions 1', 2', 5, and 6, we have a result similar to Proposition 3 with respect to the numbering according to $X_1'\tilde{\alpha} < \cdots < X_N'\tilde{\alpha}$, and the uniform convergence rate of the UC-iso-index estimator is obtained as follows.

Theorem 4. *Under Assumptions 1', 2', 5, and 6, it holds*

$$\sup_{x\in\mathcal{X}}|\tilde{p}_{\tilde{\alpha}}(x'\tilde{\alpha}) - p_0(x'\alpha_0)| = O_p\left(\frac{\log N}{N}\right)^{1/3}.$$

The existence of matching counterparts is guaranteed by an argument similar to Corollary 1.

Finally, let Z be the triple (Y, W, X), and B^- be the Moore-Penrose inverse of a square matrix B. The asymptotic properties of the isotonic propensity score matching estimator is obtained as follows.

Theorem 5. *Under Assumptions 1'-4', 5, and 6, it holds* $\tilde{\tau} \xrightarrow{p} \tau$ *and*

$$\sqrt{N}(\tilde{\tau} - \tau) \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma = \mathbb{E}[\{m(Z) + M(Z) + A(Z)\}\{m(Z) + M(Z) + A(Z)\}']$, *and*

$$\begin{aligned}
m(Z) &= \frac{YW}{p_0(X'\alpha_0)} - \frac{Y(1-W)}{1-p_0(X'\alpha_0)} - \tau, \qquad D(Z) = -\left(\frac{YW}{p_0(X'\alpha_0)^2} + \frac{Y(1-W)}{(1-p_0(X'\alpha_0))^2}\right), \\
M(Z) &= -\mathbb{E}[D(Z)|X'\alpha_0]\{W - p_0(X'\alpha_0)\} \\
A(Z) &= \mathbb{E}[\{D(Z) - \mathbb{E}(D(Z)|X'\alpha_0)\}\{X - \mathbb{E}[X|X'\alpha_0]\}'p_0^{(1)}(X'\alpha_0)] \\
&\quad \times \mathbb{E}[p_0^{(1)}(X'\alpha_0)\,Cov(X|X'\alpha_0)]^-\{X - \mathbb{E}[X|X'\alpha_0]\}\{W - p_0(X'\alpha_0)\}. \qquad (16)
\end{aligned}$$

Note that based on Newey (1994), the semiparametric efficiency bound for estimating τ with known α_0 is given by $\mathbb{E}[\{m(Z) + M(Z)\}\{m(Z) + M(Z)\}']$. The additional term $A(Z)$ can be interpreted as an influence of estimating the index coefficients α_0. In Section 5.2 below, we present a simulation result to illustrate that the proposed ATE estimator $\tilde{\tau}$ outperforms the probit matching estimator in every sample size, even for the case that the true propensity score is a probit (the correct specification). We also note that $\tilde{\tau}$ is free from tuning constants, such as bandwidths or series lengths.

4. Bootstrap inference

The asymptotic variances in Theorems 3 and 5 contain conditional mean and variance functions, such as $\mathbb{V}(Y(1)|X)$ and $\mathbb{E}[X|X'\alpha_0]$, which need to be estimated. If we use nonparametric methods to estimate them, we still need to choose some tuning parameters even though the point estimators are free from tuning. To obtain a tuning free inference method, we employ a bootstrap method to approximate the asymptotic distribution of the proposed isotonic propensity score matching estimator.

After Abadie and Imbens (2008) showed that the nonparametric bootstrap of the fixed-number matching estimator is invalid in the presence of continuous covariates, much work tried to solve this problem by proposing modified wild bootstraps, including Otsu and Rai (2017) for covariates matching estimators, and Adusumilli (2020) and Bodory *et al.* (2016) for propensity score matching estimators. In contrast, the nonparametric bootstrap of our one-to-many matching

method is valid, which is an interesting implication of Theorem 2. In this section, we discuss an asymptotically valid bootstrap procedure for the estimator $\hat{\tau}$ in Theorem 3. This result can be similarly adapted to $\tilde{\tau}$ in Theorem 5.

The nonparametric bootstrap is implemented as follows.

(1) $\{Y_i^*, W_i^*, X_i^*\}_{i=1}^N$ is a resample with replacement from $\{Y_i, W_i, X_i\}_{i=1}^N$, and the numbering is according to $X_1^* \leq \cdots \leq X_N^*$.
(2) $\tilde{p}^*(\cdot)$ is the UC-isotonic estimator based on $\{Y_i^*, W_i^*, X_i^*\}_{i=1}^N$.
(3) The bootstrap counterpart $\hat{\tau}^*$ of $\hat{\tau}$ is given by

$$\begin{aligned} \hat{\tau}^* &= \frac{1}{N}\sum_{i=1}^{N}(2W_i^* - 1)\left(Y_i^* - \frac{1}{M_i^*}\sum_{j\in\mathcal{J}^*(i)} Y_j^*\right), \\ \mathcal{J}^*(i) &= \left\{j = 1, \ldots, N : W_j^* = 1 - W_i^* \text{ and } \tilde{p}^*(X_j^*) = \tilde{p}^*(X_i^*)\right\}, \end{aligned}$$

where $M_i^* = |\mathcal{J}^*(i)|$ is the number of matches for the i-th observation in the bootstrap resample.
(4) After repeating Step (1)-(3) for B times and obtaining estimator $\hat{\tau}_1^*, \hat{\tau}_2^*, \ldots, \hat{\tau}_B^*$, we can conduct inference for τ.

The asymptotic validity of this bootstrap approximation is obtained as follows.

Theorem 6. *Under Assumptions 1-4,*

$$\sup_{t\in\mathbb{R}} |\mathbb{P}^*\{\sqrt{n}(\hat{\tau}^* - \hat{\tau}) \leq t\} - \mathbb{P}\{\sqrt{n}(\hat{\tau} - \tau) \leq t\}| \xrightarrow{p} 0,$$

where $\mathbb{P}^*$ is the bootstrap distribution conditional on the data.

5. Monte Carlo simulations

In this section, we conduct two simulation studies to assess the finite sample properties of our isotonic propensity score matching estimator.

5.1. **Univariate case.** Let $X = 0.15 + 0.7Z$, where Z and ν are independently uniformly distributed on $[0, 1]$, and

$$\begin{aligned} W &= \begin{cases} 0 & \text{if } X < \nu \\ 1 & \text{if } X \geq \nu \end{cases}, \\ Y &= 0.5W + 2X + \varepsilon, \\ \varepsilon &\sim N(0, 1). \end{aligned}$$

The true ATE is the coefficient of T, which is 0.5. The simulation results are presented in Table 1, where $\hat{\mu}_\tau$ is the Monte-Carlo mean, and the mean square errors (MSE) are rescaled by N. The number of Monte-Carlo simulations is 5000 for each sample size.

TABLE 1. Matching estimators of ATE: univariate case

with UC-isotonic			with logit and $M = 1$		
N	$\hat{\mu}_\tau$	MSE	N	$\hat{\mu}_\tau$	MSE
100	0.4977	5.2723	100	0.4997	7.1068
1000	0.4934	5.2589	1000	0.5009	7.0630
2000	0.4946	5.2158	2000	0.4999	7.0816
5000	0.4963	4.9418	5000	0.4995	6.8376
10000	0.4974	4.9785	10000	0.5000	6.8238
∞	0.5	4.94	∞	0.5	4.94

The left panel shows the simulation results of the proposed matching method based on propensity scores estimated by the UC-isotonic estimator, and the right panel shows those of the one-to-one matching estimator based on propensity scores estimated with the logit model $\mathbb{P}(W = 1|X = x) = \frac{\exp(a+bx)}{\exp(a+bx)+1}$. The last row shows the true value of ATE and the semiparametric efficiency bound of this problem calculated according to Hahn (1998).

In comparison, the logit matching estimator has a slightly smaller bias. Since our proposed UC-isotonic estimator does not truncate the data, the bias of our isotonic matching estimator is also small and is converging to zero. The MSEs of the isotonic propensity score matching estimator are considerably smaller than those of the logit matching estimator in every sample size. With the sample size growing, the MSEs of isotonic propensity score matching estimator approaches to the semiparametric efficiency bound.

5.2. **Multivariate case.** Consider the following setting:

$$\begin{aligned} Y &= X'\gamma_0 + W\tau_0 + \varepsilon, \\ W &= \begin{cases} 0 & \text{if } X'\alpha_0 < \nu \\ 1 & \text{if } X'\alpha_0 \geq \nu \end{cases}, \\ \varepsilon &\sim N(0,1), \qquad \nu \sim N(0,1), \qquad \varepsilon \perp \nu, \end{aligned}$$

where $X \sim U[-1,1]^3$, and the true parameters are set as $\alpha_0 = (1,1,1)'/\sqrt{3}$, and $\gamma_0 = (0.1, 0.2, 0.3)'$, and the ATE is $\tau_0 = 0.5$. Under this setting, we have $\mathbb{P}(W = 1|X = x) = p_0(x) = \Phi(x'\alpha_0)$, where Φ is the CDF of the standard normal distribution, i.e., the propensity score is *correctly* specified in probit estimation.

The simulation results are presented in Table 2, where $\hat{\mu}_\tau$ is the Monte-Carlo mean, and the MSEs are rescaled by N. The number of Monte-Carlo simulations is 5000 for each sample size. The left panel shows the simulation results of the proposed matching method based on propensity scores estimated by the UC-iso-index estimator, and the right panel shows those of one-to-one matching estimator based on propensity scores estimated with the correctly specified probit model.

TABLE 2. Matching estimators of ATE: multivariate case

with UC-iso-index			with probit and $M = 1$		
N	$\hat{\mu}_\tau$	MSE	N	$\hat{\mu}_\tau$	MSE
100	0.5080	5.0442	100	0.5114	7.3459
1000	0.5016	5.0014	1000	0.5030	6.9813
2000	0.4991	5.0727	2000	0.4997	7.2275
5000	0.5003	5.2115	5000	0.5010	7.2640
10000	0.5001	5.0161	10000	0.5002	7.0509

The pattern is similar to the univariate case. The isotonic matching estimator outperforms the probit matching estimator in every sample size in terms of MSE. At the same time, our proposed method has a very small bias. For larger sample sizes, its Monte-Carlo bias is even slightly smaller than that of one-to-one probit matching, and it converges to zero.

Overall, the simulation results support our proposed method.

6. Conclusion

We develop a one-to-many matching estimator of ATE based on propensity scores estimated by modified isotonic regression. We reveal that the nature of isotonic estimator can help us to fix many problems of existing matching methods, including efficiency, choice of the number of matches, choice of tuning parameter, robustness to the propensity score misspecification, and bootstrap validity. As by-products, a uniformly consistent isotonic estimator and a uniformly consistent monotone single index estimator, for both univariate and multivariate cases, are designed for our proposed isotonic matching estimator, and we study their asymptotic properties. The method can be further extended to other causal estimators based on propensity scores, such as blocking on propensity scores and regression on propensity scores.

Appendix A. Proofs

A.1. **Proof of Proposition 1.** The proof is based on the following lemma.

Lemma 1. *[Groeneboom and Jongbloed (2014, Lemma 2.1)] The vector $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_N)$ minimizes $Q(p) = \frac{1}{2}\sum_{i=1}^N (W_i - p_i)^2$ over the closed convex cone $\mathcal{C} = \{p \in \mathbb{R}^N : p_1 \le p_2 \le \cdots \le p_N\}$ if and only if*

$$\sum_{j=1}^{i} \hat{p}_j \begin{cases} \le \sum_{j=1}^{i} W_j \\ = \sum_{j=1}^{i} W_j \quad \text{if } \hat{p}_{i+1} > \hat{p}_i \text{ or } i = N \end{cases}. \tag{17}$$

We now prove Proposition 1. For any $k = 1, \ldots, K$, we have $\hat{p}_{n_k} > \hat{p}_{n_k - 1}$ and $\hat{p}_{n_{k+1}} > \hat{p}_{n_{k+1}-1}$ by (3). By (17), we have

$$\sum_{j=1}^{n_k-1} \hat{p}_j = \sum_{j=1}^{n_k-1} W_j, \qquad \sum_{j=1}^{n_{k+1}-1} \hat{p}_j = \sum_{j=1}^{n_{k+1}-1} W_j, \qquad \hat{p}_{n_k} = \hat{p}_{n_k+1} = \cdots = \hat{p}_{n_k+N_k-1}. \tag{18}$$

Since $n_k - 1 = n_{k-1} + N_{k-1} - 1$ and $n_{k+1} - 1 = n_k + N_k - 1$, (18) implies

$$\sum_{j=n_k}^{n_k+N_k-1} \hat{p}_j = \sum_{j=n_k}^{n_k+N_k-1} W_j. \tag{19}$$

Combining (18) and (19), it holds that for any $i = n_k, \ldots, n_k + N_k - 1$,

$$\hat{p}_i = \frac{1}{N_k} \sum_{j=n_k}^{n_k+N_k-1} \hat{p}_j = \frac{1}{N_k} \sum_{j=n_k}^{n_k+N_k-1} W_j.$$

A.2. **Proof of Proposition 2.** Part (i) is a direct implication of Proposition 1 and the definitions of $N_{k,1}$, N_k, and n_k. By $W_i \in \{0, 1\}$, $0 < \hat{p}(X_i) < 1$, and (4), we must have $W_i = 1$ and $W_j = 0$ for some $i, j \in \{n_k, \ldots, (n_k + N_k - 1)\}$. Thus, Part (ii) follows.

A.3. **Proof of Proposition 3.**

Proof of (i). Since the proof is similar, we focus on the proof of the first statement, $N_1 \geq \lfloor N^{2/3} \rfloor$. The isotonic estimator is written as (see, Barlow and Brunk, 1972)

$$\hat{p}(X_i) = \max_{s \leq i} \min_{t \geq i} \sum_{j=s}^{t} \frac{W_j}{t - s + 1}. \tag{20}$$

Let

$$\bar{W}_l = \frac{1}{\lfloor N^{2/3} \rfloor} \sum_{i=1}^{\lfloor N^{2/3} \rfloor} W_i, \qquad \bar{W}_u = \frac{1}{\lfloor N^{2/3} \rfloor} \sum_{i=N-\lfloor N^{2/3} \rfloor+1}^{N} W_i. \tag{21}$$

For any i with $1 \leq i \leq \lfloor N^{2/3} \rfloor$, (6), (7), and (20) imply

$$\begin{aligned}
\tilde{p}(X_i) &= \max_{s \leq i} \min_{t \geq i} \sum_{j=s}^{t} \frac{\tilde{W}_j}{t - s + 1} \\
&= \max_{s \leq i} \min_{t \geq i} \frac{\sum_{j=s}^{t \wedge \lfloor N^{2/3} \rfloor} \bar{W}_l + \mathbb{I}\{t > \lfloor N^{2/3} \rfloor\} \sum_{j=\lfloor N^{2/3} \rfloor+1}^{t} W_j}{t - s + 1} \\
&= \max_{s \leq i} \min_{t \geq i} \frac{\sum_{j=s}^{t} \bar{W}_l + \mathbb{I}\{t > \lfloor N^{2/3} \rfloor\} \left(\sum_{j=\lfloor N^{2/3} \rfloor+1}^{t} W_j - \sum_{j=\lfloor N^{2/3} \rfloor+1}^{t} \bar{W}_l \right)}{t - s + 1} \\
&= \bar{W}_l + \max_{s \leq i} \min_{t \geq i} \left[\mathbb{I}\{t > \lfloor N^{2/3} \rfloor\} \frac{t - \lfloor N^{2/3} \rfloor}{t - s + 1} \left(\frac{1}{t - \lfloor N^{2/3} \rfloor} \sum_{j=\lfloor N^{2/3} \rfloor+1}^{t} W_j - \bar{W}_l \right) \right]
\end{aligned} \tag{22}$$

Since $\mathbb{I}\{t > \lfloor N^{2/3} \rfloor\} \frac{t - \lfloor N^{2/3} \rfloor}{t - s + 1} \geq 0$, the minimizer with respect to t is determined by the sign of $\left(\frac{1}{t - \lfloor N^{2/3} \rfloor} \sum_{j=\lfloor N^{2/3} \rfloor+1}^{t} W_j - \bar{W}_l \right)$, and we split into two cases:

(I) $\min_{t > \lfloor N^{2/3} \rfloor} \frac{1}{t - \lfloor N^{2/3} \rfloor} \sum_{j=\lfloor N^{2/3} \rfloor+1}^{t} W_j > \bar{W}_l$,

(II) $\min_{t > \lfloor N^{2/3} \rfloor} \frac{1}{t - \lfloor N^{2/3} \rfloor} \sum_{j=\lfloor N^{2/3} \rfloor+1}^{t} W_j \leq \bar{W}_l$.

For Case (I), adding any terms after $\lfloor N^{2/3} \rfloor$ cannot make the average smaller, so we must have $\tilde{p}(X_i) = \bar{W}_l$ for all $1 \leq i \leq \lfloor N^{2/3} \rfloor$. Thus, it holds $N_1 = \lfloor N^{2/3} \rfloor$.

For Case (II), it makes sense to add more terms after $\lfloor N^{2/3} \rfloor$ since for any fixed s, adding more item after $\lfloor N^{2/3} \rfloor$ will lower the overall level of the sample mean (22). Define

$$t_s = \arg \min_{t \geq \lfloor N^{2/3} \rfloor} \mathbb{I}\{t > \lfloor N^{2/3} \rfloor\} \frac{t - \lfloor N^{2/3} \rfloor}{t - s + 1} \left(\frac{1}{t - \lfloor N^{2/3} \rfloor} \sum_{j=\lfloor N^{2/3} \rfloor + 1}^{t} W_j - \bar{W}_l \right). \tag{23}$$

After minimizers are chosen for each s, the maxmin operator requires to chose the maximum across different s. Since for any i smaller than $\lfloor N^{2/3} \rfloor$ and any $j \leq i$, we have $\tilde{W}_j = \bar{W}_l \geq \min_{t > \lfloor N^{2/3} \rfloor} \frac{1}{t - \lfloor N^{2/3} \rfloor} \sum_{m=\lfloor N^{2/3} \rfloor + 1}^{t} W_m$. Therefore, adding more terms before i will increase the overall level of the sample mean (22), so we must have $s = 1$. This is also justified by (23): for $s < t$, the smaller s, the greater $-\frac{t - \lfloor N^{2/3} \rfloor}{t - s + 1}$. [Note that $\min_{t > \lfloor N^{2/3} \rfloor} \frac{1}{t - \lfloor N^{2/3} \rfloor} \sum_{j=\lfloor N^{2/3} \rfloor + 1}^{t} W_j - \bar{W}_l \leq 0$ by the setup of Case (II).]

Thus, (22) can be written as

$$\tilde{p}(X_i) = \bar{W}_l + \frac{t_1 - \lfloor N^{2/3} \rfloor}{t_1} \left(\frac{1}{t_1 - \lfloor N^{2/3} \rfloor} \sum_{j=\lfloor N^{2/3} \rfloor + 1}^{t_1} W_j - \bar{W}_l \right) = \sum_{j=1}^{t_1} \frac{\tilde{W}_j}{t_1}, \tag{24}$$

with $t_1 > \lfloor N^{2/3} \rfloor$. (24) gives a common value of $\tilde{p}(X_i)$ for all $i = 1, \ldots, \lfloor N^{2/3} \rfloor$. By (3), we have $N_1 = t_1 > \lfloor N^{2/3} \rfloor$, which implies the conclusion.

Proof of (ii). Part (i) shows that all the changed treatment variable are clustered in the first and last group. Therefore, for $k = 2, 3, \ldots, K-1$, the statements of Propositions 1 and 2 (i) hold true, and we only need to show it also holds for $k = 1$ and K. Since the proof is similar, we only show the case of $k = 1$.

By using $N_1 \geq \lfloor N^{2/3} \rfloor$ from Part (i), it holds that for each $i = 1, \ldots, N_1$,

$$\begin{aligned}
\tilde{p}(X_i) &= \sum_{j=1}^{N_1} \frac{\tilde{W}_j}{N_1} = \sum_{j=1}^{\lfloor N^{2/3} \rfloor} \frac{\tilde{W}_j}{N_1} + \mathbb{I}\{N_1 > \lfloor N^{2/3} \rfloor\} \sum_{j=\lfloor N^{2/3} \rfloor + 1}^{N_1} \frac{W_j}{N_1} \\
&= \frac{1}{N_1} \left(\sum_{j=1}^{\lfloor N^{2/3} \rfloor} \left(\frac{1}{\lfloor N^{2/3} \rfloor} \sum_{i=1}^{\lfloor N^{2/3} \rfloor} W_i \right) + \mathbb{I}\{N_1 > \lfloor N^{2/3} \rfloor\} \sum_{j=\lfloor N^{2/3} \rfloor + 1}^{N_1} W_j \right) \\
&= \frac{1}{N_1} \left(\sum_{i=1}^{\lfloor N^{2/3} \rfloor} W_i + \mathbb{I}\{N_1 > \lfloor N^{2/3} \rfloor\} \sum_{j=\lfloor N^{2/3} \rfloor + 1}^{N_1} W_j \right) \\
&= \sum_{j=1}^{N_1} \frac{W_j}{N_1} = \frac{N_{1,1}}{N_1}.
\end{aligned}$$

Proof of (iii). Since the proof is similar, we focus on the first statement, $N_1 = O_p(N^{2/3})$. By definition, it is equivalent to show that there exists $c > 0$ such that $\mathbb{P}(c_1 > c) < \nu$ for any $\nu > 0$, where

$$c_1 = \frac{N_1}{N^{2/3}}. \tag{25}$$

Without loss of generality, we can set $c > 2$, and choose those sequence N such that $N^{2/3} = \lfloor N^{2/3} \rfloor$. Then,

$$\begin{aligned}
\mathbb{P}(c_1 > c) &= \mathbb{P}\left(\bar{W}_l \geq \frac{1}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} W_j, c_1 > c \right) \\
&= \mathbb{P}\left(\frac{1}{N^{2/3}} \sum_{i=1}^{N^{2/3}} W_i > \frac{1}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} W_j, c_1 > c \right) \\
&= \mathbb{P}\Bigg(\frac{1}{N^{2/3}} \sum_{i=1}^{N^{2/3}} \{W_i - p(X_i)\} + \frac{1}{N^{2/3}} \sum_{i=1}^{N^{2/3}} p(X_i) \\
&\qquad > \frac{1}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} \{W_j - p(X_j)\} + \frac{1}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} p(X_j), c_1 > c \Bigg) \\
&= \mathbb{P}\Bigg(\frac{1}{N^{1/3}} \sum_{i=1}^{N^{2/3}} \{W_i - p(X_i)\} - \frac{N^{1/3}}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} \{W_j - p(X_j)\} \\
&\qquad > \frac{N^{1/3}}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} p(X_j) - \frac{1}{N^{1/3}} \sum_{i=1}^{N^{2/3}} p(X_i), c_1 > c \Bigg) \\
&=: \mathbb{P}\left(\sum_{i=1}^{c_1 N^{2/3}} B_i > a, c_1 > c \right), \qquad (26)
\end{aligned}$$

where the first equality follows from $c_1 > c > 2$ and the implication of Case (II) of the proof of Proposition 3(i), the second equality follows from the definition of $\bar{W}_l$ in (21) and $N^{2/3} = \lfloor N^{2/3} \rfloor$, the third equality is by centering W around $p(X)$, the fourth equality follows from a rearrangement and multiplying both sides by $N^{1/3}$, and the last equality follows from the definitions

$$\begin{aligned}
B_i &= \begin{cases} \frac{W_i - p(X_i)}{N^{1/3}} & \text{for } 1 \leq i \leq N^{2/3} \\ -\frac{N^{1/3}\{W_j - p(X_j)\}}{c_1 N^{2/3} - N^{2/3}} & \text{for } \lfloor N^{2/3} \rfloor + 1 \leq i \leq c_1 N^{2/3}. \end{cases} \\
a &= \frac{N^{1/3}}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} p(X_j) - \frac{1}{N^{1/3}} \sum_{i=1}^{N^{2/3}} p(X_i). \qquad (27)
\end{aligned}$$

We now apply the following Bernstein inequality (see, e.g., van der geer, 2000) to (26).

Bernstein inequality. *Let $B_1, \ldots, B_n$ be independent random variables satisfying*

$$\begin{aligned}
&\mathbb{E}(B_i) = 0, \qquad \mathbb{E}|B_i|^m \leq \frac{m!}{2} A^{m-2} \mathbb{V}(B_i) \text{ for } m = 2, 3, \ldots, \\
&b^2 = \sum_{i=1}^{n} \mathbb{V}(B_i), \qquad (28)
\end{aligned}$$

for some constant A. Then

$$\mathbb{P}\left(\sum_{i=1}^{n} B_i \geq a \right) \leq \exp\left(-\frac{a^2}{2aA + 2b^2} \right).$$

In our problem, for any constant $c > 2$ (not c_1),

$$
\begin{aligned}
& \frac{N^{1/3}}{cN^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{cN^{2/3}} p(X_j) - \frac{1}{N^{1/3}} \sum_{i=1}^{N^{2/3}} p(X_i) \\
= \quad & N^{1/3} \left(\frac{\int_{q_{N^{-1/3}}}^{q_{c\cdot N^{-1/3}}} p(x)f(x)dx}{\int_{q_{N^{-1/3}}}^{q_{c\cdot N^{-1/3}}} f(x)dx} - \frac{\int_{x_L}^{q_{N^{-1/3}}} p(x)f(x)dx}{\int_{x_L}^{q_{N^{-1/3}}} f(x)dx} + O_p((N^{2/3})^{-1/2}) \right) \\
= \quad & N^{1/3} \Bigg(\frac{\int_{q_{N^{-1/3}}}^{q_{c\cdot N^{-1/3}}} \{p(x_L) + p^{(1)}(x_L)\cdot(x - x_L) + o(x - x_L)\} f(x)dx}{\int_{q_{N^{-1/3}}}^{q_{c\cdot N^{-1/3}}} f(x)dx} \\
& - \frac{\int_{x_L}^{q_{c\cdot N^{-1/3}}} \{p(x_L) + p^{(1)}(x_L)\cdot(x - x_L) + o(x - x_L)\} f(x)dx}{\int_{x_L}^{q_{N^{-1/3}}} f(x)dx} + O_p(N^{-1/3}) \Bigg) \\
\geq \quad & N^{1/3} \left[p(x_L) + p^{(1)}(x_L) \underline{f}(q_{c\cdot N^{-1/3}} - x_L) - \{p(x_L) + p^{(1)}(x_L) p^{(1)}(x_L) \overline{f}(q_{N^{-1/3}} - x_L) + o(N^{-1/3})\} \right] + O_p(1) \\
= \quad & p^{(1)}(x_L) N^{1/3} \{ \underline{f} q_{c\cdot N^{-1/3}}(q_{c\cdot N^{-1/3}} - x_L) - \overline{f}(q_{N^{-1/3}} - x_L)\} + O_p(1) \\
=: \quad & a_c + O_p(1), \qquad (29)
\end{aligned}
$$

where the first equality follows from the fact that the sample mean of a sample of size $N^{2/3}$ can estimate the population mean at the $O_p((N^{2/3})^{-1/2})$ rate, the second equality follows from an extension of $p(x)$ around x_L, the first inequality follows from Assumption 2(iii), and the last equality follows from the definition

$$a_c = p^{(1)}(x_L) N^{1/3} \{ \underline{f}(q_{c\cdot N^{-1/3}} - x_L) - \overline{f}(q_{N^{-1/3}} - x_L)\}. \qquad (30)$$

Now we show

$$a_c \to \infty \quad \text{as } c \to \infty. \qquad (31)$$

To this end, it is enough to show $\lim_{c\to\infty} N^{1/3} \cdot \underline{f}(q_{c\cdot N^{-1/3}} - x_L) = \infty$. By Assumption 2(iii), for or any $c \in \mathbb{N}$, we have

$$N^{-1/3}/\overline{f} \leq q_{(c+1)\cdot N^{-1/3}} - q_{(c)\cdot N^{-1/3}} \leq N^{-1/3}/\underline{f}. \qquad (32)$$

Combining (30) and (32) yields $a_c \geq p^{(1)}(x_L) \cdot c(\underline{f}/\overline{f}) + O_p(1)$, which implies (31).

By (30), we have

$$c_1 > c \Rightarrow a_{c_1} > a_c. \qquad (33)$$

On the other hand, for a defined in (27), (29) implies that

$$a \geq a_{c_1} + O_p(1). \qquad (34)$$

Now we study b in (28). Again, for any $c > 2$,

$$\begin{aligned}
& \frac{N^{2/3}}{(cN^{2/3} - N^{2/3})^2} \sum_{j=N^{2/3}+1}^{N_1} \{1 - p(X_j)\} p(X_j) + \frac{N^{2/3}}{N^{4/3}} \sum_{i=1}^{N^{2/3}} \{1 - p(X_i)\} p(X_i) \\
= & \frac{1}{c-1} \frac{(c-1)N^{2/3}}{(cN^{2/3} - N^{2/3})^2} \sum_{j=N^{2/3}+1}^{(c-1)N^{2/3}} \{1 - p(X_j)\} p(X_j) + \frac{1}{N^{2/3}} \sum_{i=1}^{N^{2/3}} \{1 - p(X_i)\} p(X_i) \\
= & \frac{c}{c-1} \{1 - p(x_L)\} p(x_L) + o_p(1) \\
=: & \ b_c^2 + o_p(1),
\end{aligned}$$

where the second equality follows from the consistency of the sample mean to the population mean, and the last equality follows by the definition $b_c^2 = \frac{c}{c-1}\{1 - p(x_L)\}p(x_L)$. Thus, we have

$$b^2 = b_{c_1}^2 + o_p(1). \tag{35}$$

Combining (34) and (33), for any $\nu > 0$, we can choose a large enough N and c such that

$$\mathbb{P}(a > \frac{1}{2} a_c, b^2 < 2b_c^2) < \frac{\nu}{2}.$$

Now we use the Bernstein inequality. Since B_i defined in (27) is a centered and normalized binary variable, we can simply choose $A = 1$ in (28), then

$$\begin{aligned}
\mathbb{P}(c_1 > c) < & \mathbb{P}(\sum B_i \geq a, a \geq \frac{1}{2} a_c, b^2 < 2b_c^2, c_1 > c) + \frac{\nu}{2} \\
& \leq \exp\left[-\frac{\frac{1}{4} a_c^2}{a_c + 4b_c^2} \right] + \frac{\nu}{2} \leq \nu.
\end{aligned} \tag{36}$$

The last inequality holds by (31), (34), and choosing a large enough c. Therefore, the conclusion $N_1 = O_p(N^{2/3})$ follows.

A.4. **Proof of Theorem 1.** Let q_α denote the α-th quantile of X. We define the following sequences of positive numbers

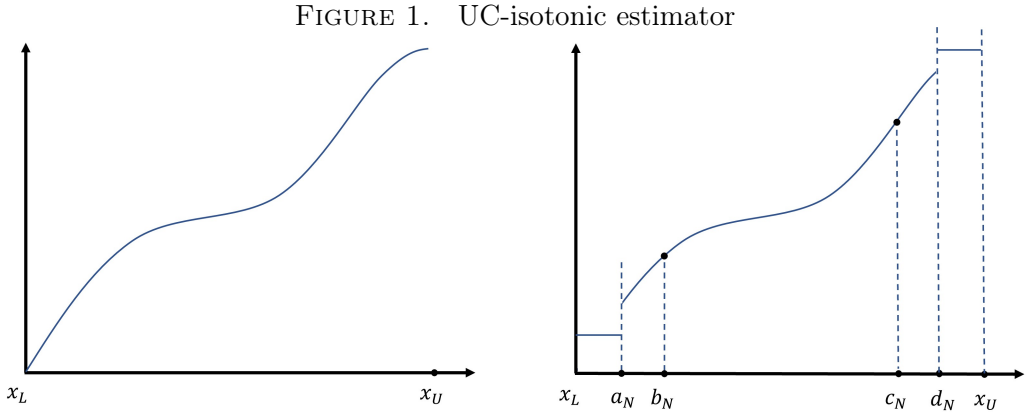
$$\begin{aligned}
a_N &= X_{N_1}, \qquad b_N = q_{(c_1+1)N^{-1/3}}, \\
c_N &= q_{1-(c_K+1)N^{-1/3}}, \qquad d_N = X_{n_K},
\end{aligned}$$

where n_K is defined in Section 2.1, which the first element of partition K (the last one). c_1 is defined in (25). c_K is defined similarly to c_1 by $c_K N^{2/3} = N_K$, where N_K is the number of elements partition K. Without a loss of generality, we assume that N is large enough to ensure that both quantiles b_N and c_N are well defined. For given N, the UC-isotonic estimator estimates the following function

$$p_N(x) = \begin{cases} \frac{\mathbb{E}[p(X)\mathbb{I}(X < a_N)]}{\mathbb{P}(X < a_N)} & \text{if } x \in [x_L, a_N) \\ p(x) & \text{if } x \in [a_N, d_N]. \\ \frac{\mathbb{E}[p(X)\mathbb{I}(X > d_N)]}{\mathbb{P}(X > d_N)} & \text{if } x \in (d_N, x_U] \end{cases}$$

It is shown in the following figure.

FIGURE 1. UC-isotonic estimator



The left panel is $p(x)$, and the right panel is $p_N(x)$.

The conclusion of Theorem 1 follows by showing these steps.

Step 1: $\sup_{x\in[x_L,a_N]}|\tilde{p}(x)-p(x)|= O_p(N^{-1/3})$ and $\sup_{x\in[d_N,x_U]}|\tilde{p}(x)-p(x)|= O_p(N^{-1/3})$.

Step 2: $\sup_{x\in[b_N,c_N]}|\tilde{p}(x)-p(x)|= O_p\left(\frac{\log N}{N}\right)^{1/3}$.

Step 3: $\sup_{x\in(a_N,b_N)}|\tilde{p}(x)-p(x)|= O_p\left(\frac{\log N}{N}\right)^{1/3}$ and $\sup_{x\in(c_N,d_N)}|\tilde{p}(x)-p(x)|= O_p\left(\frac{\log N}{N}\right)^{1/3}$.

Step 1. Note that $\tilde{p}(a_N)=\tilde{p}(x)$ for each $x\in[x_L,a_N]$. Therefore, for $\sup_{x\in[x_L,a_N]}|\tilde{p}(x)-p(x)|= O_p(N^{-1/3})$, it is enough to show that

$$\sup_{x\in[x_L,a_N]}|\tilde{p}(a_N)-p(x)|= O_p(N^{-1/3}).$$

First, by Assumption (2)(iii), we have $a_N - x_L = O_p(N^{-1/3})$,which implies

$$x - x_L = O_p(N^{-1/3}) \text{ for } x\in[x_L,a_N]. \tag{37}$$

Thus, we have

$$\begin{aligned}
\tilde{p}(a_N) &= \frac{1}{N_1}\sum_{i=1}^{N_1} W_i = \frac{1}{c_1N^{2/3}}\sum_{i=1}^{c_1N^{2/3}} W_i \\
&= \frac{\int_{x_L}^{q_{c_1\cdot N^{-1/3}}} p(x)f(x)dx}{\int_{x_L}^{q_{c_1\cdot N^{-1/3}}} f(x)dx} + O_p((c_1N^{2/3})^{-1/2}) + O_P(N^{-1/2}) \\
&= \frac{\int_{x_L}^{q_{c_1\cdot N^{-1/3}}} \{p(x_L)+p^{(1)}(x_L)\cdot(x_{c_1}-x_L)\}f(x)dx}{\int_{x_L}^{q_{c_1\cdot N^{-1/3}}} f(x)dx} + O_p(N^{-1/3}) \\
&= p(x_L) + O_p(c_1\cdot N^{-1/3}) + O_p(N^{-1/3}) \\
&= p(x) + O_p(N^{-1/3}) + O_p(c_1\cdot N^{-1/3}) + O_p(N^{-1/3}) \\
&= p(x) + O_p(N^{-1/3}),
\end{aligned}$$

where the first equality follows from Proposition (3)(ii), the $O_P(N^{-1/2})$ term in the third equality follows from that $X_{c_1N^{2/3}}$ can estimate $q_{c_1\cdot N^{-1/3}}$ at the rate of $O_P(N^{-1/2})$, x_{c_1} in the fourth equality is a number within the interval $(x_L, q_{c_1\cdot N^{-1/3}})$, the fifth equality follows from $x_{c_1}-x_L =$

$O_p(c_1 \cdot N^{-1/3})$, the sixth equality follows from (37) and Assumption 2, and the last equality follows from $c_1 = O_p(1)$, which is implied by (36).

Similarly, we can show $\sup_{x\in[d_N,x_U]}|\tilde{p}(x) - p(x)| = O_p(N^{-1/3})$.

Step 2. The result $\sup_{x\in[b_N,c_N]}|\tilde{p}(x) - p(x)| = O_p\left(\frac{\log N}{N}\right)^{1/3}$ follows by Durot, Kulikov, and Lopuhaä, (2013, Theorem 2.1). We can adapt the support from $[0,1]$ in their theorem to $[x_L, x_U]$ in our problem, and their conditions (A1)-(A3) hold under Assumptions 1 and 2.

Step 3. The statement follows directly by combining the results from Steps 1 and 2 with Assumption 2, $b_N - a_N = O_p(N^{-1/3})$, and $d_N - c_N = O_p(N^{-1/3})$.

Combining these steps, the conclusion of this theorem follows.

A.5. **Proof of Theorem 2.** If X_i is in the k-th partition given by the UC-isotonic estimator (i.e., $i \in \{n_k, \ldots, (n_k + N_k - 1)\}$), then we have $M_i = N_{k,1-W_i}$. Thus, the matching estimator $\hat{\tau}$ is written as

$$
\begin{aligned}
\hat{\tau} &= \frac{1}{N}\sum_{i=1}^{N}(2W_i-1)\left(Y_i - \frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j\right) = \frac{1}{N}\sum_{k=1}^{K}\left\{\sum_{i=n_k}^{n_k+N_k-1}(2W_i-1)\left(Y_i - \frac{1}{N_{k,1-W_i}}\sum_{j\in\mathcal{J}(i)} Y_j\right)\right\} \\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},W_i=1}\left(Y_i - \frac{1}{N_{k,0}}\sum_{j\in\{n_k,\ldots,(n_k+N_k-1)\},W_j=0} Y_j\right)\right. \\
&\qquad \left. + \sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},W_i=0}\left(\frac{1}{N_{k,1}}\sum_{j\in\{n_k,\ldots,(n_k+N_k-1)\},W_j=1} Y_j - Y_i\right)\right\} \\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\left(1+\frac{N_{k,0}}{N_{k,1}}\right)\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},W_i=1} Y_i - \left(\frac{N_{k,1}}{N_{k,0}}+1\right)\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},W_i=0} Y_i\right\} \\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\frac{N_{k,1}+N_{k,0}}{N_{k,1}}\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},W_i=1} Y_i - \frac{N_{k,1}+N_{k,0}}{N_{k,0}}\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},W_i=0} Y_i\right\} \\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},W_i=1}\frac{Y_i}{N_{k.1}/N_k} - \sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},W_i=0}\frac{Y_i}{N_{k.0}/N_k}\right\} \\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\}}\left(\frac{W_iY_i}{\tilde{p}(X_i)} - \frac{(1-W_i)\,Y_i}{1-\tilde{p}(X_i)}\right)\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{\tilde{p}(X_i)} - \frac{(1-W_i)\,Y_i}{1-\tilde{p}(X_i)}\right),
\end{aligned}
\tag{38}
$$

where the first equality is the formula of matching estimator for ATE (see, e.g., Abadie and Imbens, 2016), with a changing matched set of size M_i, the fourth equality follows from the fact that with $w \in \{0,1\}$ and $i \neq j$, we have $\sum_{i\in n_k:(n_k+N_k-1),W_i=w} Y_j = N_{k,w}\cdot Y_j$, the second last equality follows from Lemma 3, and the last equality follows from $\sum_{k=1}^{K} N_k = N$.

A.6. **Proof of Theorem 3.** Theorem 3 follows from Theorem 2 and Xu (2021), where generic results for semiparametric estimation with plug-in isotonic estimator were given and applied to the case of IPW estimator. The proof will be decomposed into the following steps:

In Appendix A.6.1, we summarize the generic results for semiparametric estimation with plug-in (standard) isotonic estimator discussed in Xu (2021), and introduce relevant assumptions and lemmas. Their proofs are in Appendix A.6.2.

Appendix A.6.3 shows that the results in Appendix A.6.1 also hold for the plug-in UC-isotonic estimator. Then, we show the asymptotic properties of the IPW with plug-in UC-isotonic estimator and conclude the proof.

A.6.1. *Relevant results on semiparametric estimation with plug-in isotonic estimator.* We start by introducing several assumptions, lemma, theorem, and corollary. Other than in the main paper, the sample size is denoted by n (not N) in this and next subsections.

Suppose we have a moment condition

$$\mathbb{E}[m(Z, \beta_0, p_0(\cdot))] = 0, \tag{39}$$

where Z is a random vector defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P}_0)$, and $\beta_0 \in \mathfrak{B} \subset \mathbb{R}^k$ is a real-valued parameter of interest. $p_0(\cdot)$ is a monotone increasing nuisance function, which is the conditional mean of some function of data. Suppose we have the just identification, i.e., $\dim(m(\cdot,\cdot,\cdot)) = \dim(\beta_0) = k$. We are aiming at showing the properties of an estimator $\hat{\beta}$ solved from

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{p}(\cdot)) = 0. \tag{40}$$

where Z_i is a vector of i-th observed data point, $\hat{p}(x)$ is a plug-in isotonic estimator of $\mathbb{E}[Y|X = x]$. (Y, X) is a sub-vector of Z. Note that Y here are no longer defined as outcomes, but the dependent variable of an isotonic regression, and it corresponds the treatment W in the main paper. In the general setting, Y can be both binary or continuous. To proceed, we impose the following assumptions.

- **A1:** X is a random scalar taking value in the space $\mathcal{X}$. The space $\mathcal{X}$ is convex with non-empty interiors, and satisfies $\mathcal{X} \subset \mathcal{B}(0, R)$ for some $R > 0$.
- **A2:** $\mathbb{E}[Y|X = x] = p_0(x)$ is monotone increasing in x. There exists $K_0 > 0$ such that $|p_0(x)| < K_0$ for all $x \in \mathcal{X}$.
- **A3:** There exist $c_0 > 0$ and $M_0 > 0$ such that $\mathbb{E}[|Y|^m|X = x] \leq m!\, M_0^{m-2} c_0$ for all integers $m \geq 2$ and almost every x.

A1 and A2 impose boundedness on the monotone function p_0 and the support of X. These conditions are used to control the entropy of the function classes that characterize (40). A3 is to restrict the size of the tail of $Y|X$. With A3, we can show that $\sup_{x \in X} \hat{p}(x) = O_p(\log n)$, which is used to obtain an entropy result associated with the $\sqrt{n}$-convergence rate in the second-stage semiparametric estimator.

Lemma 2. *Let $\hat{p}(\cdot)$ be an isotonic estimator of $\mathbb{E}[Y|X]$, and $\delta(X)$ be a bounded function of X with a finite total variation. Under A1, A2, and A3, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\delta(X_i)(Y_i-\hat{p}(X_i))=o_p(n^{-1/2}). \tag{41}$$

Let $D(z,\beta)$ be the functional derivative of $m(z,\beta,p(x))$ with respect to $p(\cdot)$. We assume:

A4: For all $\beta\in\mathfrak{B}$, $\mathbb{E}[D(Z,\beta)|X]$ is a bounded function of X with a finite total variation, and there exist $c_1>0$ and $M_1>0$ such that for each row of $D(Z,\beta)$ ($D_j(Z,\beta)$ with $j\in\{1:k\}$), $\mathbb{E}[|D_j(Z,\beta)|^m|X=x]\leq m!\,M_1^{m-2}c_1$ for all integers $m\geq 2$ and almost every x.

By (41), we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[D(Z_i,\beta_0)|X_i](Y_i-\hat{p}(X_i))=o_p(n^{-1/2}).$$

Then we add the following assumption:

A5: The first-order expansion of $m(z,\beta,p(\cdot))$ with respect to $p(\cdot)$ at $p^*(\cdot)$, $D(z,\beta,p(\cdot)-p^*(\cdot))$, is linear in $p(\cdot)-p^*(\cdot)$ (i.e., $D(z,\beta,p(x)-p^*(x))=D(z,\beta)\{p(x)-p^*(x)\}$).

The correction term $\mathbb{E}[D(Z,\beta_0)|X]\{Y-p_0(X)\}$ is also associated with efficiency. As illustrated in Newey (1994, Proposition 4), for unconditional moment condition $\mathbb{E}[m(Z,\beta,p(X))]=0$, where $p_0(X)=\mathbb{E}[Y|X]$ for some sub-vector Y, the efficient influence function ψ is

$$\psi(Z)=-\left(\frac{\partial\mathbb{E}[m(Z,\beta_0,p(X))]}{\partial\beta}\right)^{-1}[m(Z,\beta_0,p_0(X))+\mathbb{E}[D(Z,\beta_0)|X]\{Y-p_0(X)\}].$$

If we could show for an isotonic plug-in estimator $\hat{p}(\cdot)$

$$\frac{1}{n}\sum_{i=1}^{n}m(Z_i,\beta_0,\hat{p}(X_i))=\frac{1}{n}\sum_{i=1}^{n}\{m(Z_i,\beta_0,p_0(x_i))+\mathbb{E}[D(Z,\beta_0)|X_i](y_i-p_0(x_i))\}+o_p(n^{-1/2}),$$

we could show the asymptotic efficiency. To this end, we impose the following assumptions:

A6: There are $b(z)>0$ and $D(z,g)$ that (i) $||m(z,\beta,p)-m(z,\beta,p_0)-D(z,\beta,p-p_0)||\leq b(z)||p-p_0||^2$; (ii) $\mathbb{E}[b(Z)]=o_p(n^{1/6}(\log n)^{-2})$, for all $\beta\in\mathfrak{B}$, where $\mathfrak{B}$ is compact.

A7: There are ε, $b(z)$, $\tilde{b}(z)>0$ and $p(\cdot)$ with $||p||>0$. Such that (i) for all $\beta\in\mathfrak{B}$, $m(z,\beta,p_0)$ is continuous at β and $m(z,\beta,p_0)\leq b(z)$; (ii) $||m(z,\beta,p)-m(z,\beta,p_0)||\leq\tilde{b}(z)(||p-p_0||)^{\varepsilon}$.

A8: $\mathbb{E}[m(z,\beta,p_0)]=0$ has a unique solution on $\mathfrak{B}$ at β_0.

A9: For $\beta\in$ interior$(\mathfrak{B})$, (i) there are $\varepsilon>0$ and a neighborhood $\mathscr{N}$ of β_0 such that for all $||p-p_0||\leq\varepsilon$, $m(z,\beta,p)$ is differentiable in β on $\mathscr{N}$; (ii) $M_\beta=-\mathbb{E}\left[\frac{\partial m(Z,\beta_0,p_0(X))}{\partial\beta}\right]$ is nonsingular; (iii) $\mathbb{E}[||m(z,\beta,p)||^2]<\infty$; (iv) Assumption A7 is satisfied with $m(z,\beta,p)$ equaling to each row of $\frac{\partial m(Z,\beta,p)}{\partial\beta}$.

A6 is an adaption of Assumption 5.1 in Newey (1994). This assumption requires that the higher order term from a linear approximation is small. A7, A8, and A9 are adapted from Assumption 5.4, 5.5, and 5.6 in Newey (1994). They are general conditions for the consistency and asymptotical normality of the method of moment estimator.

Define $M(Z)=\mathbb{E}[D(Z,\beta_0)|X](Y-p_0(X))$. The asymptotic normality and efficiency of the plug-in estimator $\hat{\beta}$ is obtained as follows.

Lemma 3. *Let $\hat{p}(\cdot)$ be an isotonic estimator of $p_0(\cdot) = \mathbb{E}[Y|X = \cdot]$. Under A1-A9, it holds*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V),$$

where $V = M_\beta^{-1}\mathbb{E}[\{m(Z, \beta_0, p_0) + M(Z)\}\{m(z, \beta_0, p_0) + M(Z)\}']M_\beta^{-1}$.

A.6.2. *Proofs of the results in Appendix A.6.1 .*

Proof of Lemma 2. The proof here is based on the supplementary material of BGH (hereafter BGH-supp). Similar techniques can also be found in Groeneboom and Jongbloed (2014) and Groeneboom and Hendrickx (2018). For isotonic estimator $\hat{p}(\cdot)$, we have

$$||\hat{p} - p_0||^2 = O_p((\log n)^2 n^{-2/3}) = o_p(n^{-1/2}), \tag{42}$$

(see, e.g., Theorem 9.2 and Lemma 5.15 in van de Geer, S., 2000).

Let $\{x_{n_j}\}_{j=1}^k$ be the subsequence of $\{x_i\}_{i=1}^n$ representing all the jump points of $\hat{p}(\cdot)$. By the construction of $\hat{p}(\cdot)$ (see, e.g., Lemmas 2.1 and 2.3 in Groeneboom and Jongbloed, 2014), we have $\sum_{i=n_j}^{n_{j+1}-1}\{y_i - \hat{p}(x_i)\} = 0$ for each $j = 1, \ldots, k$, which implies

$$\sum_{j=1}^{k} m_j \sum_{i=n_j}^{n_{j+1}-1}\{y_i - \hat{p}(x_i)\} = 0, \tag{43}$$

for any weights $\{m_j\}_{j=1}^k$. Define the step function

$$\bar{\delta}_n(x) = \begin{cases} \delta(x_{n_j}) & \text{if } p_0(x) > \hat{p}(x_{n_j}) \text{ for all } x \in (x_{n_j}, x_{n_{j+1}}) \\ \delta(s) & \text{if } p_0(s) = \hat{p}(s) \text{ for some } s \in (x_{n_j}, x_{n_{j+1}}) \\ \delta(x_{n_{j+1}}) & \text{if } p_0(x) < \hat{p}(x_{n_j}) \text{ for all } x \in (x_{n_j}, x_{n_{j+1}}) \end{cases},$$

for $x \in [x_{n_j}, x_{n_{j+1}})$ with $j = 1, \ldots, k$ (if $j = k$, set $x_{n_{j+1}} = \max_i x_{n_i}$). By (43), it holds $\int \bar{\delta}_n(x)\{y - \hat{p}(x)\}d\mathbb{P}_n(z) = 0$, and thus

$$\frac{1}{n}\sum_{i=1}^{n}\delta(X_i)(Y_i - \hat{p}(X_i)) = \int \delta(x)\{y - \hat{p}(x)\}d\mathbb{P}_n(z) = \int [\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x))d\mathbb{P}_n(z). \tag{44}$$

By assumption, $\delta(x)$ is a bounded function with a finite total variation, so is $\bar{\delta}_n(x)$. Therefore, by a similar argument as in pp. 18-20 of BGH-supp, we have $\int [\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x))d\mathbb{P}_n(z) = o_p(n^{-1/2})$. We see that (44) can be decomposed as

$$\begin{aligned} \int [\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x))d\mathbb{P}_n(z) &= \int [\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x))d\{\mathbb{P}_n(z) - \mathbb{P}_0(z)\} \\ &\quad + \int [\delta(x) - \bar{\delta}_n(x)](y - p_0(x))d\mathbb{P}_0(z) \\ &\quad + \int [\delta(x) - \bar{\delta}_n(x)](p_0(x) - \hat{p}(x))d\mathbb{P}_0(z) \\ &= I + II + III. \end{aligned}$$

By Lemma 21 in BGH-supp, $\delta(x) - \bar{\delta}_n(x)$ are bounded functions with finite total variations. By similar arguments in Groeneboom and Jongbloed (2014), there exists a positive constant C_0 such

that for all $x \in \mathcal{X}$,

$$|\delta(x) - \bar{\delta}_n(x)| \leq C_0 |p_0(x) - \hat{p}(x)|. \tag{45}$$

For I, let us define the following function classes

$$\begin{aligned}
\mathcal{M}_{RK} &= \{\text{monotone increasing functions on } [-R, R] \text{ and bounded by } K\}, \\
\mathcal{G}_{RK} &= \{g : g(x) = p(x), x \in \mathcal{X}, p \in \mathcal{M}_{RK}\}, \\
\mathcal{D}_{RKv} &= \{d : d(x) = g_1(x) - g_2(x),\ (g_1, g_2) \in \mathcal{G}_{RK}^2,\ ||d(x)||_{P_0} \leq v\}, \\
\mathcal{H}_{RKv} &= \{h : h(y, x) = y d_1(x) - d_2(x),\ (d_1, d_2) \in \mathcal{D}_{RKv}^2,\ z \in \mathcal{Z}\}.
\end{aligned} \tag{46}$$

The integrand of I is written as

$$[\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x)) = [\delta(x) - \bar{\delta}_n(x)]y - [\delta(x) - \bar{\delta}_n(x)]\hat{p}(x). \tag{47}$$

Let $\mathcal{F}_a = \{f : f(z) = [\delta(x) - \bar{\delta}_n(x)]y - [\delta(x) - \bar{\delta}_n(x)]\hat{p}(x),\ z \in \mathcal{Z}\}$. We observe the following facts:

(i) By Lemma 21 in BGH-supp, $\delta(x) - \bar{\delta}_n(x)$ is a bounded function of x with finite total variation.

(ii) By Assumption A3, we can show $\sup_{x \in \mathcal{X}} |\hat{p}(x)| = O_p(\log n)$ (see, e.g., Lemma 7.1 in Balabdaoui, Durot, and Jankowski, 2019). Therefore, there exists $K_1 > 0$, such that $\hat{p}(x) \in \mathcal{G}_{R(K_1 \log n)}$ with probability approaching one.

(iii) By (42) and (45), we have $||\delta(x) - \bar{\delta}_n(x)||_2 \leq C_1 (\log n) n^{-1/3}$, for some $C_1 > 0$. Thus, there exists a positive constant C_2 that is larger than twice the bound of $\delta(x)$, and $v_1 = C_1 (\log n) n^{-1/3}$, such that $[\delta(x) - \bar{\delta}_n(x)] \in \mathcal{D}_{RC_2 v_1}$.

(iv) By (ii), a similar argument of (iii), (42), and Jensen's inequality, we have $[\delta(x) - \bar{\delta}_n(x)]\hat{p}(x) \in \mathcal{D}_{R(K_2 \log n) v_2}$ for a large enough constant $K_2 > 0$ and $v_2 = C_3 (\log n)^2 n^{-1/3}$ for some $C_3 > 0$, with probability approaching one.

We choose $K = \max\{C_2, K_2 \log n\}$ and $v = \max\{v_1, v_2\}$. Now we have (47)$\in \mathcal{H}_{RKv}$.

Define $||\mathbb{G}_n||_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\sqrt{n}(\mathbb{P}_n - P_0) f|$. Let $N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||)$ be the ε-bracketing number of the function class $\mathcal{F}$ under the norm $||\cdot||$, and $H_B(\varepsilon, \mathcal{F}, ||\cdot||) = \log N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||)$ be the entropy of $N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||)$, and $J_n(\delta, \mathcal{F}, ||\cdot||) = \int_0^\delta \sqrt{1 + H_B(\varepsilon, \mathcal{F}, ||\cdot||)} d\varepsilon$. Let $||\cdot||_{B,P_0}$ be the Bernstein norm under a measure P_0. In this section, we use $J_n(\delta)$ to denote $J_n(\delta, \mathcal{F}, ||\cdot||_{B,P_0})$. By similar arguments in Lemma 13 of BGH-supp (in our case, we can ignore the single-index coefficients), we have, with probability approaching one:

$$H_B(\varepsilon, \tilde{\mathcal{F}}_a, ||\cdot||_{B,P_0}) \leq \frac{C_3}{\varepsilon}, \tag{48}$$

for some $C_3 > 0$, where $\tilde{\mathcal{F}}_a = (C_4 \log n)^{-1} \mathcal{F}_a$ with some $C_4 > 0$. Also, there exists a constant $C_5 > 0$ such that

$$||\tilde{f}||_{B,P_0} \leq C_5 (\log n) n^{-1/3}, \tag{49}$$

for all $\tilde{f}_a \in \tilde{\mathcal{F}}_a$, with probability approaching one. We use $\mathscr{E}$ to denote the event that both (48) and (49) happen, and we have $\lim_{n \to \infty} P(\mathscr{E}) = 1$.

Let $\delta_n = C_5(\log n)n^{-1/3}$ and I_j be the j-th component of I. For any positive constants A and ν, there exist positive constants B_1, and B_2 such that

$$\begin{aligned}
\mathbb{P}\{|I_j| > An^{-1/2}\} &\leq \mathbb{P}\left\{|I_j| > An^{-1/2}, \mathscr{E}\right\} + P(\mathscr{E}^c) \leq \mathbb{P}\left\{||\mathbb{G}_n||_{\mathcal{F}_a} > A, \mathscr{E}\right\} + \frac{\nu}{2} \\
&\leq \frac{\mathbb{E}[||\mathbb{G}_n||_{\mathcal{F}_a} \quad |\mathscr{E}]}{A} + \frac{\nu}{2} = \frac{C_4 \log n}{A}\mathbb{E}[||\mathbb{G}_n||_{\tilde{\mathcal{F}}_a} \quad |\mathscr{E}] + \frac{\nu}{2} \\
&\lesssim \frac{C_4 \log n}{A} J_n(\delta_n)\left(1 + \frac{J_n(\delta_n)}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A}(\delta_n + 2B_1^{1/2}\delta_n^{1/2})\left(1 + \frac{\delta_n + 2B_1^{1/2}\delta_n^{1/2}}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{1}{A}(\log n)^{3/2} n^{-1/6}\left(1 + \frac{B_2}{(\log n)^{3/2}}\right) + \frac{\nu}{2} \lesssim \nu, \qquad (50)
\end{aligned}$$

for all n large enough, where the second inequality follows from the definition of $\mathcal{F}_a$, the third inequality follows from the Markov inequality, the first equality follows from the definition of $\tilde{\mathcal{F}}_a$, the first wave inequality ($\lesssim$) comes from Lemma 3.4.3 of van der Vaart and Wellner (1996) and the definition of δ_n, the second wave inequality comes from (48) and Equation (.2) in BGH-supp, and the third wave inequality follows from $\delta_n \lesssim \delta_n^{1/2}$ and the definition of δ_n. Therefore,

$$I = o_p(n^{-1/2}). \qquad (51)$$

For II, the law of iterated expectation yields

$$II = \int [\delta(x) - \bar{\delta}_n(x)](y - p_0(x))d\mathbb{P}_0(z) = 0.$$

For III, we have

$$\begin{aligned}
III &= \int [\delta(x) - \bar{\delta}_n(x)](p_0(x) - \hat{p}(x))d\mathbb{P}_0(z) \lesssim \int (p_0(x) - \hat{p}(x))^2 d\mathbb{P}_0(z) \\
&= O_p((\log)^2 n^{-2/3}) = o_p(n^{-1/2}),
\end{aligned}$$

where the first wave inequality follows from (45), and the second equality follows from (42).

Combining the rates for I, II, and III, the conclusion of Lemma 2 follows.

Proof of Lemma 3. The proof is a combination of the techniques for isotonic regression applied in Groeneboom and Hendrickx (2018) and BGH, and the framework of Newey (1994). Let $u = y - p_0(x)$ and $M(z) = \delta(x)u$. We verify Assumptions 5.1-5.6 of Newey (1994).

Step 1: Verify Assumption 5.1 of Newey (1994). (i) There is a function $D(z,p)$ that is linear in p such that for all p with $||p-p_0||$ small enough, $||m(z,p)-m(z,p_0)-D(z,p-p_0)|| \leq b(z)||p-p_0||^2$, and (ii) $\mathbb{E}(b(Z))\sqrt{n}||\hat{p}-p_0||^2 \xrightarrow{p} 0$.

Assumption 5.1(i) is a restatement of A6 (i). Assumption 5.1(ii) can be derived by A6(ii) and the fact $||\hat{p} - p_0||^2 = O_p((\log n)^2 n^{-2/3})$ (see, e.g., Theorem 9.2 and Lemma 5.15 in van de Geer, S., 2000).

Step 2: Verify Assumption 5.2 of Newey (1994). It holds $\frac{1}{n}\sum_{i=1}^{n} D(Z_i, \hat{p}(X_i) - p_0(X_i)) - \int D(z, \hat{p}(x) - p_0(x)) d\mathbb{P}_0(z) = o_p(n^{-1/2})$.

By A5, we have

$$\frac{1}{n}\sum_{i=1}^{n} D(Z_i, \beta_0, \hat{p}(X_i) - p_0(X_i)) - \int D(z, \beta_0, \hat{p}(x) - p_0(x)) d\mathbb{P}_0(z) = \int D(z, \beta_0)(p_0(x) - \hat{p}(x)) d(\mathbb{P}_n - \mathbb{P}_0)(z). \tag{52}$$

Let

$$\mathcal{F}_b = \{f : f(z) = D(z, \beta_0)(p_0(x) - \hat{p}(x)),\ x \in \mathcal{X}\}.$$

To avoid heavy notations, we re-define some constant terms in this subsection, such as A_i, C_i, K_i, δ_n, and v, etc.. They are not related to those constants with the same names in other sections.

By similar arguments as in the proof of Lemma 2, for some $C_1, C_2 > 0$, we have

$$p_0(x) - \hat{p}(x) \in \mathcal{D}_{R(C_1 \log n)(C_2 n^{-1/3} \log n)}, \tag{53}$$

with probability approaching one. By Theorem 2.7.5 in van der Vaart and Wellner (1996) and Lemma 11 in BGH-supp, with $R, C, v > 0$, we have

$$H_B(\varepsilon, \mathcal{D}_{RCv}, ||\cdot||_{P_0}) \leq \frac{AC}{\varepsilon},$$

for some $A > 0$.

Now we define

$$\mathcal{H}^{(2)}_{RKv} = \{h : h(z) = D(z, \beta_0) d(x),\ d(\cdot) \in \mathcal{D}_{RKv},\ z \in \mathcal{Z}\},$$

and let $D(z, \beta_0) \in \mathbb{R}^1$. This is just to simplify the notation of the following proof, i.e., the following steps hold for any $D_j(z, \beta_0)$ with $j \in \{1 : k\}$, the j-th row of $D(z, \beta_0)$. Let (d^L, d^U) be any ϵ-bracket of the function class $\mathcal{D}_{RKv}$, and

$$h^L = \begin{cases} D(z, \beta_0) d^L(x) & \text{if } D(z, \beta_0) \geq 0 \\ D(z, \beta_0) d^U(x) & \text{if } D(z, \beta_0) < 0 \end{cases},$$

$$h^U = \begin{cases} D(z, \beta_0) d^U(x) & \text{if } D(z, \beta_0) \geq 0 \\ D(z, \beta_0) d^L(x) & \text{if } D(z, \beta_0) < 0 \end{cases}.$$

We see that (h^L, h^U) is a bracket of h whose size is

$$\begin{aligned} \int_{\mathcal{Z}} \{h^U(z) - h^L(z)\}^2 d\mathbb{P}_0(z) &= \int_{\mathcal{Z}} D(z, \beta_0)^2 \{d^U(x) - d^L(x)\}^2 d\mathbb{P}_0(z) \\ &= \int_{\mathcal{X}} \mathbb{E}[D(z, \beta_0)^2 | x] \{d^U(x) - d^L(x)\}^2 d\mathbb{P}_0(x) \\ &= A_1 \epsilon^2, \end{aligned}$$

for some $A_1 > 0$. The last equality follows from A4 and the definition of ϵ-bracket. Thus, for some $\tilde{A} > 0$, we have

$$H_B(\varepsilon, \mathcal{H}^{(2)}_{RCv}, ||\cdot||_{P_0}) \leq \frac{\tilde{A}C}{\varepsilon}. \tag{54}$$

Now we switch to Bernstein norm since we do not want to put a bound on $D(z,\beta_0)$. By the definition of Bernstein norm

$$||h||^2_{B,P_0} = 2\mathbb{E}_0[\exp(|h|) - |h| - 1] = 2\int \sum_{k=2}^{\infty} \frac{1}{k!}|h|^k d\mathbb{P}_0(z).$$

We bound the Bernstein norm of $\frac{h(\cdot)}{H}$, where H is some positive constant chosen in the following steps to achieve a finite upper bound. Observe that

$$\begin{aligned}
||H^{-1}h||^2_{B,P_0} &= 2\int \sum_{k=2}^{\infty} \frac{1}{H^k}\frac{1}{k!}|D(z,\beta_0)d(x)|^k d\mathbb{P}_0(z) \leq 2\int \sum_{k=2}^{\infty} \frac{1}{H^k}\frac{1}{k!}|D(z,\beta_0)|^k |d(x)|^k d\mathbb{P}_0(z) \\
&\leq 2\sum_{k=2}^{\infty} \frac{1}{H^k}\frac{(2C)^{k-2}}{k!} k!\, M_1^{k-2} c_1 \int |d(x)|^2 d\mathbb{P}_0(z) \\
&= \frac{2}{H^2}\sum_{k=2}^{\infty} \frac{(2M_1C)^{k-2}}{H^{k-2}} c_1 \int |d(x)|^2 d\mathbb{P}_0(z) = \frac{2}{H^2}\sum_{k=2}^{\infty} \left(\frac{2M_1C}{H}\right)^{k-2} c_1 v^2 \\
&= \left(\frac{2}{H}\right)^2 c_1 v^2,
\end{aligned}$$

where the second inequality follows from A4 and the fact $d(\cdot) \in \mathcal{D}_{RCv}$, where c_1 and M_1 are the same constants in A4, the third equality follows from the definition of v in $\mathcal{D}_{RCv}$, and the last equality follows by choosing $H = 4M_1C$. This implies

$$||H^{-1}h||_{B,P_0} \lesssim H^{-1}v. \tag{55}$$

Now by setting $C = C_1 \log n$, $v = C_2 n^{-1/3}\log n$ so that $\mathcal{F}_b \subset \mathcal{H}^{(2)}_{R(C_1\log n)(C_2 n^{-1/3}\log n)}$, and $\tilde{H} = 4M_1C_1\log n$, we have

$$\tilde{\mathcal{F}}_b = \tilde{H}^{-1}\mathcal{F}_b \quad \text{for some } C_3 > 0.$$

Thus, combining (54) and (55), it holds that with probability approaching one

$$H_B(\varepsilon, \tilde{\mathcal{F}}_b, ||\cdot||_{B,P_0}) \leq \frac{C_3}{\varepsilon}, \tag{56}$$

for some $C_3 > 0$, and

$$||\tilde{f}||_{B,P_0} \leq C_4 n^{-1/3}, \tag{57}$$

for all $\tilde{f}_b \in \tilde{\mathcal{F}}_b$ and some $C_4 > 0$.

We use $\mathscr{E}_1$ to denote the events defined in (56) and (57), and use S to denote the value of (52). Let $\delta_n = C_4 n^{-1/3}$. For any $A_2 > 0$, it holds

$$\begin{aligned}
\mathbb{P}\{|S| > A_2 n^{-1/2}\} &\leq \mathbb{P}\{|S| > A_2 n^{-1/2}, \mathscr{E}_1\} + \mathbb{P}\{\mathscr{E}_1^c\} \leq \mathbb{P}\{||\mathbb{G}_n||_{\mathcal{F}_b} > A_2, \mathscr{E}_1\} + \frac{\nu}{2} \\
&\leq \frac{\mathbb{E}[||\mathbb{G}_n||_{\mathcal{F}_b}|\mathscr{E}_1]}{A_2} + \frac{\nu}{2} \lesssim \frac{\log n}{A_2}\mathbb{E}[||\mathbb{G}_n||_{\tilde{\mathcal{F}}_b}|\mathscr{E}_1] + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A_2} J_n(\delta_n)\left(1 + \frac{J_n(\delta_n)}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \lesssim \frac{\log n}{A_2}(\delta_n + 2B_1^{1/2}\delta_n^{1/2})\left(1 + \frac{\delta_n + 2B_1^{1/2}\delta_n^{1/2}}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{1}{A}(\log n)^{3/2} n^{-1/6}\left(1 + \frac{B_2}{(\log n)^{3/2}}\right) + \frac{\nu}{2} \lesssim \frac{\log n}{A_2} n^{-1/6} B_2 + \frac{\nu}{2} \\
&\lesssim \nu. \qquad (58)
\end{aligned}$$

These steps are similar to those of (50). Thus, we have $\int D(z, \beta_0)\{p_0(x) - \hat{p}(x)\}d(\mathbb{P}_n - \mathbb{P}_0)(z) = o_p(n^{-1/2})$ and Assumption 5.2 in Newey (1994) is satisfied.

Step 3: Verify Assumption 5.2 of Newey (1994).[4] *It holds* $\int D(z, \hat{p}(x) - p_0(x))d\mathbb{P}_0(z) = \frac{1}{n}\sum_{i=1}^n M(Z_i) + o_p(n^{-1/2})$.

Note that

$$\begin{aligned}
\int D(z, \beta_0, \hat{p}(x) - p_0(x))d\mathbb{P}_0(z) &= \int D(z, \beta_0)(\hat{p}(x) - p_0(x))d\mathbb{P}_0(x) \\
&= \int \mathbb{E}(D(Z, \beta_0)|X = x)(\hat{p}(x) - p_0(x))d\mathbb{P}_0(x) \\
&= \int \delta(x)(\hat{p}(x) - p_0(x))d\mathbb{P}_0(x),
\end{aligned}$$

where the first equality follows from A5, and the last equality follows by setting $\mathbb{E}[D(Z, \beta_0)|X = x] = \delta(x)$. By plugging in $M(z) = \delta(x)u$, we obtain

$$\begin{aligned}
&\int D(z, \hat{p}(x) - p_0(x))d\mathbb{P}_0(z) - \frac{1}{n}\sum_{i=1}^n M(Z_i) \\
= &\int \delta(x)(\hat{p}(x) - p_0(x))d\mathbb{P}_0(x) - \frac{1}{n}\sum_{i=1}^n \delta(X_i)(Y_i - p_0(X_i)) \\
= &\int \delta(x)(\hat{p}(x) - p_0(x))d\mathbb{P}_0(x) - \int \delta(x)(y - \hat{p}(x) + \hat{p}(x) - p_0(x))d\mathbb{P}_n(z) \\
= &\int -\delta(x)(y - \hat{p}(x))d\mathbb{P}_n(z) + \int -\delta(x)(\hat{p}(x) - p_0(x))d(\mathbb{P}_n - \mathbb{P}_0)(x) \\
= &\ I + II. \qquad (59)
\end{aligned}$$

Lemma 2 guarantees $I = o_p(n^{-1/2})$. For II, by A4 and a similar argument as in p. 23 of BGH-supp, we have $II = o_p(n^{-1/2})$. Thus, Assumption 5.3 in Newey (1994) is satisfied.

Step 4: Conclude the proof. Assumptions 5.4-5.6 in Newey (1994) are adapted as A7-A9 in this paper. Therefore, all the assumptions in Newey (1994, Assumptions 5.1-5.6) are satisfied in our setup.

[4]This is a simplified version of Assumption 5.3, which is mentioned in p.1366 in Newey (1994).

Now the consistency can be proved by similar arguments as in Lemma 5.2 of Newey (1994). Also by Lemma 5.3 of Newey (1994), we obtain $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V)$. Finally, the asymptotic efficiency is proved according to Proposition 4 of Newey (1994).

A.6.3. *Asymptotic properties of $\hat{\tau}$.* The results in Appendix A.6.1 also hold for the UC-isotonic estimator $\tilde{p}(\cdot)$. Recall that in Appendices A.6.1 and A.6.2, Y represents the dependent variable of isotonic regression, and for the sample size n, $\tilde{p}(\cdot)$ is obtained by regression $\tilde{Y}$ on X, where

$$\tilde{Y}_i = \begin{cases} \frac{1}{\lfloor n^{2/3} \rfloor} \sum_{i=1}^{\lfloor n^{2/3} \rfloor} Y_i & \text{for } i \leq \lfloor n^{2/3} \rfloor \\ Y_i & \text{for } \lfloor n^{2/3} \rfloor < i \leq n - \lfloor n^{2/3} \rfloor \\ \frac{1}{\lfloor n^{2/3} \rfloor} \sum_{i=n-\lfloor n^{2/3} \rfloor+1}^{N} Y_i, & \text{for } i > n - \lfloor n^{2/3} \rfloor \end{cases}$$

where the ordering is according to $X_1 < X_2 < \cdots < X_n$.

Lemma 4. *Let $\tilde{p}(\cdot)$ be the UC-isotonic estimator of the conditional mean $\mathbb{E}[Y|X]$, and $\delta(X)$ be a bounded function of X with a finite total variation. Under A1, A2, and A3, it holds*

$$\frac{1}{n}\sum_{i=1}^{n} \delta(X_i)(Y_i - \tilde{p}(X_i)) = o_p(n^{-1/2}). \tag{60}$$

Proof. Note that two key results, (42) and (43) also hold for $\tilde{p}(\cdot)$:

$$||\tilde{p} - p_0||^2 = O_p\left[\left(\frac{\log n}{n}\right)^{2/3}\right] = o_p(n^{-1/2}), \tag{61}$$

$$\sum_{j=1}^{k} m_j \sum_{i=n_j}^{n_{j+1}-1} \{Y_i - \tilde{p}(\cdot)\} = 0, \tag{62}$$

where (61) follows from Theorem 1, and (62) is by Proposition 3(i). Based on (61) and (62), all other arguments in the proof of Lemma 2 in Appendix A.6.2 hold for the UC-isotonic estimator $\tilde{p}(\cdot)$. □

As a result, Lemma 3 also holds for $\tilde{p}(\cdot)$, which is presented as follows.

Lemma 5. *Let $\tilde{p}(\cdot)$ be the UC-isotonic estimator of the conditional mean $\mathbb{E}[Y|X]$, and $\tilde{\beta}$ be the solution of the moment condition $\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \tilde{\beta}, \tilde{p}(\cdot)) = 0$. Under A1-A9, it holds*

$$\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, V),$$

where $V = M_\beta^{-1}\mathbb{E}[\{m(Z,\beta_0,p_0) + M(Z)\}\{m(z,\beta_0,p_0) + M(Z)\}']M_\beta^{-1}$.

We apply Lemma 5 to the ATE model by setting $Z = (Y, W, X)$, $\beta_0 = \mathbb{E}\left[\frac{Y \cdot W}{p_0(X)} - \frac{Y \cdot (1-W)}{1-p(X)}\right]$, $m(Z, \beta, p(\cdot)) = \frac{YW}{p_0(X)} - \frac{Y(1-W)}{1-p(X)} - \beta$, and $p_0(x) = \mathbb{E}[W|X=x] = \mathbb{P}\{W=1|X=x\}$. Let $\tilde{p}(\cdot)$ is the UC-isotonic estimator of the propensity score $p_0(x)$. The plug-in estimator is written as

$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i \cdot W_i}{\tilde{p}(X_i)} - \frac{Y_i \cdot (1-W_i)}{1-\tilde{p}(X_i)}\right\}. \tag{63}$$

The asymptotic distribution of $\tilde{\beta}$ is obtained as follows.

Lemma 6. *Under Assumptions 1-4, it holds* $\tilde{\beta} \xrightarrow{p} \beta_0$ *and*

$$\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, \Omega),$$

where $\Omega = \mathbb{V}(\mathbb{E}[Y(1) - Y(0)]|X) + \mathbb{E}[\mathbb{V}(Y(1)|X)/p_0(X)] + \mathbb{E}[\mathbb{V}(Y(0)|X)/(1 - p_0(X))]$ *is the semiparametric efficiency bound.*

Proof. It is sufficient to check A1-A9 of Lemma 5 in the present setup. Assumption 1 directly implies A1. Assumption 2 and $W \in \{0, 1\}$ imply A2. A3 is satisfied by the fact that $W \in \{0, 1\}$ (note: Y in A3 corresponds to W in this proof). For A4, note that $\mathbb{E}[D(Z, \beta)|X] = -\left(\frac{\mathbb{E}[Y(1)|X]}{p_0(X)} + \frac{\mathbb{E}[Y(0)|X]}{1-p_0(X)}\right)$. It is a bounded function of X with finite total variation by Assumptions 3 and 4 since the continuous differentiability implies a finite total variation. A5 is satisfied since $D(z, \beta, p(x) - p^*(x)) = -\left(\frac{y \cdot w}{p^*(x)^2} + \frac{y(1-w)}{(1-p^*(x))^2}\right)\{p(x) - p^*(x)\}$. A6-A9 are satisfied by the same arguments in pp.26-33 of Hirano, Imbens and Ridder (2000).

Therefore, all assumptions for Lemma 5 are satisfied. The asymptotic variance matrix Ω can be obtained in the same way as pp.34-35 of Hirano, Imbens and Ridder (2000). □

Setting the sample size as $n = N$, Theorem 3 follows from Theorem 2 and Lemma 6, and the conclusion is obtained.

A.7. **Proof of Theorem 4.** The proof is similar to that of Theorem 1. Noting that $\tilde{\alpha} - \alpha_0 = O_p(N^{-1/2}) = o_p(N^{-1/3})$, so the estimation of α_0 does not affect the uniform convergence rate. All the steps in Section A.4 can be similarly applied with plugged-in $\tilde{\alpha}$.

A.8. **Proof of Theorem 5.** Note that Proposition 3 and Corollary 1 also hold for the UC-iso-index estimator $\tilde{p}_{\tilde{\alpha}}$. By a similar argument for the proof of Theorem 2 (in Appendix A.6),

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(Y_i - \frac{1}{M_i}\sum_{j \in \mathcal{J}(i)} Y_j\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i Y_i}{\tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha})} - \frac{(1 - W_i)\, Y_i}{1 - \tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha})}\right). \tag{64}$$

It remains to derive the asymptotic properties of (64). In Appendix A.8.1, we summarize the generic results for semiparametric estimation with plug-in (standard) monotone single index estimator discussed in Xu (2021), and introduce relevant assumptions and lemmas. Their proofs are presented in Appendix A.8.2.

Appendix A.8.3 shows that the results in Appendix A.8.1 also hold for the plug-in UC-iso-index estimator. Then we show the asymptotic properties of (64) and conclude the proof.

A.8.1. *Relevant results on semiparametric estimation with plug-in monotone single index estimator.* For the moment condition (39), we think about a general case that $\mathbb{E}[T(Z, \beta_0)|X] = p_0(X) = F_0(X'\alpha_0)$, and $F_0(u)$ is a monotone increasing function of its index u. This structure encompasses many semiparametric model where the dependent variable of a monotone conditional mean function might contain β_0. For example, in a partially linear monotone index model $Y = X_1'\beta_0 + F_0(X_2'\alpha_0) + \varepsilon$, we have $T(Z, \beta_0) = Y - X_1'\beta_0$. It also encompasses simpler cases: in a monotone single index model, $Y = F_0(X'\alpha_0) + \varepsilon$, we set $T(Z, \beta_0) = Y$; in Section 3, we set $T(Z, \beta_0) = W$.

For identification, α_0 is a k_α-dimensional vector normalized with $||\alpha_0||= 1$. Let k_β be the dimension of β and the moment function $m(Z, \beta_0, p_0(\cdot))$. To implement isotonic estimation to the link function $p_0(\cdot) = F_0(\cdot'\alpha_0)$, we need that the monotonicity holds in the neighbors of the true values α_0 and β_0. We denote $\theta = (\alpha', \beta')' \in \Theta \equiv \mathcal{S}_{k_\alpha - 1} \times \mathbb{R}^{k_\beta}$. For a fixed θ, we define $F_\theta(u) = F_{\alpha,\beta}(u) = \mathbb{E}(T(Z, \beta)|\alpha' X = u)$. Let $F_\theta(\cdot) = F_\theta(\cdot'\alpha)$, and we have by definition $F_0(\cdot) = F_{\theta_0}(\cdot'\alpha_0)$. For the sample n (not N), the problem is to solve

$$\hat{F}_{\alpha,\beta} = \arg\min_{F \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \{T(Z_i, \beta) - F(X_i'\alpha)\}^2, \tag{65}$$

$$\hat{\theta} = (\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} ||\frac{1}{n} \sum_{i=1}^{n} m(Z_i, \beta, \hat{F}_{\alpha,\beta}(X_i'\alpha))||^2. \tag{66}$$

Assumptions A1-A7 in Appendix A.6.1 are modified as follows.

A1': X is a random vector taking value in the space $\mathcal{X} \subset \mathbb{R}^{k_\alpha}$. The space $\mathcal{X}$ is convex with non-empty interiors, and satisfies $\mathcal{X} \subset \mathcal{B}(0, R)$ for some $R > 0$.

A2': There exists $\delta_0 > 0$ that for each $\theta \in \mathcal{B}(\theta_0, \delta_0)$, the function $u \mapsto \mathbb{E}[T(Z, \beta)|X'\alpha = u]$ is monotone increasing in u and differentiable in θ. There exists $K_0 > 0$ such that $|F_0(\cdot)|< K_0$ for all $x \in \mathcal{X}$.

A3': There exist $c_0 > 0$ and $M_0 > 0$ such that $\mathbb{E}[|T(Z, \beta)|^m|X = x] \leq m!\, M_0^{m-2} c_0$ for all integers $m \geq 2$ and almost every x and $\beta \in \mathcal{B}(\beta_0, \delta_0)$.

Similar to Lemma 2, we have the following lemma.

Lemma 7. *For a fixed $\theta \in \mathcal{B}(\theta_0, \delta_0)$, $\hat{F}_{\alpha,\beta}(\cdot)$ are solved by solving* (65)*, and $\delta(u)$ is a bounded function of u with a finite total variation. Under A1'-A3', it holds*

$$\frac{1}{n} \sum_{i=1}^{n} \delta(X_i'\alpha)\{T(Z_i, \beta) - \hat{F}_{\alpha,\beta}(X_i'\alpha)\} = o_p(n^{-1/2}).$$

Furthermore, we add these assumptions.

A4': For all $\theta \in \Theta$, $u \mapsto \mathbb{E}[D(Z, \beta)|X'\alpha = u]$ is a bounded function of u with a finite total variation. There exist $c_1 > 0$ and $M_1 > 0$ such that for each row of $D(Z, \beta)$ ($D_j(Z, \beta)$ with $j \in \{1 : k_\beta\}$), $\mathbb{E}[|D_j(Z, \beta)|^m|X = x] \leq m!\, M_1^{m-2} c_1$ for all integers $m \geq 2$ and almost every x.

A5': The first-order expansion of $m(z, \beta, p(\cdot))$ with respect to $p(\cdot)$ at $p^*(\cdot)$, $D(z, \beta, p(\cdot) - p^*(\cdot))$, is linear in $p(\cdot) - p^*(\cdot)$ (i.e., $D(z, \beta, p(x) - p^*(x)) = D(z, \beta)\{p(x) - p^*(x)\}$).

A6': There are $b(z) > 0$ and $D(z, g)$ that (i) $||m(z, \beta, F_\theta) - m(z, \beta, F_0) - D(z, \beta, F_\theta - F_0)|| \leq b(z)||F_\theta - F_0||^2$; (ii) $\mathbb{E}[b(Z)] = o_p(n^{1/6}(\log n)^{-2})$, for all $\theta \in \Theta$, where Θ is compact.

A7': There are ε, $b(z)$, $\tilde{b}(z) > 0$ and $F(\cdot)$ with $||F||> 0$. Such that (i) for all $\theta \in \Theta$, $m(z, \beta, F_\theta)$ is continuous at θ and $m(z, \beta, F_\theta) \leq b(z)$; (ii) $||m(z, \beta, F) - m(z, \beta, F_\theta)|| \leq \tilde{b}(z)(||F - F_\theta||)^\varepsilon$.

Let $m_1(z,\beta,F_\theta) = x\{T(z,\beta) - F_\theta(x'\alpha)\}$, $m^*(z,\beta,F_\theta) = \begin{pmatrix} m(z,\beta,F_\theta) \\ m_1(z,\beta,F_\theta) \end{pmatrix}$, and

$$\begin{aligned}
M_\alpha &= -\mathbb{E}\left\{[D(Z,\beta_0) - \mathbb{E}[D(Z,\beta_0)|X'\alpha_0]]\{X - \mathbb{E}[X|X'\alpha_0]\}'F_0^{(1)}(X'\alpha_0)\right\}, \\
M_\beta &= -\mathbb{E}\left\{\frac{\partial m(Z,\beta_0,F_0(X'\alpha_0))}{\partial\beta} + \mathbb{E}[D(Z,\beta_0)|X'\alpha_0]\frac{\partial T(Z,\beta_0)}{\partial\beta}\right\}, \\
M_\theta &= -\mathbb{E}\left\{\frac{dm^*(Z,\beta_0,F_{\theta_0})}{d\theta}\right\}, \\
M(Z) &= \mathbb{E}(D(Z,\beta_0)|X'\alpha_0)(T(Z,\beta_0) - F_0(X'\alpha_0)),
\end{aligned} \tag{67}$$

and denote $M_{\alpha,1}$ as M_α corresponding to the moment function m_1. The modified versions of A8 and A9 are presented as follows.

A8': $\mathbb{E}[m^*(z,\beta,F_\theta)] = 0$ has a unique solution on Θ at θ_0.

A9': For $\theta \in$ interior(Θ), (i) there are $\varepsilon > 0$ and a neighborhood $\mathscr{N}$ of β_0 such that for all $||F - F_0|| \leq \varepsilon$, $m(z,\beta,F)$ is differentiable in β on $\mathscr{N}$; (ii) M_β is nonsingular; (iii) $M_{\alpha,1}$ has rank $k_\alpha - 1$, and M_θ has rank $k_\alpha + k_\beta - 1$ (iv) $\mathbb{E}[||m^*(Z,\beta,F_\theta)||^2] < \infty$; (v) Assumption A7 is satisfied with $m(z,\beta,p)$ equaling to each row of $\frac{dm^*(z,\beta,F_\theta)}{d\theta}$.

Under these assumptions, the asymptotic distributions of $\hat{\alpha}$ and $\hat{\beta}$ are obtained as follows.

Lemma 8. *Under A1'-A9', it holds*

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N(0, V_\alpha), \qquad \sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_\beta),$$

where

$$\begin{aligned}
V_\beta &= M_\beta^{-1}\mathbb{E}[\{m(Z,\beta_0,p_0) + M(Z) + A(Z)\}\{m(Z,\beta_0,p_0) + M(Z) + A(Z)\}']M_\beta^{-1}, \\
V_\alpha &= M_{\alpha,1}^{-}\mathbb{E}[\{m_1(Z,\beta_0,p_0) + M_1(Z) + B_1(Z)\}\{m_1(Z,\beta_0,p_0) + M_1(Z) + B_1(Z)\}']M_{\alpha,1}^{-},
\end{aligned}$$

$M_{\alpha,1}^{-}$ is the Moore-Penrose inverse of $M_{\alpha,1}$, and A, B_1, and M_1 are defined in Appendix A.8.2.

A.8.2. *Proofs of the results in Appendix A.8.1.*

Proof of Lemma 7. The proof is similar to that on pp. 18-20 of the supplementary material of BGH (hereafter, BGH-supp) and that for Lemma 2. We replace $\mathbb{E}[X|S(\beta)'X]$ and Y in BGH-supp with $\delta(X'\alpha)$ and $T(Z,\beta)$ in our setting, respectively.

Proof of Lemma 8. Now the nuisance function $\hat{F}_{\hat{\alpha},\hat{\beta}}(x'\hat{\alpha})$ depends on $\hat{\alpha}$ and $\hat{\beta}$. A similar argument to Balabdaoui and Groeneboom (2021) yields[5]

$$\left\| \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) \right\| = o_p(n^{-1/2}). \tag{68}$$

[5]This equation mainly concerns the case where the isotonic index estimator depends on the parameter β such that the sample moment condition $\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) = 0$ does not have a root for given n. See Balabdaoui and Groeneboom (2021) for more discussion. In the main paper, the link function estimated in the first stage does not involve the second stage τ, so the right hand side of (68) can take zero.

Under A7' and A8', the consistency of $\hat{\theta}$ can be proved by similar arguments as in Lemma 5.2 of Newey (1994). Define $\mathbb{E}[\cdot|u] = \mathbb{E}[\cdot|X'\hat{\alpha} = u]$,

$$\begin{aligned} M_{n,\beta} &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial m(Z_i,\beta_0,F_0(X_i'\alpha_0))}{\partial\beta} + \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\frac{\partial T(Z_i,\beta_0)}{\partial\beta}\right\}, \\ M_{\beta} &= -\mathbb{E}\left\{\frac{\partial m(Z,\beta_0,F_0(X'\alpha_0))}{\partial\beta} + \mathbb{E}[D(Z,\beta_0)|X'\alpha_0]\frac{\partial T(Z,\beta_0)}{\partial\beta}\right\}. \end{aligned}$$

Observe that

$$\begin{aligned} o_p(n^{-1/2}) &= \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\hat{\beta},\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{m(Z_i,\hat{\beta},\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) + \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\{T(Z_i,\hat{\beta}) - \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})\}\right\} + o_p(n^{-1/2}) \\ &= -M_{n,\beta}(\hat{\beta}-\beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta_0,F_0(X_i'\alpha_0)) + o_p(n^{-1/2} + (\hat{\beta}-\beta_0)) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left\{D(Z_i,\beta_0)\{\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}) - F_0(X_i'\alpha_0)\} + \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\{T(Z_i,\beta_0) - \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})\}\right\} \\ &= -M_{\beta}(\hat{\beta}-\beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta_0,F_0) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\{D(Z_i,\beta_0) - \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\}\{\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}) - F_0(X_i'\alpha_0)\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[D(Z_i,\beta_0)|X_i'\alpha_0]\{T(Z_i,\beta_0) - F_0(X_i'\alpha_0)\} + o_p(n^{-1/2} + (\hat{\beta}-\beta_0)), \end{aligned} \tag{69}$$

where the first equality follows from (68), the second equality follows from Lemma 7, the third equality follows from expanding $m(Z_i,\hat{\beta},\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) + \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]T(Z_i,\hat{\beta})$ around β_0 and F_0 with the derivatives $M_{n,\beta}$ and $D(Z_i,\beta_0)$, and some rearrangements, and the last equality follows from $M_{n,\beta} - M_{\beta} = o_p(1)$ and

$$\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}[D(Z_i,\beta_0)|X_i'\alpha_0] - \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\}\{T(Z_i,\beta_0) - F_0(X_i'\alpha_0)\} = o_p(n^{-1/2}),$$

which can be shown by a similar argument about (C.20) in pp.21-22 of BGH-supp.

The second term in the last equality of (69) can be rewritten as:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0) - \mathbb{E}(D(Z_i,\beta_0)|X_i'\hat{\alpha})](\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}) - F_0(X_i'\alpha_0))\right\} \\ =\ &\frac{1}{n}\sum_{i=1}^{n}\{D(Z_i,\beta_0) - \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\}\{\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}) - F_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})\} \\ &+\frac{1}{n}\sum_{i=1}^{n}\{D(Z_i,\beta_0) - \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\}\{F_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}) - F_0(X_i'\alpha_0)\} \\ =\ &I_m + II_m. \end{aligned}$$

A similar argument about (C.22) in p.23 of BGH-supp implies $I_m = o_p(n^{-1/2})$. For II_m, Lemma 17 of BGH-supp implies

$$\begin{aligned}\frac{\partial}{\partial \alpha^j} F_{\alpha,\hat{\beta}}(X'\alpha)\Big|_{\alpha=\alpha_0} &= (x^j - \mathbb{E}[X^j|X'\alpha_0 = x'\alpha_0])F^{(1)}_{0,\hat{\beta}}(x'\alpha_0), \\ &= (x^j - \mathbb{E}[X^j|X'\alpha_0 = x'\alpha_0])F^{(1)}_{0}(x'\alpha_0) + O_p(\hat{\beta} - \beta_0),\end{aligned}$$

where α^j and x^j are the j-th elements of α and x, respectively. Thus, an expansion of II_m around α_0 yields

$$\begin{aligned}II_m &= \frac{1}{n}\sum_{i=1}^n \{D(Z_i,\beta_0) - \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\}\{X_i - \mathbb{E}[X_i|X_i'\alpha_0]\}'F_0^{(1)}(X_i'\alpha_0) + O_p(\hat{\beta}-\beta_0)(\hat{\alpha}-\alpha_0) \\ &\quad + o_p(\hat{\alpha}-\alpha_0) \\ &= \frac{1}{n}\sum_{i=1}^n \{D(Z_i,\beta_0) - \mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\}\{X_i - \mathbb{E}[X_i|X_i'\alpha_0]\}'F_0^{(1)}(X_i'\alpha_0)(\hat{\alpha}-\alpha_0) + o_p(\hat{\alpha}-\alpha_0) \\ &= \mathbb{E}\left[\{D(Z,\beta_0) - \mathbb{E}[D(Z,\beta_0)|X'\alpha_0]\}\{X - \mathbb{E}\left(X|X'\alpha_0\right)\}'F_0^{(1)}(X'\alpha_0)\right](\hat{\alpha}-\alpha_0) + o_p(\hat{\alpha}-\alpha_0), \quad (70)\end{aligned}$$

where the second equality follows from $\hat{\beta} - \beta_0 = o_p(1)$, and the last equality follows from $\hat{\alpha} - \alpha_0 = o_p(1)$ and $\mathbb{E}[D(Z_i,\beta_0)|X_i'\hat{\alpha}] - \mathbb{E}[D(Z_i,\beta_0)|X_i'\alpha_0] = o_p(1)$. Now let us define

$$M(Z) = \mathbb{E}[D(Z,\beta_0)|X'\alpha_0]\{T(Z,\beta_0) - F_0(X'\alpha_0)\} \quad (71)$$

$$M_\alpha = -\mathbb{E}\left[\{D(Z,\beta_0) - \mathbb{E}[D(Z,\beta_0)|X'\alpha_0]\}\{X - \mathbb{E}[X|X'\alpha_0]\}'F_0^{(1)}(X'\alpha_0)\right]. \quad (72)$$

By (70) and (72) with (69), we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n m(Z_i,\hat{\beta},\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) \\ =\ & -M_\beta(\hat{\beta}-\beta_0) - M_\alpha(\hat{\alpha}-\alpha_0) + \frac{1}{n}\sum_{i=1}^n m(z_i,\beta_0,F_0) \\ &+\frac{1}{n}\sum_{i=1}^n M(Z_i) + o_p(n^{-1/2} + (\hat{\beta}-\beta_0) + (\hat{\alpha}-\alpha_0)). \quad (73)\end{aligned}$$

Combining the facts $\mathbb{E}[m(Z,\beta_0,F_0)] = 0$ and $\mathbb{E}[M(Z)] = 0$ with A3', A4', and A9', we have $\frac{1}{n}\sum_{i=1}^n m(z_i,\beta_0,F_0) + \frac{1}{n}\sum_{i=1}^n M(Z_i) = O_p(n^{-1/2})$. Then (69) and (73) imply $\hat{\alpha}-\alpha_0 = O_p(n^{-1/2})$ and $\hat{\beta}-\beta_0 = O_p(n^{-1/2})$. Besides, from (73) we can see that $\hat{\alpha}-\alpha_0$ and $\hat{\beta}-\beta_0$ are asymptotically linear. Thus, we can rewrite the first term in the last row as

$$-M_\alpha(\hat{\alpha}-\alpha_0) = \frac{1}{n}\sum_{i=1}^n A(Z_i) + o_p(n^{-1/2}), \quad (74)$$

with $\mathbb{E}[A(Z_i)] = 0$. Similarly, we can rewrite

$$-M_\beta(\hat{\beta}-\beta_0) = \frac{1}{n}\sum_{i=1}^n B(Z_i) + o_p(n^{-1/2}),$$

with $\mathbb{E}[B(Z_i)] = 0$.

Now we can rewrite (73) to obtain asymptotic expressions of $\hat{\alpha}$ and $\hat{\beta}$. Note that given β, $\hat{\alpha}$ solves $\hat{\alpha} = \arg\min_\alpha ||\frac{1}{n}\sum_{i=1}^n X_i'\{T(Z_i,\beta) - \hat{F}_\alpha(X_i'\alpha)\}||^2$, which corresponds to the solution of the moment condition

$$m_1(Z,\beta,F(X'\alpha)) = X\{T(Z,\beta) - F(X'\alpha)\}. \tag{75}$$

We can express $\sqrt{n}(\hat{\alpha}-\alpha_0)$ by replacing m in (73) by m_1, i.e.,

$$\begin{aligned}\sqrt{n}(\hat{\alpha}-\alpha_0) &= M_{\alpha,1}^- \frac{1}{\sqrt{n}}\sum_{i=1}^n \{m_1(Z_i,\beta_0,p_0) + B_1(Z_i) + M_1(Z_i)\} \\ &= M_{\alpha,1}^- \frac{1}{\sqrt{n}}\sum_{i=1}^n \{X_i - \mathbb{E}[X_i|X_i'\alpha_0]\}\left\{T(Z_i,\beta_0) + \frac{\partial T(Z_i,\beta_0)}{\partial\beta}(\hat{\beta}-\beta_0) - F_0(X'\alpha_0)\right\}, \end{aligned} \tag{76}$$

where $M_{\alpha,1}$, B_1, and M_1 are counterparts of M_α, B, and M, respectively, for the moment function m_1, and $M_{\alpha,1}^-$ is the Moore-Penrose inverse of $M_{\alpha,1}$. Combining this with (73), we obtain the conclusion $\sqrt{n}(\hat{\alpha}-\alpha_0) \xrightarrow{d} N(0,V_\alpha)$ and $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V_\beta)$.

A.8.3. *Asymptotic properties of $\tilde{\tau}$.* The results in Appendix A.8.1 also hold for the plug-in UC-iso-index estimator by applying similar arguments in Appendix A.6.3. Define

$$\tilde{T}_i(Z_i,\beta) = \begin{cases} \frac{1}{\lfloor n^{2/3}\rfloor}\sum_{i=1}^{\lfloor n^{2/3}\rfloor} T(Z_i,\beta) & \text{for } i \le \lfloor n^{2/3}\rfloor \\ T(Z_i,\beta) & \text{for } \lfloor n^{2/3}\rfloor < i \le n - \lfloor n^{2/3}\rfloor \\ \frac{1}{\lfloor n^{2/3}\rfloor}\sum_{i=n-\lfloor n^{2/3}\rfloor+1}^{n} T(Z_i,\beta) & \text{for } i > n - \lfloor n^{2/3}\rfloor \end{cases}, \tag{77}$$

which are ordered by $X_1'\alpha < X_2'\alpha < \cdots < X_n'\alpha$.

Lemma 9. *For a given $\theta \in \mathcal{B}(\theta_0,\delta_0)$, let $\tilde{F}_{\alpha,\beta}(\cdot) = \arg\min_{F\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^n\{\tilde{T}_i(Z_i,\beta) - F(X_i'\alpha)\}^2$ and $\delta(u)$ be a bounded function of u with a finite total variation. Under A1'-A3', it holds*

$$\frac{1}{n}\sum_{i=1}^n \delta(X_i'\alpha)\{T(Z_i,\beta) - \tilde{F}_{\alpha,\beta}(X_i'\alpha)\} = o_p(n^{-1/2}).$$

Proof. The proof is similar to that of Lemma 4. We note that two key results hold for the UC-iso-index estimator

$$||\tilde{F}_{\alpha,\beta}(\cdot) - F_{\alpha,\beta}(\cdot)||^2 = O_p\left(\left(\frac{\log n}{n}\right)^{2/3}\right) = o_p(n^{-1/2}), \tag{78}$$

$$\sum_{j=1}^k m_j \sum_{i=n_j}^{n_{j+1}-1}\{T(Z_i,\beta) - \tilde{F}_{\alpha,\beta}(X_i'\alpha)\} = 0. \tag{79}$$

By taking $T(Z_i,\beta)$ and $F_{\alpha,\beta}(X_i'\alpha)$ as dependent variables for the conditional mean functions, (78) follows from Theorem 1 and Proposition 4 of BGH, and (79) follows from Proposition 3(i).

The rest arguments are similar to that on pp. 18-20 of BGH-supp and that for Lemma 2, where $\mathbb{E}[X|S(\beta)'X_i]$ and Y_i in BGH-supp correspond to $\delta(X_i'\alpha)$ and $T(Z_i,\beta)$ in our setting, respectively. □

The semiparametric estimator with plug-in UC-iso-index estimator is obtained as follows.

(1) Compute $\hat{\alpha}$ and $\hat{\beta}$ by solving (65) and (66).

(2) Let $\tilde{\alpha} = \hat{\alpha}$, and transform the sample $\{T(Z_i, \beta), X_i\}_{i=1}^n$ indexed by $X_1'\tilde{\alpha} < \cdots < X_n'\tilde{\alpha}$ into $\{\tilde{T}_i(Z_i, \beta), X_i\}_{i=1}^n$ with (77).

(3) Compute the UC-iso-index estimator $\tilde{F}_{\tilde{\alpha},\tilde{\beta}}$ by solving

$$\tilde{F}_{\tilde{\alpha},\beta} = \arg\min_{F\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^n \{\tilde{T}_i(Z_i,\beta) - F(X_i'\alpha)\}^2, \tag{80}$$

$$\tilde{\beta} = \arg\min_{\beta} ||\frac{1}{n}\sum_{i=1}^n m(Z_i, \beta, \tilde{F}_{\tilde{\alpha},\beta}(X_i'\alpha))||^2. \tag{81}$$

Similar to Lemma 8, the asymptotic properties of the UC-iso-index estimator $(\tilde{\alpha}, \tilde{\beta})$ are obtained as follows.

Lemma 10. *Under A1'-A9', it holds*

$$\sqrt{n}(\tilde{\alpha} - \alpha_0) \xrightarrow{d} N(0, V_\alpha), \qquad \sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, V_\beta),$$

where

$$\begin{aligned} V_\beta &= M_\beta^{-1}\mathbb{E}[\{m(Z,\beta_0,p_0) + M(Z) + A(Z)\}\{m(z,\beta_0,p_0) + M(Z) + A(Z)\}']M_\beta^{-1}, \\ V_\alpha &= M_{\alpha,1}^{-}\mathbb{E}[\{m_1(Z,\beta_0,p_0) + M_1(Z) + B_1(Z)\}\{m_1(Z,\beta_0,p_0) + M_1(Z) + B_1(Z)\}']M_{\alpha,1}^{-}, \end{aligned}$$

$M_{\alpha,1}^-$ is the Moore-Penrose inverse of $M_{\alpha,1}$, and A, B_1, and M_1 are defined in Appendix A.8.2.

Now we apply Lemma 10 to the ATE model. In the following equations, the left hand sides are notations for Appendices A.8.1 and A.8.2, and the right hand sides are notations in Section 3:

$$\begin{aligned} Z &= (Y, W, X), \\ \beta_0 &= \mathbb{E}\left[\frac{YW}{p_0(X'\alpha_0)} - \frac{Y(1-W)}{1-p(X'\alpha_0)}\right], \\ F_\theta(\cdot'\alpha) = F_{\alpha,\beta}(\cdot'\alpha) &= p_\alpha(\cdot'\alpha), \\ m(Z,\beta,F_\theta(X'\alpha)) &= \frac{YW}{p_\alpha(X'\alpha)} - \frac{Y(1-W)}{1-p_\alpha(X'\alpha)} - \beta, \end{aligned} \tag{82}$$

and the propensity score

$$p_0(X'\alpha_0) = \mathbb{E}[W|X = x] = \mathbb{P}(W = 1|X = x),$$

has a monotone increasing link function p_0.

From (82), we see that this case is simpler than the general framework discussed in Appendix A.8.1 in that the monotone single index estimator does not depend on β. Therefore, we can simply write the UC-iso-index estimator as $\tilde{p}_{\tilde{\alpha}}(\cdot)$, and the plug-in estimator of interest is written as

$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^n \left\{\frac{Y_iW_i}{\tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha})} - \frac{Y_i(1-W_i)}{1-\tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha})}\right\}. \tag{83}$$

The asymptotic properties of $\tilde{\beta}$ is obtained as follows.

Lemma 11. *Under Assumptions 1'-4', 5, and 6, it holds* $\tilde{\beta} \xrightarrow{p} \beta_0$ *and*

$$\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma = \mathbb{E}\left[\{m(Z) + M(Z) + A(Z)\}\{m(Z) + M(Z) + A(Z)\}'\right]$, *and*

$$\begin{aligned}
m(Z) &= \frac{YW}{p_0(X'\alpha_0)} - \frac{Y(1-W)}{1-p_0(X'\alpha_0)} - \beta_0, \\
D(Z) &= -\left(\frac{YW}{p_0(X'\alpha_0)^2} + \frac{Y(1-W)}{(1-p_0(X'\alpha_0))^2}\right), \qquad M(Z) = -\mathbb{E}[D(Z)|X'\alpha_0]\{W - p_0(X'\alpha_0)\}, \\
A(Z) &= \mathbb{E}\left[\{D(Z) - \mathbb{E}[D(Z)|X'\alpha_0]\}\{X - \mathbb{E}[X|X'\alpha_0]\}' p_0^{(1)}(X'\alpha_0)\right] \\
&\quad \times \mathbb{E}[p_0^{(1)}(X'\alpha_0) Cov(X|X'\alpha_0)]^{-}\{X - \mathbb{E}[X|X'\alpha_0]\}\{W - p_0(X'\alpha_0)\}.
\end{aligned}$$

Proof. It is sufficient to check A1'-A9' of Lemma 10 for $m(Z, \beta_0, p(\cdot'\alpha_0)) = \frac{Y \cdot W}{p_0(X'\alpha_0)} - \frac{Y \cdot (1-W)}{1-p_0(X'\alpha_0)} - \beta_0$. Note that (Y, X) in Lemma 10 corresponds to (W, X) in this lemma.

Assumption 1' directly implies A1'. Assumption 2' and $W \in \{0, 1\}$ imply A2'. A3' is satisfied by the fact that $W \in \{0, 1\}$. For A4', note that $\mathbb{E}[D(Z, \beta)|X'\alpha = u] = -\mathbb{E}\left(\frac{\mathbb{E}[Y(1)|X]}{p_0(X'\alpha_0)} + \frac{\mathbb{E}[Y(0)|X]}{1-p_0(X'\alpha_0)}|X'\alpha = u\right)$ in this lemma. It is a bounded function of X with finite total variation by Assumptions 3 and 4, since the continuous differentiability implies a finite total variation. A5' is satisfied since we have $D(z, \beta, p(x) - p^*(x)) = -\left(\frac{yw}{p^*(x)^2} + \frac{y(1-w)}{(1-p^*(x))^2}\right)\{p(x) - p^*(x)\}$. A6' and A7' are satisfied by the same arguments in pp.26-33 of Hirano, Imbens and Ridder (2000). A8' and A9' are satisfied by Assumptions 5 and 6 and the same arguments in pp.26-33 of Hirano, Imbens and Ridder (2000).

Since all the assumptions for Lemma 10 are satisfied, we can apply Lemma 10 to the estimator (83).

It remains to derive the concrete expressions of $m(Z)$, $M(Z)$, and $A(Z)$. $m(Z)$ is obtained directly from $\beta_0 = \mathbb{E}\left[\frac{YW}{p_0(X'\alpha_0)} - \frac{Y(1-W)}{1-p(X'\alpha_0)}\right]$, and $M(Z)$ can be obtained directly from (71). Now we derive $A(Z)$. Since $\tilde{\alpha} = \hat{\alpha}$, combining (72), (75), and (76) yields

$$\hat{\alpha} - \alpha_0 = \mathbb{E}[p_0^{(1)}(X'\alpha_0)\mathrm{Cov}(X|X'\alpha_0)]^{-}\frac{1}{n}\sum_{i=1}^{n}\{X_i - \mathbb{E}[X|X_i'\alpha_0]\}\{W_i - p_0(X_i'\alpha_0)\} + o_p(n^{-1/2}).$$

Combining this with (74) and (72), we have the expression of $A(Z)$. □

By setting the sample size as $n = N$, Theorem 5 follows from Theorem 2 and Lemma 11.

A.9. **Proof of Theorem 6.** The proof is based on Groeneboom and Hendrickx (2017) (hereafter GH). By Theorem 2, it is sufficient to show validity of the bootstrap approximation for $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i Y_i}{\tilde{p}(X_i)} - \frac{(1-W_i)Y_i}{1-\tilde{p}(X_i)}\right)$. Define

$$\begin{aligned}
m(Z, \tau, p(\cdot)) &= \frac{YW}{p(X)} - \frac{Y(1-W)}{1-p(X)} - \tau, \\
D(Z) &= -\left(\frac{YW}{p(X)^2} + \frac{Y(1-W)}{(1-p(X))^2}\right), \qquad M(Z) = \mathbb{E}[D(Z)|X]\{W - p(X)\}.
\end{aligned}$$

Let $\{Z_i\}_{i=1}^N = \{Y_i, W_i, X_i\}_{i=1}^N$ be the original sample and $\{Z_i^*\}_{i=1}^N$ be its bootstrap resample. Define $\tilde{p}^*(\cdot)$ and $\hat{\tau}^*$ as the UC-isotonic estimator of the propensity score and the corresponding

ATE estimator by $\{Z_i^*\}_{i=1}^N$, respectively. By (38), $\hat{\tau}^*$ solves

$$\frac{1}{N}\sum_{i=1}^{N} m(Z_i^*, \hat{\tau}^*, \tilde{p}^*(X_i^*)) = 0. \tag{84}$$

In the bootstrap world, the L_2-convergence result is obtained as

$$\frac{1}{N}\sum_{i=1}^{N}\{\tilde{p}^*(X_i^*) - p(X_i^*)\}^2 = O_{P_M}((\log n)^2 n^{-2/3}) = o_{P_M}(N^{-1/2}), \tag{85}$$

where P_M is the probability measure in the bootstrap world defined in p. 3450 of GH. Furthermore, a similar result to Lemma 4 is obtained as

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[D(Z_i^*)|X_i^*]\{W_i^* - \tilde{p}(X_i^*)\} = o_{P_M}(N^{-1/2}). \tag{86}$$

By expanding (84), we have

$$\begin{aligned}
0 &= \frac{1}{N}\sum_{i=1}^{N} m(Z_i^*, \hat{\tau}^*, \tilde{p}^*(X_i^*)) \\
&= \frac{1}{N}\sum_{i=1}^{N} m(Z_i^*, \hat{\tau}^*, \tilde{p}^*(X_i^*)) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(D(Z_i^*)|X_i^*)\{W_i^* - \tilde{p}^*(X_i^*)\} + o_{P_M}(N^{-1/2}) \\
&= -(\hat{\tau}^* - \tau) + \frac{1}{N}\sum_{i=1}^{N} m(Z_i^*, \tau, \tilde{p}^*(X_i^*)) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[D(Z_i^*)|X_i^*]\{W_i^* - \tilde{p}^*(X_i^*)\} + o_{P_M}(N^{-1/2} + (\hat{\tau}^* - \tau)) \\
&= -(\hat{\tau}^* - \tau) + \frac{1}{N}\sum_{i=1}^{N} m(Z_i^*, \tau, p(X_i^*)) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[D(Z_i^*)|X_i^*]\{W_i^* - p(X_i^*)\} + o_{P_M}(N^{-1/2} + (\hat{\tau}^* - \tau)) \\
&= -(\hat{\tau}^* - \tau) + \frac{1}{N}\sum_{i=1}^{N}\{m(Z_i^*, \tau, p(X_i^*)) + M(Z_i^*)\} + o_{P_M}(N^{-1/2} + (\hat{\tau}^* - \tau)),
\end{aligned}$$

where the fourth equality follows from (86) and the second step of the proof of Lemma 3 (in Appendix A.6.2).

This result implies

$$\begin{aligned}
\hat{\tau}^* - \tau &= \left\{\frac{1}{N}\sum_{i=1}^{N} m(Z_i^*, \tau, p(X_i^*)) - \frac{1}{N}\sum_{i=1}^{N} m(Z_i, \tau, p(X_i))\right\} + \left\{\frac{1}{N}\sum_{i=1}^{N} M(Z_i^*) - \frac{1}{N}\sum_{i=1}^{N} M(Z_i)\right\} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\{m(Z_i, \tau, p(X_i) + M(Z_i)\} + o_{P_M}(n^{-1/2} + (\hat{\tau}^* - \tau)).
\end{aligned} \tag{87}$$

From the proof of Lemma 3, we also have

$$\hat{\tau} - \tau = \frac{1}{N}\sum_{i=1}^{N}\{m(Z_i, \tau, p(X_i)) + M(Z_i)\} + o_p(N^{-1/2}). \tag{88}$$

(Recall: by Theorem 2 and setting $n = N$, $\hat{\beta}$ in Lemma 3 is equivalent to $\hat{\tau}$ in the above equation.) Subtracting (88) from (87),

$$\begin{aligned}\hat{\tau}^* - \hat{\tau} &= \left\{\frac{1}{N}\sum_{i=1}^{N} m(Z_i^*, \tau, p(X_i^*)) - \frac{1}{N}\sum_{i=1}^{N} m(Z_i, \tau, p(X_i))\right\} + \left\{\frac{1}{N}\sum_{i=1}^{N} M(Z_i^*) - \frac{1}{N}\sum_{i=1}^{N} M(Z_i)\right\} \\ &\quad + o_{P_M}((\hat{\tau}^* - \tau) + N^{-1/2}) + o_p(N^{-1/2}),\end{aligned}$$

Note that $\mathbb{E}_{P_M}[m(Z_i^*, \tau, p(X_i^*))] = \frac{1}{N}\sum_{i=1}^{N} m(Z_i, \tau, p(X_i))$ and $\mathbb{E}_{P_M}[M(Z_i^*)] = \frac{1}{N}\sum_{i=1}^{N} M(Z_i)$, where $\mathbb{E}_{P_M}[\cdot]$ is expectation under P_M. Therefore, a central limit theorem yields $\sqrt{n}(\hat{\tau}^* - \hat{\tau}) \xrightarrow{d} N(0, \Omega)$, where Ω is defined in Theorem 3, and the conclusion is obtained.

References

[1] Abadie, A. and G. W. Imbens (2006) Large sample properties of matching estimators for average treatment effects, *Econometrica*, 74, 235-267.

[2] Abadie, A. and G. W. Imbens (2008) On the failure of the bootstrap for matching estimators, *Econometrica*, 76, 1537-1557.

[3] Abadie, A. and G. W. Imbens (2011) Bias-corrected matching estimators for average treatment effects, *Journal of Business & Economic Statistics*, 29, 1-11.

[4] Abadie, A. and G. W. Imbens (2016) Matching on the estimated propensity score, *Econometrica*, 84, 781-807.

[5] Adusumilli, K. (2020) Bootstrap inference for propensity score matching, Working paper.

[6] Ai, C. and Chen, X. (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6), 1795-1843.

[7] Andrews, D. W. K. (1994) Asymptotics for semiparametric econometric models via stochastic equicontinuity, *Econometrica*, 62, 43-72.

[8] Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and E. Silverman (1955) An empirical distribution function for sampling with incomplete information, *Annals of Mathematical Statistics*, 26, 641-647.

[9] Babii, A. and R. Kumar (2021) Isotonic regression discontinuity designs, forthcoming in *Journal of Econometrics*.

[10] Balabdaoui, F., Durot, C. and H. Jankowski (2019) Least squares estimation in the monotone single index model, *Bernoulli*, 25, 3276-3310.

[11] Balabdaoui, F., Groeneboom, P. and K. Hendrickx (2019) Score estimation in the monotone single index model, *Scandinavian Journal of Statistics*, 46, 517-544.

[12] Balabdaoui, F. and P. Groeneboom (2021) Profile least squares estimators in the monotone single index model, in *Advances in Contemporary Statistics and Econometrics*, pp. 3-22, Springer.

[13] Barlow, R. and H. Brunk (1972) The Isotonic regression problem and its dual, *Journal of the American Statistical Association*, 67, 140-147.

[14] Bickel, P. J. and Y. A. Ritov (2003) Nonparametric estimators which can be "plugged-in", *Annals of Statistics*, 31, 1033-1053.

[15] Bodory, H., Camponovo, L., Huber, M. and M. Lechner (2016) A wild bootstrap algorithm for propensity score matching estimators, Working paper.

[16] Cattaneo, M. D. and M. H. Farrell (2013) Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators, *Journal of Econometrics,* 174, 127-143.

[17] Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions, *Journal of Econometrics*, 34, 305-334.

[18] Chen, X., Linton, O. and Van Keilegom, I. (2003) Estimation of semiparametric models when the criterion function is not smooth, *Econometrica*, 71, 1591-1608.

[19] Chen, X. and Santos, A. (2018) Overidentification in regular models. *Econometrica*, 86(5), 1771-1817.

[20] Cheng, G. (2009) Semiparametric additive isotonic regression, *Journal of Statistical Planning and Inference*, 139, 1980-1991.

[21] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and J. Robins (2018) Double/debiased machine learning for treatment and structural parameters, *Econometrics Journal*, 21, C1-C68.

[22] Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. and J. M. Robins (2016) Locally robust semiparametric estimation, arXiv:1608.00033.

[23] Cosslett, S. R. (1983) Distribution-free maximum likelihood estimator of the binary choice model, *Econometrica*, 51 765-782.

[24] Cosslett, S. R. (1987) Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models, *Econometrica*, 55, 559-585.

[25] Cosslett, S. R. (2007) Efficient estimation of semiparametric models by smoothed maximum likelihood, *International Economic Review*, 48, 1245-1272.

[26] Dehejia, R. H. and S. Wahba (1999) Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American statistical Association*, 94, 1053-1062.

[27] Durot, C., Kulikov, V. N. and H. P. Lopuhaä (2012) The limit distribution of the L_∞-error of Grenander-type estimators, *Annals of Statistics*, 40, 1578-1608.

[28] Frölich, M. (2004) Finite-sample properties of propensity-score matching and weighting estimators, *Review of Economics and Statistics*, 86, 77-90.

[29] Frölich, M., Huber, M. and M. Wiesenfarth (2017) The finite sample performance of semi-and non-parametric estimators for treatment effects and policy evaluation, *Computational Statistics & Data Analysis*, 115, 91-102.

[30] Goel, P. K. and T. Ramalingam (2012) The matching methodology: some statistical properties, *Springer Science & Business Media,* Vol. 52.

[31] Grenander, U. (1956) On the theory of mortality measurement. II, *Skand. Aktuarietidskr.*, 39, 125-153.

[32] Groeneboom, P. and K. Hendrickx (2017) The nonparametric bootstrap for the current status model, *Electronic Journal of Statistics*, 11, 3446-3484.

[33] Groeneboom, P. and K. Hendrickx (2018) Current status linear regression, *Annals of Statistics*, 46, 1415-1444.

[34] Groeneboom, P. and G. Jongbloed (2014) *Nonparametric Estimation under Shape Constraints*, Cambridge University Press.

[35] Györfi, L., Kohler, M., Krzyzak, A. and H. Walk (2002) A distribution-free theory of nonparametric regression, *Springer Science & Business Media,* Vol. 1.

[36] Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects, *Econometrica*, 66, 315-331.

[37] Heckman, J. J., Ichimura, H. and P. E. Todd (1997) Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme, *Review of Economic Studies*, 64, 605-654.

[38] Heckman, J. J., Ichimura, H. and P. Todd (1998) Matching as an econometric evaluation estimator, *Review of Economic Studies*, 65, 261-294.

[39] Heckman, J., Ichimura, H., Smith, J. and P. Todd (1998) Characterizing selection bias using experimental data, *Econometrica*, 66, 1017-1098.

[40] Hirano, K., Imbens, G. W. and G. Ridder (2000) Efficient estimation of average treatment effects using the estimated propensity score, *NBER Technical Working Paper No. 251.*

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

[42] Huang, J. (2002) A note on estimating a partly linear model under monotonicity constraints, *Journal of Statistical Planning and Inference*, 107, 343-351

[43] Imbens, G. W. (2004) Nonparametric estimation of average treatment effects under exogeneity: a review, *Review of Economics and Statistics,* 86, 4-29.

[44] Khan, S. and E. Tamer (2010) Irregular identification, support conditions, and inverse weight estimation, *Econometrica*, 78, 2021-2042.

[45] Klein, R. W. and R. H. Spady (1993) An efficient semiparametric estimator for binary response models, *Econometrica*, 61 387-421.

[46] Liu, Y. and Qin, J. (2022) Tuning-parameter-free optimal propensity score matching approach for causal inference. arXiv preprint arXiv:2205.13200.

[47] Matzkin R. L. (1992) Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models, *Econometrica,* 60, 239-70.

[48] Meyer, M. C. (2006) Consistency and power in tests with shape-restricted alternatives, *Journal of Statistical Planning and Inferenc*e, 136, 3931-3947.

[49] Newey, W. K. (1990) Semiparametric efficiency bounds, *Journal of Applied Econometrics*, 5, 99-135.

[50] Newey, W. K. (1994) The asymptotic variance of semiparametric estimators, *Econometrica,* 62, 1349-1382.

[51] Otsu, T. and Y. Rai (2017) Bootstrap inference of matching estimators for average treatment effects, *Journal of the American Statistical Association*, 112, 1720-1732.

[52] Qin, J., Yu, T., Li, P., Liu, H. and B. Chen (2019) Using a monotone single-index model to stabilize the propensity score in missing data problems and causal inference, *Statistics in Medicine,* 38, 1442-1458.

[53] Rao, B. P. (1969) Estimation of a unimodal density, *Sankhyā*, A 31, 23-36.

[54] Rao, B. P. (1970) Estimation for distributions with monotone failure rate, *Annals of Mathematical Statistics*, 41, 507-519.

[55] Robins, J. M. and Y. A. Ritov (1997) Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models, *Statistics in Medicine*, 16, 285-319.

[56] Robins, J. and A. Rotnitzky (1995) Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association,* 90, 122-129.

[57] Robinson, P. M. (1988) Root-N-consistent semiparametric regression, *Econometrica*, 56, 931-954.

[58] Rosenbaum, P. R. (1989) Optimal matching for observational studies, *Journal of the American Statistical Association*, 84, 1024-1032.

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

[60] Rosenbaum, P. R. and D. B. Rubin (1984) Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American statistical Association*, 79, 516-524.

[61] Rothe, C. (2017) Robust confidence intervals for average treatment effects under limited overlap, *Econometrica*, 85, 645-660.

[62] Rothe, C. and S. Firpo (2019) Properties of doubly robust estimators when nuisance functions are estimated nonparametrically, *Econometric Theory*, 35, 1048-1087.

[63] Scharfstein, D. O., Rotnitzky, A. and J. M. Robins (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models, *Journal of the American Statistical Association*, 94, 1096-1120.

[64] van de Geer, S. (2000) Empirical Processes in M-Estimation, *Cambridge University Press.*

[65] van der Vaart, A. (1991) On differentiable functionals, *Annals of Statistics*, 19, 178-204.

[66] Wright, F. T. (1981) The asymptotic behavior of monotone regression estimates, *Annals of Statistics*, 9, 443-448.

[67] Xu, M. (2021) Essays in semiparametric estimation and inference with monotonicity constraints, *Doctoral dissertation,* The London School of Economics and Political Science (LSE).

[68] Yu, K. (2014) On partial linear additive isotonic regression, *Journal of the Korean Statistical Society*, 43, 11-17.

[69] Yuan, A., Yin, A. and M. T. Tan (2021) Enhanced doubly robust procedure for causal inference, *Statistics in Biosciences,* 13, 454-478.

Department of Economics, University of Mannheim, L7 3-5, 68161, Mannheim, Germany.
Email address: mengshan.xu@uni-mannheim.de

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.
Email address: t.otsu@lse.ac.uk