Econometric Theory, **0**, 2025, 1–51.
doi:10.1017/S0266466625100133

ISOTONIC PROPENSITY SCORE MATCHING

MENGSHAN XU
University of Mannheim

TAISUKE OTSU
London School of Economics

We propose a one-to-many matching estimator of the average treatment effect based on propensity scores estimated by isotonic regression. This approach is predicated on the assumption of monotonicity in the propensity score function, a condition that can be justified in many economic applications. 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 article 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 the outcome Y, the binary treatment W, the covariate X, and a sample of size N, the matched set for the i-th unit is defined as

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

where $\tilde{p}(\cdot)$ is a uniformly consistent isotonic estimator of the propensity score 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 of the propensity score developed in Section 4.

We are grateful to Markus Frölich, Daniel Gutknecht, Phillip Heiler, Lihua Lei, Yoshi Rai, Christoph Rothe, Carsten Trenkler, and participants at the econometrics seminar at Mannheim 2022, NASMES 2023, and IAAE 2023, for helpful comments and discussions. We also would like to thank a co-editor and anonymous referees for helpful comments to revise the paper. Address correspondence to Taisuke Otsu, Department of Economics, London School of Economics, London, U.K., e-mail: t.otsu@lse.ac.uk.

Remarkably, the isotonic estimator proves to be especially well-suited as the initial nonparametric estimator in a two-stage semiparametric approach to estimating the ATE. It incorporates features of both matching and weighting estimators into the second-stage ATE estimator, addressing at least five issues commonly encountered by existing matching methods in the causal inference literature.

First, it is well known that the existing matching estimators of the ATE with a fixed number of matches are inefficient (Abadie and Imbens, 2006) since they do not balance bias and variance in the second-stage estimation. In comparison, our isotonic matching estimator is more efficient. In the univariate case, our method attains the semiparametric efficiency bound; in the multivariate case, where the efficiency bound becomes more complicated, we show that our proposed estimator performs better than those based on a fixed number of matches with propensity scores derived from widely used parametric models, such as probit and logit, which are prevalent in applied research.

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 have been open questions. However, these issues are addressed by recent papers by Armstrong and Kolesár (2021) and Lin, Ding, and Han (2023). By specifying a large enough Lipschitz constant, Armstrong and Kolesár (2021) showed that the matching estimator with the number of matches set to one is minimax optimal if the conditional mean is restricted to be Lipschitz; by adding an estimated correction term. Lin et al. (2023) gave the optimal number of matches for a bias-corrected matching estimator. In this article, we argue that the isotonic estimator can provide an alternative 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 the need of 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 the (empirical) density function (e.g., Imbens, 2004) of covariates. Surprisingly, the set of the matching counterparts adaptively selected by the 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 more practical 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 does not involve smoothing parameters of conventional nonparametric methods, such as series length or bandwidth. In general, choosing the tuning parameters of a first-stage nonparametric estimator remains a difficult open question in the

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

semiparametric estimation literature. The MSE optimal tuning parameter is usually not a good choice 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). To ensure the $\sqrt{N}-$consistency of a semiparametric estimator, "undersmoothed" tuning parameters should be applied (Newey, 1994). But it is difficult to find a clear standard for shrinking tuning parameters below their MSE optimal values. The non-smooth nature of the isotonic estimator, on the other hand, turns out to automatically render an adequate amount of undersmoothing. 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, such as $\sqrt{N}-$consistency or efficiency.

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, the 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 (compared to parametric methods) 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 been done 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 Bodory et al. (2016) and Adusumilli (2020) to propensity score matching estimators. In our article, 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 the utility functions, which are often monotone increasing in quantities. Furthermore, many functions derived from cumulative distribution functions (CDFs) inherit the monotonicity from the latter. For example, in a threshold-crossing binary choice model,

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

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

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 $\varepsilon \sim \text{Logistic}(0, \frac{\pi^2}{3})$, it becomes a logit model. Although both parametric models are widely applied in estimating the probability of treatments, we can relax the distributional assumptions on ε and express (1) with a semiparametric model $Y = F_\varepsilon(X'\beta_0) + \nu$, where $F_\varepsilon(\cdot)$ is a nonparametric link function. We emphasize that the link function is monotone increasing by construction. See Cosslett (1983, 1987, 2007), Matzkin (1992), and Klein and Spady (1993) for more discussions of the model (1).[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 a similar 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 toward zero or one, resulting in a detrimental effect on the ATE estimator similar to the one caused by limited overlaps (see Khan and Tamer, 2010; Rothe, 2017, among others). 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 the 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; Babii and Kumar, 2023). 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.[2] Second, as discussed in Appendix A.6.2, one of the key conditions for $\sqrt{N}$-consistency and efficient estimation of ATE is (A40) 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 isotonic (hereafter, UC-isotonic) estimator,

[1] Although this article explores monotonicity of the propensity score function, our isotonic regression approach may be extended to the regression-based estimators with monotonicity constraints on the expected outcome functions $\mathbb{E}[Y(1)|X]$ and $\mathbb{E}[Y(0)|X]$. However, it should be noted that if monotonicity is imposed on the link functions of index models, the regression-based approach is clearly more restrictive than the propensity-score-based approach (because monotonicity on $F_\varepsilon(\cdot)$ is not substantive).

[2] 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 $\sqrt{N}$-rates 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.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

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 the 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 et al., 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; Bodory et al., 2016; Adusumilli, 2020, among others). The propensity score matching estimators studied in the literature mainly use 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 a monotonicity constraint. Suppose that the conditional expectation $\mathbb{E}[Y|X] = p_0(X)$ is monotone increasing. Then, for an independent and identically distributed (iid) random sample $\{Y_i, X_i\}_{i=1}^N$, the isotonic estimator is the minimizer of the sum of squared errors, $\min_{p\in\mathcal{M}} \sum_{i=1}^N \{Y_i - p(X_i)\}^2$, 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, 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 does not involve smoothing parameters. On the other hand, the double machine learning (DML) estimators (see, e.g., Robins and Rotnitzky, 1995; Chernozhukov et al., 2017, 2018, among others) provide efficient estimators of the ATE that do not rely on subjective choices of smoothing parameters, thereby, to some extent, sharing many advantages of our approach. We provide a detailed comparison between the isotonic propensity score matching estimator and the DML for the ATE in Section 3.4.

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

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

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) (the last two papers are called BGH hereafter) developed a score-type approach for the monotone single-index model and showed 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 estimators, monotone single-index estimators, or monotone additive estimators, 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, and thus differ from our approach.

In terms of applying isotonic regression to estimate the ATE, the primary difference between this article and Chapter 3 of Xu (2021) is that we address the boundary issue inherent in the isotonic estimator, while Xu (2021) relies on a stronger assumption adapted from Assumption 5.1 in Newey (1994). In the process of writing this article, we have gradually realized that this assumption does not automatically apply to the IPW estimator, although it straightforwardly holds for some other semiparametric models, such as the monotone partially linear model and the monotone single index model, wherein the plugged-in isotonic estimator is not in the denominator. Compared to Chapter 3 of Xu (2021), the main contributions of this article are: (i) proposing a UC-isotonic estimator that is suitable as the first-stage estimator in a propensity score matching estimator of the ATE; (ii) revealing the equivalence between the matching estimator and the IPW estimator when the first-stage propensity score is estimated via UC-isotonic regression; and (iii) based on this equivalence, enriching the literature on propensity score matching by introducing a new approach that addresses several problems of the existing matching methods, as detailed at the beginning of this introduction.[3]

The rest of the article 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. Section 3 compares our approach with existing matching estimators as well as the DML

[3] At almost the same time, an independent work by Liu and Qin (2022) derived a similar equivalence result for the ATE on treated (ATT). Recently, a revised version of Liu and Qin (2022) is published as Liu and Qin (2024). There are two main differences between our article and their papers. First, we formally address the boundary problem of the isotonic estimator and achieve the $\sqrt{N}$-normality of the ATE estimator by proposing a uniformly consistent isotonic estimator. Second, our asymptotic analysis of the model with multivariate covariates in Section 4 focuses on a more general case, where the influence of the estimation errors from the parametric component of the first-stage monotone single-index model is maintained.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

estimator for the ATE. The univariate results are extended to the case of multivariate covariates in Section 4, where the propensity score is modeled by a semiparametric single-index model with an unknown monotone increasing link function. In Section 5, we establish the validity of the nonparametric bootstrap. Monte-Carlo simulation studies are presented in Section 6. All proofs are presented in the Appendix, while additional theoretical details and simulation comparisons are provided in the Supplementary Material.

2. MAIN RESULTS

2.1. Setup and Isotonic Propensity Score

Suppose we observe the triple (Y, W, X) drawn randomly from the product space $\mathcal{Z} = \mathbb{R} \times \{0,1\} \times \mathcal{X}$. Within the triple, $W \in \{0,1\}$ 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 domain $\mathcal{X} = [x_L, x_U] \subset \mathbb{R}$. In this section, we tentatively assume X is scalar, and discuss extensions for multivariate X in Section 4. Without loss of generality, $\{Y_i, W_i, X_i\}_{i=1}^N$ is an iid sample of (Y, W, X) and is ordered by X. If X is continuously distributed, we should have $X_1 < X_2 < \cdots < X_N$ with probability one (i.e., no ties).

In this section, we consider the estimation of the 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

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

where $\mathcal{M}$ is the class of all monotone increasing functions defined on $\mathcal{X}$. Since Brunk (1958), this isotonic regression estimator has been extensively studied in the statistics literature (see, e.g., Barlow et al. (1972) and Groeneboom and Jongbloed (2014) for an overview). One of the well-known features of isotonic regression is that the estimator $\hat{p}(\cdot)$ is a monotone increasing piecewise constant function with jump points at $\{X_{n_k}\}_{k=1}^K$ for some integer K with $1 \le K \le N$. By these jump points, the sample is divided into K disjoint groups, with $\{n_k\}_{k=1}^K$ denoting the first indices of these K groups. Further, we let N_k denote the number of observations belonging to the k-th group. Based on these definitions, it holds that $n_k + N_k = n_{k+1}$ for each $k = 1, \ldots, K-1$, and $\sum_{k=1}^K N_k = N$. Note that the integer K and the corresponding disjoint groups are automatically determined by the isotonic estimation algorithm (see formula (7) below), rather than being chosen by the user.

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

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

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

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 $k+1$). Then, 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 $\{Y_i, W_i, X_i\}_{i=1}^N$ is indexed according to $X_1 < X_2 < \cdots < X_N$.

PROPOSITION 1. *Under Assumption 1, the isotonic estimator $\hat{p}(\cdot)$ satisfying (3) partitions the sample into K disjoint groups in the sense that for each $k = 1, \ldots, K$ 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 groups, and each treated (controlled) observation in group k will be matched with its $N_{k,0}$ ($N_{k,1}$) counterparts, which belong to 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.*

(i) [*Isotonic estimator*] *For each 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 $i \in \{1, \ldots, N\}$ 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 the potential lack of matching counterparts 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 imprecise at the boundary. If we apply the isotonic estimator to the binary dependent variable W,

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

there is a 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 $x \in \mathcal{X}$. For example, if $\hat{p}(X_1) = 0$, (5) implies $N_{1,1} = 0$, i.e., there are no matching counterparts for the treated units.

To fix this problem, we propose a modified isotonic estimator that is uniformly consistent in its domain at a $(\log N)^{1/3} N^{-1/3}$ 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:

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

Our proposed UC-isotonic estimator is obtained by implementing the standard isotonic regression of $\tilde{W}$ on X:

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

A similarly 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 the isotonic estimator itself, while we are interested in the performance of the second-stage matching estimator. To achieve an $N^{-1/2}$ rate at the second stage, we need the isotonic estimator to be uniformly consistent at a rate faster than $N^{-1/4}$, which won't be achieved under Meyer's choice. Meyer (2006) presented a theorem regarding the consistency of the modified estimator at the boundary; however, a proof of consistency was not provided, nor was the rate of convergence discussed.

In this article, 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 $\bar{f}$ and $\underline{f}$, it holds $\underline{f} < f(x) < \bar{f}$ all $x \in \mathcal{X}$.

Assumption 2 (i) is our main assumption, the monotonicity of $p(\cdot)$. Assumption 2 (ii) is required for the $\sqrt{N}-$consistency of the second-stage matching estimator. The same assumption has been adopted by Groeneboom and Hendrickx (2018, Assump. A2) and by BGH (Assumption A3 and its accompanying remark; see also Lemma 22 in the Supplementary Material of BGH) in the context of the monotone single-index model. If we believe that the underlying propensity score function has some flat parts where $p^{(1)}(x) = 0$, we could first run an isotonic estimation of $\tilde{W}_i + c \cdot X_i$ on X_i, where c is a positive constant, to obtain $\tilde{p}_c(x)$. Then,

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

by subtracting the linear trend $c \cdot x$ from $\tilde{p}_c(x)$, we obtain a consistent estimator of $p(x)$.[4] 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 observations 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 $\tilde{p}(\cdot)$ ($N_{k,1}$ is 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 the last group. 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 sizes of the first and the last group, 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 the 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 $0 < \underline{p} \leq p(x) \leq \bar{p} < 1$ 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, the existence of the matching counterparts by $\tilde{p}(\cdot)$ is obtained as follows.

[4]Technically, if $p(\cdot)$ has some flat parts where $p^{(1)}(x) = 0$, then the original estimator $\tilde{p}(\cdot)$ may not satisfy the requirement in (A32) in Appendix A.6, $|\delta(x) - \tilde{\delta}_N(x)| \leq C_0 |p(x) - \tilde{p}(x)|$. Flat parts in $p(\cdot)$ imply that $p(x) - \tilde{p}(x) = 0$ might hold within an entire interval, potentially leading to a violation of (A32). In contrast, if $p(\cdot)$ is strictly monotone increasing, then $p(\cdot)$ and $\tilde{p}(\cdot)$ will cross at most once within each partition given by the isotonic estimator, since $\tilde{p}(\cdot)$ is a piecewise flat function. We refer to Sections 10.2 and 10.3 and Figure 10.1 of Groeneboom and Jongbloed (2014) for more details.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

COROLLARY 1. *Suppose Assumptions 1–3 hold. For each $i = 1, \ldots, N$, the set of its matching counterparts $\{j = 1, \ldots, N : W_j = 1 - W_i, \tilde{p}(x_j) = \tilde{p}(x_i)\}$ 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 $\{Y_i, \tilde{W}_i, X_i\}_{i=1}^N$ using (6).
2. Compute the UC-isotonic estimator $\tilde{p}(\cdot)$ using (7).
3. For each $i = 1, \ldots, N$, compute the matching counterparts

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

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. [Data generating process] (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 $x \in \mathcal{X}$, (iii) for $D(Z) = \frac{WY}{p(X)^2} + \frac{Y(1-W)}{\{1-p(X)\}^2}$, there exist positive constants c_0 and M_0 such that $\mathbb{E}[|D(Z)|^m | X = x] \leq m! M_0^{m-2} c_0$ holds for all integers $m \geq 2$ and every x, and (iv) $Y(1), Y(0) \perp W|X$ almost surely.

Assumption 4 (i)–(iii) regulates the tail behaviors of the (conditional functions of) potential outcomes, which are necessary for $\sqrt{N}$-consistent estimation. Assumption 4 (iv) 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 equal to the corresponding IPW estimator.*

2.3.1. *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 equal to the weighting estimator in each finite sample. This equivalence is tightly 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). See Section 3.1 below for a further comparison of isotonic and partitioning estimators

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

within the context of a two-stage matching estimator of the ATE. Additionally, our method is related to the propensity score methods of blocking, stratification, and radius matching (see Rosenbaum and Rubin, 1983, 1985; Dehejia and Wahba, 1999, 2002, among others). See Section 3.2 for a comparison with these methods.

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.

We notice that the threshold $\lfloor N^{2/3} \rfloor$ in the algorithm (6) can be interpreted as an implicit tuning parameter. We would like to point out that, first, it is convenient to choose since it depends only on the sample size N; second, it is aimed at correcting the boundary problem, which is also faced by other semiparametric and even parametric matching methods. In practice, trimming estimated propensity scores is widely adopted, and the amount of trimming is chosen subjectively in most cases. Our proposed method provides transparent guidance for correcting this common boundary issue. Furthermore, to investigate the impact of different threshold choices, we have included both theoretical analysis and simulation evidence in Sections S1 and S2.3 of the Supplementary Material, respectively.

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 (see, e.g., Hahn, 1998; Hirano, Imbens, and Ridder, 2003). Although we may conduct inference based on an estimator of Ω, we suggest a bootstrap inference method, which will be discussed in Section 5.

3. COMPARISON TO RELATED PROPENSITY SCORE METHODS

In this section, we draw comparisons of our approach with a range of related estimators for the ATE. The comparison with matching methods based on propensity score estimated by partitioning estimator is presented in Section 3.1, the comparison with propensity score methods of blocking, stratification, and radius matching is presented in Section 3.2, the comparison with matching methods based on propensity score estimated by regression trees is presented in Section 3.3, and the comparison with the DML estimator for the ATE can be found in Section 3.4.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

3.1. Propensity Score Estimated by Partitioning Estimator

One notable feature of the proposed isotonic propensity score matching method is that it is a one-to-many matching method that provides exact matches, as illustrated by formula (8). This is attributed to the isotonic estimator being considered a special type of partitioning estimator, in which the volume sizes of partitions are automatically chosen by the monotonicity constraint, and a simple average is implemented within each partition.

The partitioning estimator is a nonparametric method for estimating regression functions.[5] It divides the domain of the running variables into disjoint partitions. Within each partition, a local estimator is implemented by the user, such as the sample mean, a linear estimator, or a series estimator. Each sample point is exclusively used in the estimation within the partition to which it belongs. This feature simplifies the complex correlation structure of a matching estimator such that it achieves equivalence with an IPW estimator. For the UC-isotonic estimator, this equivalence is presented by equation (A24) in Appendix A.5. In the resulting matching estimator of ATE, the same set of partitions serves both the first- and the second-stage nonparametric estimation. Usually, these two stages are not associated with each other since they have distinct objects, the propensity score and the potential outcomes. Certainly, a matching estimator of the ATE that utilizes propensity scores estimated with a partitioning estimator should exhibit a similar equivalence to the weighting estimator. However, the selection of the number of partitions and their sizes necessitates careful consideration, as they must meet specific undersmoothing conditions to secure the desired asymptotic properties of the second-stage ATE estimator. The challenge of selecting an appropriate undersmoothed bandwidth or volume size, as mentioned in the introduction, remains a difficult open question in the semiparametric estimation. In contrast, our proposed isotonic matching estimator automatically chooses these tuning parameters, leading to the efficient estimation of the ATE, as demonstrated in Theorem 3.

3.2. Propensity Score Methods of Blocking, Stratification, and Radius Matching

Our proposed isotonic matching estimator is also related to some of the seminal ideas introduced at the outset of the propensity score methods: blocking, stratification (Rosenbaum and Rubin, 1983; Dehejia and Wahba, 1999, 2002), and radius matching (Rosenbaum and Rubin, 1985).

The isotonic propensity score matching shares similarities with blocking and stratification matching on propensity scores, notably: (i) they initially categorize data points into distinct groups (or strata, blocks, and partitions) according to estimated propensity scores and (ii) within each group, they calculate the conditional

[5] We refer to Györfi et al. (2002) and Cattaneo and Farrell (2013) for comprehensive discussions of the partitioning estimator.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

ATE as the simple difference in means of outcomes between the treatment and comparison groups. The primary distinction lies in the grouping mechanism: for isotonic propensity score matching, the groups are determined adaptively in a data-driven manner through isotonic regression, whereas for the stratification estimator of the ATE, the strata must be explicitly specified by the user. Another distinction is that for the isotonic propensity score matching method, the same set of partitions is utilized for both the first and second stages of nonparametric estimation. As presented by Theorem 2, this characteristic leads to the equivalence between the matching and IPW estimator, resulting in the efficient estimation of the ATE. In contrast, in the case of blocking or stratification matching methods, particularly when the propensity score is estimated using parametric models, this equivalence cannot generally be established, and efficiency cannot be assured without implementing some bias-correction method.

The case for the radius matching estimator is similar to the stratified matching estimator. The difference is that for stratified matching, each unit is matched solely with units from the opposite treatment group within the same stratum, while radius matching allows each unit to be matched to several local balls, the centers of which belong to the opposite treatment group. For both radius and stratified matching estimators, the sizes of strata or the radii act as tuning parameters, which must be chosen by the users when the propensity score is estimated via parametric or nonparametric methods dependent on smoothing parameters (such as kernel or series estimation). These smoothing parameters play a key role in balancing the bias and variance, thereby significantly affecting the second-stage estimator of ATE. In contrast, isotonic regression distinguishes itself by automatically generating these partitions through the application of the monotonicity constraint.

3.3. Propensity Score Estimated by Regression Trees

As methods of estimating the propensity scores, the isotonic estimator and regression trees share several similarities. First, both are nonparametric estimators that do not impose restrictive parametric structures on the underlying response function. Second, both approaches partition the domain of running variables (the feature space in regression tree terminology) into several regions and use the sample average within each region as estimators. As a result, both estimators take the form of piecewise-constant functions. Third, both methods form their piecewise-constant functions in data-adaptive manners. In particular, the partitions created by both methods depend on the dependent variable (the response), which differentiates them from regular nonparametric methods, such as the kernel estimator.

On the other hand, there are notable distinctions between the two methods. First, both approaches construct their piecewise-constant functions differently: the partitions in a regression tree are obtained in a stepwise manner. In each step, a partition is chosen to achieve the maximum marginal reduction of the mean square error (MSE), without imposing any shape constraints during this process. In contrast, the isotonic estimator employs a one-step approach that

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

determines partitions to minimize the MSE over the class of monotone functions. Second, although both approaches are data-driven, the isotonic estimator is free of smoothing parameters, whereas the regression tree depends on the user to specify the tree's length. (When the tree length is determined by cross-validation, the user must select the penalty parameter.) Third, the regression tree is inherently designed for multi-dimensional problems, whereas the canonical form of isotonic regression addresses one-dimensional issues, given that the traditional definition of monotonicity characterizes the relationship between two variables. Nevertheless, the isotonic estimation can be extended to multivariate cases by being incorporated into a partially linear model or a monotone single-index model. The latter is illustrated in Section 4 below.

To summarize, the isotonic estimator necessitates the monotonicity assumption in the underlying response function, a requirement not shared by regression trees. This assumption, however, enables the isotonic estimation algorithm to automatically regulate the trade-off between bias and variance. Conversely, when using regression trees, practitioners are tasked with the challenge of selecting an appropriate tree length to effectively manage the balance between bias and the risk of overfitting. The strength of regression trees is their natural aptitude for tackling multivariate problems. When employing regression trees in the preliminary stage of propensity score estimation as part of a two-stage approach to estimating the ATE, it is commonly combined with methods for bias correction and sample splitting, as discussed by Chernozhukov et al. (2018). See Section 3.4 below for more details about the comparison of our approach with the DML estimator.

3.4. DML Estimator

The isotonic propensity score matching estimator and the DML estimator for the ATE both share the benefit of not requiring subjective choices of tuning parameters. For estimating the ATE, a typical example of a DML estimator is given by applying the sample splitting to the AIPW estimator. In the following, we abstract from sample splitting to simplify notation:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\{\hat{\psi}_1(X_i)-\hat{\psi}_0(X_i)\}+\frac{1}{N}\sum_{i=1}^{N}\left[\frac{W_i(Y_i-\hat{\psi}_1(X_i))}{\hat{p}(X_i)}-\frac{(1-W_i)(Y_i-\hat{\psi}_0(X_i))}{1-\hat{p}(X_i)}\right]\\
&\quad=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{Y_iW_i}{\hat{p}(X_i)}-\frac{Y_i(1-W_i)}{1-\hat{p}(X_i)}\right]\\
&\qquad-\frac{1}{N}\sum_{i=1}^{N}\left[\frac{W_i-\hat{p}(X_i)}{\hat{p}(X_i)}\hat{\psi}_1(X_i)-\frac{W_i-\hat{p}(X_i)}{1-\hat{p}(X_i)}\hat{\psi}_0(X_i)\right], \qquad \textbf{(10)}
\end{aligned}$$

where $\hat{\psi}_1(\cdot)$ and $\hat{\psi}_0(\cdot)$ are estimators of $\mathbb{E}[Y(1)|X=\cdot]$ and $\mathbb{E}[Y(0)|X=\cdot]$, respectively. The first and second lines of (10) present two formulations of the

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

DML estimator for the ATE. The first terms in both lines correspond to the standard regression and IPW estimators, respectively, while the subsequent terms represent their bias-correction components.

The AIPW has been extensively studied since the seminal work of Robins, Rotnitzky, and Zhao (1995) and Robins and Rotnitzky (1995) (see also Newey, Hsieh, and Robins, 1998, 2004; Scharfstein, Rotnitzky, and Robins, 1999; Rothe and Firpo, 2019, among others). In an influential work, Chernozhukov et al. (2018) combined orthogonal moment functions – of which formula (10) is a specific case for the ATE – with sample splitting, accommodating a broad array of the first-stage machine learners that are prone to bias due to regularization or model selection. Recent developments by Chernozhukov et al. (2022) and Chernozhukov, Newey, and Singh (2022) have proposed methods for constructing the correction term without requiring an explicit function form for the bias correction.

Both estimators have their own advantages and comparative strengths. From a practical standpoint, the isotonic propensity score matching method stands out for its simplicity and ease of implementation: it does not require the correction terms, thereby sparing the effort of estimating the conditional means of potential outcomes and sidesteps the challenges associated with their correct specification. In contrast, the DML estimator's efficiency relies on correctly specifying and effectively estimating both the propensity score and the conditional means of potential outcomes. A misstep in either leads to a consistent yet inefficient estimator. On the other hand, the DML estimator exhibits great flexibility: through the use of sample splitting, it supports a variety of first-stage estimators, accommodating high-dimensional data or highly complex function classes, such as random forest, neural networks, and other advanced machine learning technologies.

From a technical standpoint, the isotonic propensity score matching and the DML for the ATE represent two distinct pathways of semiparametric estimation: undersmoothing and bias correction. Both strategies aim for $\sqrt{N}$-consistent (or efficient in certain cases) estimators (see Newey (1994) for a relevant discussion). The undersmoothing strategy depends on a first-stage estimator with reduced bias, achievable in nonparametric estimators by selecting smoothing parameters smaller than the MSE-optimal levels. Conversely, the bias-correction method addresses bias by incorporating an estimated correction term into the second-stage sample moment function, rather than concentrating on the first stage.

The proposed isotonic propensity score matching estimator utilizes the isotonic estimator, which achieves a similar effect of "undersmoothing," and this effect is automatically rendered by enforcing monotonicity. The isotonic estimator does not really shrink its bias to a level lower than $N^{-1/2}$. However, when combined with the monotonicity, it eventually achieves a deviation from the efficient influence function that decays at a rate faster than $N^{-1/2}$ (see (A40) in the Appendix). In contrast, the DML for the ATE represents a typical bias-correction approach. The second terms in both lines of (10), while achieving the "doubly robust" effect, also serve as bias-correction components. At the cost of computing additional

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

correction terms and some efficiency loss due to sample splitting, the DML approach manages to mitigate potential bias and prevent overfitting risks, while being less restrictive on the first-stage estimation. It is not only less sensitive to the choice of the smoothing parameter for the traditional first-stage nonparametric estimator but can also accommodate many black-box machine learning methods, whose asymptotic properties remain to be fully understood. Consequently, the theoretical development of the isotonic propensity score matching and the DML for the ATE differs substantially. The DML approach significantly reduces the effort needed to address issues arising from the complexity of function classes, which is associated either with the correlation brought by plug-in estimators or with the choice of smoothing parameter. In contrast, this article needs to address the impact of the plug-in estimator in the theoretical development of the isotonic propensity score matching estimator.

Finally, we would like to emphasize that our proposed method represents a targeted advancement within the matching estimation literature, specifically addressing several limitations present in existing matching techniques for estimating the ATE. In contrast, the DML is a versatile tool designed for broader semiparametric estimation tasks, which include a wide array of econometric problems, such as average derivatives, partially linear models, and parameters of economic structural models. Our approach, therefore, complements rather than competes with the expansive toolkit that DML offers, by providing subtle yet significant improvements in the specialized area of matching estimation.

4. 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 with the monotone single-index model:

$$W = p_0(X'\alpha_0) + \varepsilon, \qquad \mathbb{E}[\varepsilon|X] = 0, \tag{11}$$

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 $||\alpha_0|| = 1$.[6]

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), and Klein and Spady (1993), among others. In the case where $p_0(\cdot)$ is estimated with isotonic regression, Balabdaoui et al. (2019) studied (11) with the monotone least square method, and Groeneboom and Hendrickx (2018), Balabdaoui et al. (2019), and Balabdaoui and Groeneboom (2021) (BGH) estimated α_0 and $p_0(\cdot)$ by solving a score-type sample moment

[6] 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.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

condition of

$$\mathbb{E}[X\{W - p_0(X'\alpha_0)\}] = 0. \tag{12}$$

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

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

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

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha} \left\| \frac{1}{N}\sum_{i=1}^{N} X_i'\{W_i - \hat{p}_\alpha(X_i'\alpha)\} \right\|^2. \tag{14}$$

BGH showed that under certain assumptions, $\hat{\alpha}$ is a $\sqrt{N}$-consistent estimator for α_0,[7] 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 the 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, which is denoted by $\tilde{p}_{\tilde{\alpha}}$. The matching procedure can be implemented as follows.

1. Compute $\hat{\alpha}$ by (13) and (14).
2. Define $\tilde{\alpha} = \hat{\alpha}$, 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. 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

$$\mathcal{J}(i) = \left\{j = 1, \ldots, N : W_j = 1 - W_i \text{ and } \tilde{p}_{\tilde{\alpha}}(X_j'\tilde{\alpha}) = \tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha})\right\}. \tag{15}$$

[7] BGH proposed solving a "zero-crossing" root of $\frac{1}{N}\sum_{i=1}^N X\{W_i - \hat{p}_\alpha(X_i'\alpha)\} = 0$. Then, they realized that there was an issue with the existence of the zero-crossing root for a finite sample (due to the discreteness of $\hat{p}_\alpha$). To fix this problem, Balabdaoui and Groeneboom (2021) replaced this objective function with (14), where a minimizer always exists. If there are multiple minimizers, any of them is a $\sqrt{N}$-consistent estimator for α_0. (See a discussion on p. 1426 of Groeneboom and Hendrickx (2018).) BGH also proposed an efficient estimator of α_0 by solving a kernel-adjusted score function. Since our aim is the second-stage ATE τ instead of the first-stage propensity score p, we do not apply BGH's efficient estimator. It will introduce additional tuning parameters without improving the second-stage ATE.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

5. Calculate the matching estimator for the ATE τ by

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(Y_i - \frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j\right), \tag{16}$$

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 a nonempty interior. There exists $R > 0$ such that $\mathcal{X} \subset \mathcal{B}(0,R) = \{x : ||x|| \le R\}$.

Given α, we define the true link function of (13):

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

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 $\underline{f}$ and $\bar{f}$, it holds $\underline{f} < f(x) < \bar{f}$ all $x \in \mathcal{X}$.

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

Assumption 4' [Data-generating process]. (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 $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$, (iii) for $D(Z) = \frac{WY}{p_0(X'\alpha_0)^2} + \frac{Y(1-W)}{\{1-p_0(X'\alpha_0)\}^2}$, there exist positive constants c_0 and M_0 such that $\mathbb{E}[|D(Z)|^m|X = x] \le m! M_0^{m-2} c_0$ holds for all integers $m \ge 2$ and every x, and (iv) $Y(1), Y(0) \perp W|X$ almost surely.

Let Z denote the triple (Y, W, X), and $\mathcal{Z}$ denote the space of the random vector Z. For each $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$, $u \in I_\alpha = \{x'\alpha : x \in \mathcal{X}\}$, and a function $f(\cdot)$ defined on $\mathcal{Z}$, we define $\mathbb{E}_\alpha[f(Z)|u] = \mathbb{E}[f(Z)|X'\alpha = u]$. Similarly, we define the conditional covariance $\mathrm{Cov}_{\alpha_0}(f(Z), X|u)$. The following two assumptions are adapted from BGH, which ensure that the score estimators (13) and (14) have desirable properties.

Assumption 5. For all $\alpha \neq \alpha_0$ such that $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$, the random variable $\mathrm{Cov}[(\alpha - \alpha_0)'X, p_0(X'\alpha_0)|X'\alpha]$ 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 $\mathbb{E}[p_0^{(1)}(X'\alpha_0)\mathrm{Cov}(X|X'\alpha_0)]$ has rank $k-1$.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

Based on Assumptions 1', 2', 5, and 6, we have a result similar to Proposition 3, but the numbering is according to $X_1'\tilde{\alpha} < \cdots < X_N'\tilde{\alpha}$. 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 $\mathbf{B}^-$ denote the Moore–Penrose inverse of a square matrix $\mathbf{B}$. The asymptotic properties of the isotonic propensity score matching estimator are 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*

$$\begin{aligned} m(Z) &= \frac{YW}{p_0(X'\alpha_0)} - \frac{Y(1-W)}{1-p_0(X'\alpha_0)} - \tau, \qquad D(Z) = \frac{YW}{p_0(X'\alpha_0)^2} + \frac{Y(1-W)}{(1-p_0(X'\alpha_0))^2}, \\ M(Z) &= -\mathbb{E}_{\alpha_0}[D(Z)|X'\alpha_0]\{W - p_0(X'\alpha_0)\}, \\ A(Z) &= -\mathbb{E}[\mathrm{Cov}_{\alpha_0}(D(Z), X|X'\alpha_0)p_0^{(1)}(X'\alpha_0)], \\ &\quad \times \mathbb{E}[p_0^{(1)}(X'\alpha_0)\mathrm{Cov}_{\alpha_0}(X|X'\alpha_0)]^-\{X - \mathbb{E}_{\alpha_0}[X|X'\alpha_0]\}\{W - p_0(X'\alpha_0)\}. \end{aligned} \tag{17}$$

Note that the semiparametric efficiency bound for estimating τ with known α_0 is given by $\mathbb{E}[\{m(Z) + M(Z)\}\{m(Z) + M(Z)\}']$ (see, e.g., Newey, 1994). The additional term $A(Z)$ can be interpreted as the influence of estimating the index coefficients α_0. This influence is also faced by parametric matching estimators. In general, our proposed method uses the matched sets, in which the number of matches increases to infinite, so it better balances the variance and bias in the second stage and should asymptotically outperform any matching method with fixed numbers of matches. In Section 6.2 below, we present simulation results to illustrate that the proposed ATE estimator $\tilde{\tau}$ outperforms the probit matching estimator in every sample size, even in the case that the true propensity score is a probit (the correct specification).

Theoretically, the additional term $A(Z)$ can be avoided by using a semiparametric weighting estimator. However, the costs are strong assumptions on the smoothness of the propensity scores (typically, $7 \cdot \dim(X)$-th continuous differentiability; see Hirano et al., 2003) and a proper choice of smoothing parameters. Our proposed method only requires the propensity score to be once continuously differentiable, and it does not involve smoothing parameters, such as bandwidths or series lengths.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

5. 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 have to choose some smoothing parameters even though the point estimators are free from smoothing. To avoid the estimation of such nonparametric components, 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 Bodory et al. (2016) and Adusumilli (2020) 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 bootstrap sample 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

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

   where $M_i^* = |\mathcal{J}^*(i)|$ is the number of matches for the i-th observation in the bootstrap sample.
4. After repeating Steps (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. *Let $\mathbb{P}^*$ be the bootstrap distribution conditional on the data, and $c_{1-\alpha}^*$ be the $(1-\alpha)$-th sample quantile of $(\sqrt{N}(\hat{\tau}_1^* - \hat{\tau}), \sqrt{N}(\hat{\tau}_2^* - \hat{\tau}), \ldots, \sqrt{N}(\hat{\tau}_B^* - \hat{\tau}))$. Under Assumptions 1–4, it holds*

(i) $\sup_{t\in\mathbb{R}} |\mathbb{P}^*\{\sqrt{N}(\hat{\tau}^* - \hat{\tau}) \leq t\} - \mathbb{P}\{\sqrt{N}(\hat{\tau} - \tau) \leq t\}| \xrightarrow{p} 0$;
(ii) $\mathbb{P}\{\sqrt{N}(\hat{\tau} - \tau) \leq c_{1-\alpha}^*\} \xrightarrow{p} 1 - \alpha$.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

6. MONTE-CARLO SIMULATIONS

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

6.1. Univariate Case

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

$$W = \begin{cases} 0 & \text{if } X < \nu \\ 1 & \text{if } X \geq \nu, \end{cases}$$
$$Y = 0.5W + 2X + \varepsilon,$$
$$\varepsilon \sim N(0, 1). \tag{18}$$

The true ATE is the coefficient of W, which is 0.5. The simulation results are presented in Table 1, where $\hat{\mu}_\tau$ is the Monte-Carlo mean, and the MSEs are rescaled by N. The number of Monte-Carlo simulations is 5,000 for each sample size.

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):

$$\begin{aligned} \Omega_{\text{SEB}} &= \text{Var}(\mathbb{E}[Y(1) - Y(0)|X]) + \mathbb{E}[\text{Var}(Y(1)|X)/p_0(X)] \\ &\quad + \mathbb{E}[\text{Var}(Y(0)|X)/(1 - p_0(X))] \\ &= \text{Var}(0.5) + \mathbb{E}[1/p_0(X)] + \mathbb{E}[1/(1 - p_0(X))] \\ &= 0 + \int_{0.15}^{0.85} \frac{1}{x} \frac{1}{0.7} dx + \int_{0.15}^{0.85} \frac{1}{1-x} \frac{1}{0.7} dx \approx 4.96. \end{aligned}$$

TABLE 1. Matching estimators of ATE: the 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
1,000	0.4934	5.2589	1,000	0.5009	7.0630
2,000	0.4946	5.2158	2,000	0.4999	7.0816
5,000	0.4963	4.9418	5,000	0.4995	6.8376
10,000	0.4974	4.9785	10,000	0.5000	6.8238
∞	0.5	4.96	∞	0.5	4.96

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

In comparison, the logit matching estimator has a slightly smaller bias, and it seems that both estimators are asymptotically unbiased. 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.

6.2. Multivariate Case

Consider the following setting:

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

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 5,000 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 the one-to-one matching estimator based on propensity scores estimated with the correctly specified probit model.

The pattern is similar to the univariate case. The biases of both estimators are small and converge to zero. The isotonic matching estimator outperforms the probit matching estimator in every sample size in terms of MSE.

TABLE 2. Matching estimators of ATE: The 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
1,000	0.5016	5.0014	1,000	0.5030	6.9813
2,000	0.4991	5.0727	2,000	0.4997	7.2275
5,000	0.5003	5.2115	5,000	0.5010	7.2640
10,000	0.5001	5.0161	10,000	0.5002	7.0509

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

Table 3. Bootstrap coverage rates

n	90% CI	95% CI
100	0.860	0.918
1,000	0.889	0.938
2,000	0.881	0.940
5,000	0.901	0.945
10,000	0.891	0.948
∞	0.90	0.95

6.3. Bootstrap

Table 3 shows the bootstrap coverage rates. We draw 2,000 Monte-Carlo simulations, and for each simulation, we draw 500 bootstrap samples. The coverage rates are calculated with these 2,000 sets of confidence intervals for both 90% and 95% confidence levels. From Table 3, we see clear trends that the bootstrap coverage rates are converging to their theoretical limits.

Overall, the simulation outcomes of the univariate case, the multivariate case, and the bootstrap encourage the proposed isotonic propensity score matching method. Additionally, for further simulation comparisons of our approach with propensity score methods of one-to-many matching and radius matching, as well as the impact of thresholds for averaging treatment variables at boundaries, see Section S2 of the Supplementary Material.

7. 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 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 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.

A. PROOFS

A.1. Proof of Proposition 1

The proof is based on the following lemma.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

LEMMA A1 (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*

$$\sum_{j=1}^{i}\hat{p}_j \begin{cases} \le \sum_{j=1}^{i} W_j \\ = \sum_{j=1}^{i} W_j \quad \text{if } \hat{p}_{i+1} > \hat{p}_i \text{ or } i = N. \end{cases} \tag{A1}$$

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 (A1), we have

$$\sum_{j=1}^{n_k-1}\hat{p}_j = \sum_{j=1}^{n_k-1} W_j, \qquad \sum_{j=1}^{n_{k+1}-1}\hat{p}_j = \sum_{j=1}^{n_{k+1}-1} W_j, \qquad \hat{p}_{n_k} = \hat{p}_{n_k+1} = \cdots = \hat{p}_{n_k+N_k-1}. \tag{A2}$$

Since $n_k - 1 = n_{k-1} + N_{k-1} - 1$ and $n_{k+1} - 1 = n_k + N_k - 1$, (A2) implies

$$\sum_{j=n_k}^{n_k+N_k-1}\hat{p}_j = \sum_{j=n_k}^{n_k+N_k-1} W_j. \tag{A3}$$

Combining (A2) and (A3), 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

A.3.1. *Proof of (i).* Since the proof is similar, we focus on the proof of the first statement, $N_1 \ge \lfloor N^{2/3} \rfloor$. The isotonic estimator can be written as (see, Barlow and Brunk, 1972):

$$\hat{p}(X_i) = \max_{s \le i}\min_{t \ge i}\sum_{j=s}^{t}\frac{W_j}{t-s+1}. \tag{A4}$$

Let

$$\bar{W}_l = \frac{1}{\lfloor N^{2/3} \rfloor}\sum_{i=1}^{\lfloor N^{2/3} \rfloor} W_i, \qquad \bar{W}_u = \frac{1}{\lfloor N^{2/3} \rfloor}\sum_{i=N-\lfloor N^{2/3} \rfloor+1}^{N} W_i. \tag{A5}$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

For any i with $1 \le i \le \lfloor N^{2/3} \rfloor$, (6), (7), and (A4) imply

$$\begin{aligned}\tilde{p}(X_i) &= \max_{s\le i}\min_{t\ge i}\sum_{j=s}^{t}\frac{\tilde{W}_j}{t-s+1}\\ &= \max_{s\le i}\min_{t\ge i}\frac{\sum_{j=s}^{t\wedge\lfloor N^{2/3}\rfloor}\bar{W}_l+\mathbb{I}\{t>\lfloor N^{2/3}\rfloor\}\sum_{j=\lfloor N^{2/3}\rfloor+1}^{t}W_j}{t-s+1}\\ &= \max_{s\le i}\min_{t\ge i}\frac{\sum_{j=s}^{t}\bar{W}_l+\mathbb{I}\{t>\lfloor N^{2/3}\rfloor\}\left(\sum_{j=\lfloor N^{2/3}\rfloor+1}^{t}W_j-\sum_{j=\lfloor N^{2/3}\rfloor+1}^{t}\bar{W}_l\right)}{t-s+1}\\ &= \bar{W}_l+\max_{s\le i}\min_{t\ge i}\left[\mathbb{I}\{t>\lfloor N^{2/3}\rfloor\}\frac{t-\lfloor N^{2/3}\rfloor}{t-s+1}\left(\frac{1}{t-\lfloor N^{2/3}\rfloor}\sum_{j=\lfloor N^{2/3}\rfloor+1}^{t}W_j-\bar{W}_l\right)\right].\end{aligned} \tag{A6}$$

Since $\mathbb{I}\{t>\lfloor N^{2/3}\rfloor\}\frac{t-\lfloor N^{2/3}\rfloor}{t-s+1}\ge 0$, 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 discuss 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) $\min_{t>\lfloor N^{2/3}\rfloor}\frac{1}{t-\lfloor N^{2/3}\rfloor}\sum_{j=\lfloor N^{2/3}\rfloor+1}^{t}W_j\le\bar{W}_l$.

For Case (I), adding any terms after $\lfloor N^{2/3}\rfloor$ cannot make the average smaller. Thus, we have $\tilde{p}(X_i)=\bar{W}_l$ for all $1\le i\le\lfloor N^{2/3}\rfloor$, and 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 items after $\lfloor N^{2/3}\rfloor$ will lower the overall level of the sample mean (A6). Define

$$t_s=\arg\min_{t\ge\lfloor N^{2/3}\rfloor}\mathbb{I}\{t>\lfloor N^{2/3}\rfloor\}\frac{t-\lfloor N^{2/3}\rfloor}{t-s+1}\left(\frac{1}{t-\lfloor N^{2/3}\rfloor}\sum_{j=\lfloor N^{2/3}\rfloor+1}^{t}W_j-\bar{W}_l\right). \tag{A7}$$

After minimizers are chosen for each s, the maxmin operator requires to choose the maximum across different s. For any i smaller than $\lfloor N^{2/3}\rfloor$ and any $j\le i$, we have $\tilde{W}_j=\bar{W}_l\ge\min_{t>\lfloor N^{2/3}\rfloor}\frac{1}{t-\lfloor N^{2/3}\rfloor}\sum_{m=\lfloor N^{2/3}\rfloor+1}^{t}W_m$. Therefore, adding more terms before i will increase the overall level of the sample mean (A6), so we must have $s=1$. (This is also justified by (A7): for $s<t$, the smaller s, the greater $-\frac{t-\lfloor N^{2/3}\rfloor}{t-s+1}$. Note that $\min_{t>\lfloor N^{2/3}\rfloor}\frac{1}{t-\lfloor N^{2/3}\rfloor}\sum_{j=\lfloor N^{2/3}\rfloor+1}^{t}W_j-\bar{W}_l\le 0$ by the setup of Case (II).)

Consequently, (A6) can be written as

$$\tilde{p}(X_i)=\bar{W}_l+\frac{t_1-\lfloor N^{2/3}\rfloor}{t_1}\left(\frac{1}{t_1-\lfloor N^{2/3}\rfloor}\sum_{j=\lfloor N^{2/3}\rfloor+1}^{t_1}W_j-\bar{W}_l\right)=\sum_{j=1}^{t_1}\frac{\tilde{W}_j}{t_1}, \tag{A8}$$

with $t_1>\lfloor N^{2/3}\rfloor$. (A8) gives a common value of $\tilde{p}(X_i)$ for all $i=1,\dots,\lfloor N^{2/3}\rfloor$. By (3), we have $N_1=t_1>\lfloor N^{2/3}\rfloor$, which implies the conclusion.

A.3.2. *Proof of (ii).* Part (i) shows that all the changed treatment variables are clustered in the first and the last group. Therefore, for $k=2,3,\dots,K-1$, Part (ii) holds

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

by the same arguments for Propositions 1 and 2 (i), so it remains to show the cases for for $k = 1$ and K. Since the proof is similar, we only present the proof for $k = 1$.

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

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

A.3.3. *Proof of (iii).* Since the proofs are similar, we focus on the first statement, $N_1 = O_p(N^{2/3})$. The idea of this proof is based on the intuition that under Assumption 2 (ii), the treatment propensity should be higher after $\lfloor N^{2/3} \rfloor$ than before this point. As a result, it becomes increasingly unlikely for the points to the right of $\lfloor N^{2/3} \rfloor$ to be allocated to the first partition.

By definition, it is equivalent to show that for any $\nu > 0$, there exists $c > 0$ such that $\mathbb{P}(c_1 > c) < \nu$, where

$$c_1 = \frac{N_1}{N^{2/3}}. \tag{A9}$$

Without loss of generality, we choose N such that $N^{2/3} = \lfloor N^{2/3} \rfloor$; and we can set c to be a positive integer and $c > 2$; note that $N_1 = c_1 N^{2/3}$, and we have

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

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

$$= \mathbb{P}\Bigg(\frac{1}{N^{1/3}} \sum_{i=1}^{N^{2/3}} \{W_i - p(X_i)\} - \frac{N^{1/3}}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} \{W_j - p(X_j)\}$$

$$> \frac{N^{1/3}}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} p(X_j) - \frac{1}{N^{1/3}} \sum_{i=1}^{N^{2/3}} p(X_i),\ c_1 > c \Bigg)$$

$$= \mathbb{P}\left(\sum_{i=1}^{c_1 N^{2/3}} B_i > a,\ c_1 > c \right), \tag{A10}$$

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 (A5) and $N^{2/3} = \lfloor N^{2/3} \rfloor$, the third equality follows 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 is given by the definitions:

$$B_i = \begin{cases} \frac{W_i - p(X_i)}{N^{1/3}} & \text{for } 1 \le i \le N^{2/3} \\ -\frac{N^{1/3}\{W_i - p(X_i)\}}{c_1 N^{2/3} - N^{2/3}} & \text{for } \lfloor N^{2/3} \rfloor + 1 \le i \le c_1 N^{2/3}, \end{cases}$$

$$a = \frac{N^{1/3}}{c_1 N^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{c_1 N^{2/3}} p(X_j) - \frac{1}{N^{1/3}} \sum_{i=1}^{N^{2/3}} p(X_i). \tag{A11}$$

Note that $\sum_{i=1}^{c_1 N^{2/3}} B_i$ is an average centered around zero; the term a, due to Assumption 2 (ii), should be strictly positive. We now apply the following Bernstein inequality (see, e.g., van de Geer, 2000) to (A10).

Bernstein Inequality: *Let $B_1, \ldots, B_n$ be independent random variables satisfying*

$$\mathbb{E}[B_i] = 0, \qquad \mathbb{E}[|B_i|^m] \le \frac{m!}{2} A^{m-2} \mathbb{V}(B_i) \quad \text{for each } m = 2, 3, \ldots,$$

$$b^2 = \sum_{i=1}^{n} \mathbb{V}(B_i), \tag{A12}$$

for some constant A. Then,

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

Let q_α denote the α-th quantile of X. For any positive integer $c > 2$,

$$\frac{N^{1/3}}{cN^{2/3} - N^{2/3}} \sum_{j=N^{2/3}+1}^{cN^{2/3}} p(X_j) - \frac{1}{N^{1/3}} \sum_{i=1}^{N^{2/3}} p(X_i)$$

$$= N^{1/3} \left(\frac{\int_{q_{N^{-1/3}}}^{q_{c \cdot N^{-1/3}}} p(x) f(x) dx}{\int_{q_{N^{-1/3}}}^{q_{c \cdot N^{-1/3}}} f(x) dx} - \frac{\int_{x_L}^{q_{N^{-1/3}}} p(x) f(x) dx}{\int_{x_L}^{q_{N^{-1/3}}} f(x) dx} + O_p((N^{2/3})^{-1/2}) \right)$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

$$
\begin{aligned}
&= N^{1/3}\left(\frac{\int_{q_{N^{-1/3}}}^{q_{c\cdot N^{-1/3}}} \{p(x_L)+p^{(1)}(x_L)\cdot(x-x_L)+o(x-x_L)\}f(x)dx}{\int_{q_{N^{-1/3}}}^{q_{c\cdot N^{-1/3}}} f(x)dx} \right. \\
&\quad \left. - \frac{\int_{x_L}^{q_{c\cdot N^{-1/3}}} \{p(x_L)+p^{(1)}(x_L)\cdot(x-x_L)+o(x-x_L)\}f(x)dx}{\int_{x_L}^{q_{N^{-1/3}}} f(x)dx} + O_p(N^{-1/3}) \right) \\
&\geq N^{1/3}\Big[p(x_L)+p^{(1)}(x_L)\underline{f}(q_{c\cdot N^{-1/3}}-x_L) \\
&\quad - \{p(x_L)+p^{(1)}(x_L)p^{(1)}(x_L)\bar{f}(q_{N^{-1/3}}-x_L)+o(N^{-1/3})\}\Big] + O_p(1) \\
&= p^{(1)}(x_L)N^{1/3}\{\underline{f}_{q_{c\cdot N^{-1/3}}}(q_{c\cdot N^{-1/3}}-x_L)-\bar{f}(q_{N^{-1/3}}-x_L)\}+O_p(1) \\
&=: a_c + O_p(1), \qquad \textbf{(A13)}
\end{aligned}
$$

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 is given by the definition

$$a_c = p^{(1)}(x_L)N^{1/3}\{\underline{f}(q_{c\cdot N^{-1/3}}-x_L)-\bar{f}(q_{N^{-1/3}}-x_L)\}. \qquad \textbf{(A14)}$$

Now, we show

$$a_c \to \infty \quad \text{as } c \to \infty. \qquad \textbf{(A15)}$$

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

$$N^{-1/3}/\bar{f} \leq q_{(c+1)\cdot N^{-1/3}} - q_{(c)\cdot N^{-1/3}} \leq N^{-1/3}/\underline{f}. \qquad \textbf{(A16)}$$

Combining (A14) and (A16) yields $a_c \geq p^{(1)}(x_L)\cdot c(\underline{f}/\bar{f})+O(1)$, which implies (A15).

Furthermore, by (A14), we have

$$c_1 > c \Rightarrow a_{c_1} > a_c. \qquad \textbf{(A17)}$$

On the other hand, for a defined in (A11), applying (A13) to a yields

$$a = a_{c_1} + O_p(1). \qquad \textbf{(A18)}$$

Now, we study b in (A12). Note that for a binary W and $p(X)=\mathbb{E}(W|X)$, we have $\mathbb{V}(W-p(X))=\{1-p(X)\}p(X)$. Thus, for any $c>2$,

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

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

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

$$\sum_{i=1}^{N_1} \mathbb{V}(B_i) = \sum_{i=1}^{c_1 N^{2/3}} \mathbb{V}(B_i) = b_{c_1}^2 + o_p(1). \tag{A19}$$

Due to (A15), (A17), and (A18), for any $\nu > 0$, we can choose a large enough N and c such that

$$\mathbb{P}\left(a < \frac{1}{2}a_c,\ c_1 > c\right) \le \mathbb{P}\left(a < \frac{1}{2}a_c,\ a_{c_1} > a_c\right) < \frac{\nu}{4}. \tag{A20}$$

Further, note that $b_c^2 = \frac{c}{c-1}\{1-p(x_L)\}p(x_L)$ is decreasing in c when $c > 1$. By (A19), we have

$$\mathbb{P}\left(\sum_{i=1}^{c_1 N^{2/3}} \mathbb{V}(B_i) \ge 2b_c^2,\ c_1 > c\right) \le \mathbb{P}\left(\sum_{i=1}^{c_1 N^{2/3}} \mathbb{V}(B_i) \ge 2b_{c_1}^2,\ c_1 > c\right) < \frac{\nu}{4}. \tag{A21}$$

Now, we use the Bernstein inequality. Since B_i defined in (A11) is a centered and normalized binary variable, we can simply choose $A = 1$ in (A12), then

$$\begin{aligned}
\mathbb{P}(c_1 > c) \le& \mathbb{P}\left(\sum_{i=1}^{c_1 N^{2/3}} B_i > a,\ c_1 > c\right) \\
\le& \mathbb{P}\left(\sum_{i=1}^{N_1} B_i \ge a,\ a \ge \frac{1}{2}a_c,\ c_1 > c\right) + \frac{\nu}{4} \\
\le& \mathbb{P}\left(\sum_{i=1}^{N_1} B_i \ge a,\ a \ge \frac{1}{2}a_c,\ c_1 > c,\ \sum_{i=1}^{c_1 N^{2/3}} \mathbb{V}(B_i) < 2b_c^2\right) + \frac{\nu}{2} \\
\le& \mathbb{P}\left(\sum_{i=1}^{N_1} B_i \ge \frac{1}{2}a_c,\ \sum_{i=1}^{N_1} \mathbb{V}(B_i) < 2b_c^2\right) + \frac{\nu}{2} \\
\le& \exp\left(-\frac{\frac{1}{4}a_c^2}{a_c + 4b_c^2}\right) + \frac{\nu}{2} \le \nu.
\end{aligned} \tag{A22}$$

After choosing a large enough N and c, the second inequality follows from (A20); the third inequality follows from (A21); the fourth inequality follows from (A9) and $\sum_{i=1}^{N_1} B_i \ge a,\ a \ge \frac{1}{2}a_c \Rightarrow \sum_{i=1}^{N_1} B_i \ge \frac{1}{2}a_c$; and the fifth inequality follows from Bernstein inequality. Consequently, the conclusion $N_1 = O_p(N^{2/3})$ follows.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

A.4. Proof of Theorem 1

Let q_α denote the α-th quantile of X. We define the following sequences of positive numbers:

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

where n_K is defined in Section 2.1, which is the first element of partition K (the last partition). c_1 is defined in (A9). 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 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 A1).

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

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

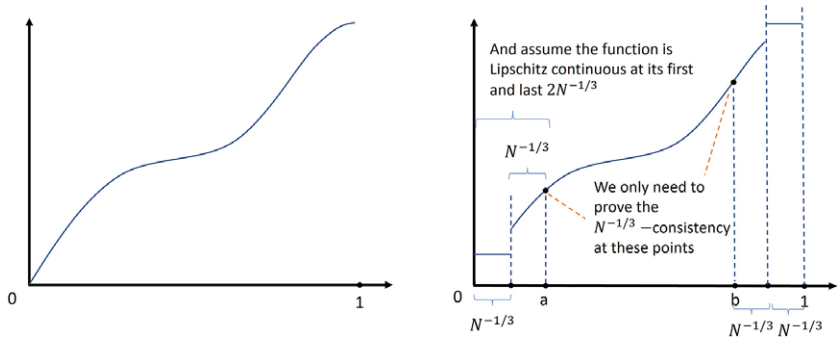

FIGURE A1. UC-isotonic estimator. The left panel is $p(x)$, and the right panel is $p_N(x)$.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

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

$$x - x_L = O_p(N^{-1/3}) \quad \text{for } x \in [x_L, a_N]. \tag{A23}$$

Thus, we have

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

where the first equality follows from Proposition 3 (ii); the $O_p(N^{-1/2})$ term in the third equality follows from that $X_{c_1 N^{2/3}}$ 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 (A23) and Assumption 2; and the last equality follows from $c_1 = O_p(1)$, which is implied by (A22).

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 (2012, Thm. 2.1). We can adapt the domain 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

$$\begin{aligned}
\hat{\tau} &= \frac{1}{N}\sum_{i=1}^{N} (2W_i - 1)\left(Y_i - \frac{1}{M_i}\sum_{j \in \mathcal{J}(i)} Y_j\right) \\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\sum_{i=n_k}^{n_k+N_k-1} (2W_i - 1)\left(Y_i - \frac{1}{N_{k,1-W_i}}\sum_{j \in \mathcal{J}(i)} Y_j\right)\right\}
\end{aligned}$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

$$
\begin{aligned}
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_i=1}\left(Y_i-\frac{1}{N_{k,0}}\sum_{j\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_j=0}Y_j\right)\right.\\
&\quad\left.+\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_i=0}\left(\frac{1}{N_{k,1}}\sum_{j\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_j=1}Y_j-Y_i\right)\right\}\\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\left(1+\frac{N_{k,0}}{N_{k,1}}\right)\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_i=1}Y_i\right.\\
&\quad\left.-\left(\frac{N_{k,1}}{N_{k,0}}+1\right)\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_i=0}Y_i\right\}\\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\frac{N_{k,1}+N_{k,0}}{N_{k,1}}\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_i=1}Y_i\right.\\
&\quad\left.-\frac{N_{k,1}+N_{k,0}}{N_{k,0}}\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_i=0}Y_i\right\}\\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_i=1}\frac{Y_i}{N_{k.1}/N_k}-\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\},\,W_i=0}\frac{Y_i}{N_{k.0}/N_k}\right\}\\
&= \frac{1}{N}\sum_{k=1}^{K}\left\{\sum_{i\in\{n_k,\ldots,(n_k+N_k-1)\}}\left(\frac{W_iY_i}{\tilde{p}(X_i)}-\frac{(1-W_i)Y_i}{1-\tilde{p}(X_i)}\right)\right\}\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{\tilde{p}(X_i)}-\frac{(1-W_i)Y_i}{1-\tilde{p}(X_i)}\right), \qquad \textbf{(A24)}
\end{aligned}
$$

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

Given Theorem 2, it is sufficient to show that the last line of (A24) has the desired properties. By the Taylor extension,

$$
\begin{aligned}
\frac{W_iY_i}{\tilde{p}(X_i)}-\frac{(1-W_i)Y_i}{1-\tilde{p}(X_i)} &= \left(\frac{W_iY_i}{p(X_i)}-\frac{(1-W_i)Y_i}{1-p(X_i)}\right)-\left(\frac{W_iY_i}{p(X_i)^2}+\frac{Y_i(1-W_i)}{\{1-p(X_i)\}^2}\right)\{\tilde{p}(X_i)-p(X_i)\}\\
&\quad+2\left(\frac{W_iY_i}{\check{p}_i^3}-\frac{Y_i(1-W_i)}{\{1-\check{p}_i\}^3}\right)\{\tilde{p}(X_i)-p(X_i)\}^2, \qquad \textbf{(A25)}
\end{aligned}
$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

where the random variable $\check{p}_i$ takes values in the interval between $\tilde{p}(X_i)$ and $p(X_i)$. For $Z=(Y,X,W)$, we define

$$D(Z)=\frac{WY}{p(X)^2}+\frac{Y(1-W)}{\{1-p(X)\}^2}. \tag{A26}$$

For any $x\in\mathcal{X}$, we denote $\mathbb{E}[D(Z)|X=x]$ by $\mathbb{E}[D(Z)|x]$. Then, the second term in the first line of (A25) can be written as:

$$\begin{aligned}
&-D(Z_i)\{\tilde{p}(X_i)-p(X_i)\}\\
=&-\mathbb{E}[D(Z_i)|X_i]\{\tilde{p}(X_i)-p(X_i)\}-\{D(Z_i)-\mathbb{E}[D(Z_i)|X_i]\}\{\tilde{p}(X_i)-p(X_i)\}\\
=&-\mathbb{E}[D(Z_i)|X_i][\{W_i-p(X_i)\}-\{W_i-\tilde{p}(X_i)\}]\\
&-\{D(Z_i)-\mathbb{E}[D(Z_i)|X_i]\}\{\tilde{p}(X_i)-p(X_i)\}\\
=&-\mathbb{E}[D(Z_i)|X_i]\{W_i-p(X_i)\}+\mathbb{E}[D(Z_i)|X_i]\{W_i-\tilde{p}(X_i)\}\\
&-\{D(Z_i)-\mathbb{E}[D(Z_i)|X_i]\}\{\tilde{p}(X_i)-p(X_i)\}.
\end{aligned}$$

Plugging it back to (A25), we have

$$\begin{aligned}
\hat{\tau}-\tau=&\frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{W_iY_i}{p(X_i)}-\frac{(1-W_i)Y_i}{1-p(X_i)}-\tau\right)-\mathbb{E}[D(Z_i)|X_i]\{W_i-p(X_i)\}\right]\\
&+\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[D(Z_i)|X_i]\{W_i-\tilde{p}(X_i)\}\\
&-\frac{1}{N}\sum_{i=1}^{N}\{D(Z_i)-\mathbb{E}[D(Z_i)|X_i]\}\{\tilde{p}(X_i)-p(X_i)\}\\
&+\frac{2}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{\check{p}_i^3}-\frac{Y_i(1-W_i)}{\{1-\check{p}_i\}^3}\right)\{\tilde{p}(X_i)-p(X_i)\}^2\\
=:&I+II-III+IV.
\end{aligned} \tag{A27}$$

We will show the asymptotic properties of these four terms in the subsequent subsections.

A.6.1. *The Limit of I.* Note that $\mathbb{E}[D(Z_i)|X_i]=\mathbb{E}\left[\frac{W_iY_i}{p(X_i)^2}+\frac{(1-W_i)Y_i}{\{1-p(X_i)\}^2}\middle|X_i\right]=\frac{\mathbb{E}[Y(1)]}{p(X_i)}+\frac{\mathbb{E}[Y(0)]}{1-p(X_i)}$. Therefore, by Theorem 1 of Hirano et al. (2003) (see also their equations (12) and (38)), it holds

$$\sqrt{N}\cdot I\xrightarrow{d}N(0,\Omega), \tag{A28}$$

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

A.6.2. *The Rate of II.* Since $\tilde{p}(\cdot)$ is the isotonic estimator of regressing $\{\tilde{W}_i\}_{i=1}^N$ on $\{X_i\}_{i=1}^N$, by the construction of the isotonic estimator (see, e.g., Groeneboom and Jongbloed, 2014, Lemmas 2.1 and 2.3; see also Barlow and Brunk, 1972), we have $\sum_{i=n_k}^{n_{k+1}-1}\{\tilde{W}_i-\tilde{p}(X_i)\}=0$ for each $k=1,\ldots,K$. (For the last summand, we can simply set $n_{K+1}=N+1$.)

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

By Proposition 3 (i) and the construction of $\tilde{W}$ (given by (6)), it holds that $\sum_{i=n_k}^{n_{k+1}-1}\{W_i - \tilde{p}(X_i)\} = 0$ for each $k = 1, \ldots, K$. As a result,

$$\sum_{k=1}^{K} m_k \sum_{i=n_k}^{n_{k+1}-1} \{W_i - \tilde{p}(X_i)\} = 0 \tag{A29}$$

holds for any weights $\{m_k\}_{k=1}^K$. To proceed, we define the function $\delta(x)$ as

$$\delta(x) = \mathbb{E}[D(Z)|X = x], \tag{A30}$$

and its corresponding step function $\bar{\delta}_N(x)$ as

$$\bar{\delta}_N(x) = \begin{cases} \delta(X_{n_k}) & \text{if } p(x) > \tilde{p}(X_{n_k}) \text{ for all } x \in (X_{n_k}, X_{n_{k+1}}) \\ \delta(s) & \text{if } p(s) = \tilde{p}(s) \text{ for some } s \in (X_{n_k}, X_{n_{k+1}}) \\ \delta(X_{n_{k+1}}) & \text{if } p(x) < \tilde{p}(X_{n_k}) \text{ for all } x \in (X_{n_k}, X_{n_{k+1}}), \end{cases}$$

for each $x \in [X_{n_k}, X_{n_{k+1}})$ with $k = 1, \ldots, K$ (if $k = K$, set $X_{n_{k+1}} = \max_k X_{n_k}$). Further, we define $z = (y, w, x)$ as a given vector belonging to the domain of the random vector $Z = (Y, W, X)$. By (A29), it holds

$$\int \bar{\delta}_N(x)\{w - \tilde{p}(x)\} d\mathbb{P}_N(z) = 0,$$

where $\mathbb{P}_N$ is the empirical measure. Thus, we have

$$\begin{aligned} II &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[D(Z_i)|X_i]\{W_i - \tilde{p}(X_i)\} = \int \delta(x)\{w - \tilde{p}(x)\} d\mathbb{P}_N(z) \\ &= \int \{\delta(x) - \bar{\delta}_N(x)\}\{w - \tilde{p}(x)\} d\mathbb{P}_N(z). \end{aligned} \tag{A31}$$

By definition, $\delta(x)$ is a bounded function with a finite total variation, so is $\bar{\delta}_N(x)$. For $\mathbb{P}_0$ denoting the joint probability measure of (Y, W, X), the last row of (A31) can be decomposed as:

$$\begin{aligned} &\int \{\delta(x) - \bar{\delta}_N(x)\}\{w - \tilde{p}(x)\} d\mathbb{P}_n(z). \\ =& \int \{\delta(x) - \bar{\delta}_N(x)\}\{w - \tilde{p}(x)\} d(\mathbb{P}_n(z) - \mathbb{P}_0(z)) \\ &+ \int \{\delta(x) - \bar{\delta}_N(x)\}\{w - p(x)\} d\mathbb{P}_0(z) + \int \{\delta(x) - \bar{\delta}_N(x)\}\{p(x) - \tilde{p}(x)\} d\mathbb{P}_0(z) \\ =:& II_1 + II_2 + II_3. \end{aligned}$$

It shall be noted that by Assumption 4, $\delta(x) - \bar{\delta}_N(x)$ is a bounded function with a finite total variation, and $\delta(x)$ is continuously differentiable in x. Under Assumption 2 (ii) and similar arguments following (10.64) of Groeneboom and Jongbloed (2014), it holds that for some $C_0 > 0$ and all $x \in \mathcal{X}$,

$$|\delta(x) - \bar{\delta}_N(x)| \leq C_0 |p(x) - \tilde{p}(x)|. \tag{A32}$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

A.6.3. *The Rate of II_1.* For $R := \max\{|x_L|, |x_u|\}$ and a positive constant K, let us define

$$\begin{aligned}
\mathcal{M}_{RK} &= \{\text{monotone increasing functions on} [-R, R] \text{and bounded by} K\},\\
\mathcal{G}_{RK} &= \{g : g(x) = p(x), x \in \mathcal{X}, p \in \mathcal{M}_{RK}\},\\
\mathcal{D}_{RKv} &= \{d : d(x) = g_1(x) - g_2(x),\ (g_1, g_2) \in \mathcal{G}_{RK}^2,\ \|d(\cdot)\|_{\mathbb{P}_0} \le v\},\\
\mathcal{H}_{RKv} &= \{h : h(y, x) = wd_1(x) - d_2(x),\ (d_1, d_2) \in \mathcal{D}_{RKv}^2,\ z \in \mathcal{Z}\},
\end{aligned} \tag{A33}$$

where $\|f\|_{\mathbb{P}} = \sqrt{\int |f(x)|^2 d\mathbb{P}(x)}$ denotes the $L_2(\mathbb{P})$ norm of function f, given the probability measure $\mathbb{P}$. Then, the integrand of II_1 can be written as

$$\{\delta(x) - \bar{\delta}_N(x)\}\{w - \tilde{p}(x)\} = \{\delta(x) - \bar{\delta}_N(x)\}w - \{\delta(x) - \bar{\delta}_N(x))\}\tilde{p}(x). \tag{A34}$$

Furthermore, we define

$$\mathcal{F}_a = \left\{f : f(z) = \{\delta(x) - \bar{\delta}_N(x)\}w - \{\delta(x) - \bar{\delta}_n(x)\}\tilde{p}(x),\ z \in \mathcal{Z}\right\}.$$

We note the following points:

(i) By Assumption 2 and the construction of $\bar{\delta}_N(x)$, $\{\delta(x) - \bar{\delta}_N(x)\}$ is a bounded function of x with a finite total variation.

(ii) $\tilde{W} \in [0, 1]$ implies that $\sup_{x \in \mathcal{X}} |\tilde{p}(x)| \le 1$. Therefore, for any constant $K_1 > 1$, it holds that $\tilde{p}(x) \in \mathcal{G}_{RK_1}$.

(iii) By Theorem 1 and (A32), there exists some constant $C_1 > 0$, such that $\left\|\delta(x) - \bar{\delta}_N(x)\right\|_{\mathbb{P}} \le C_1 \left(\frac{\log N}{N}\right)^{1/3}$ holds with the probability arbitrarily close to one (hereafter denoted as w.p.a.1) if we choose a large enough C_1. Therefore, by (i) and Lemma 21 in the Supplementary Material of BGH (BGH-supp hereafter),[8] for a constant C_2 that is larger than twice the bound of $\delta(x)$ (which is guaranteed by Assumption 4) and $v_1 = C_1 \left(\frac{\log N}{N}\right)^{1/3}$, it holds that $\{\delta(x) - \bar{\delta}_N(x)\} \in \mathcal{D}_{RC_2v_1}$, w.p.a.1.

(iv) By (i), (ii), a similar argument to that (iii), Theorem 1, Jensen's inequality, and the fact that the product of two monotone increasing functions remains monotone increasing, we have $\{\delta(x) - \bar{\delta}_n(x)\}\tilde{p}(x) \in \mathcal{D}_{RKv}$ holds for constants $K = C_2K_1$ and $v = v_1K_1$, w.p.a.1.

(v) By definition of function classes presented by (A33), we have $\mathcal{F}_a \subseteq \mathcal{H}_{RKv}$, w.p.a.1. Let $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ be the ϵ-bracketing number of the function class $\mathcal{F}$ under the norm $\|\cdot\|$, and

$$H_B(\epsilon, \mathcal{F}, \|\cdot\|) = \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$$

be the entropy of $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$. Furthermore, let us define

$$J_B(\delta, \mathcal{F}, \|\cdot\|) := \int_0^{\delta} \sqrt{1 + H_B(\epsilon, \mathcal{F}, \|\cdot\|)}\, d\epsilon.$$

By Theorem 2.7.5 in van der Vaart and Wellner (1996) and Lemma 11 in BGH-supp, given a positive constant C and a bracket size ϵ, there exists a constant $A > 0$, such that the entropy

[8] The lemma states that a function f, which is bounded and has a finite total variation, can be decomposed as $f = f_1 - f_2$, where both f_1 and f_2 are bounded and monotone increasing.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

of the function class $\mathcal{D}_{RCv}$ satisfies

$$H_B(\epsilon, \mathcal{D}_{RCv}, \|\cdot\|_{\mathbb{P}_0}) \leq \frac{AC}{\epsilon}.$$

Let (d_1^L, d_1^U) and (d_2^L, d_2^U) be two ϵ-brackets for the function class $\mathcal{D}_{RKv}$. Now, we calculate the entropy of $\mathcal{H}_{RKv}$, by using a set of brackets derived from the ϵ-brackets of $\mathcal{D}_{RKv}$. Note that $w \in \{0, 1\}$ is nonnegative. Then, we can define a bracket (h^L, h^U) within $\mathcal{H}_{RKv}$ as

$$h^L = wd_1^L(x) - d_2^U(x), \qquad h^U = wd_1^U(x) - d_2^L(x),$$

and its size is

$$\begin{aligned}
&\sqrt{\int \{h^U(z) - h^L(z)\}^2 d\mathbb{P}_0(z)} \\
\leq &\sqrt{2\left[\int w^2\{d^U(x) - d^L(x)\}^2 d\mathbb{P}_0(z) + \int \{d^U(x) - d^L(x)\}^2 d\mathbb{P}_0(z)\right]} \\
\leq &\sqrt{4\int \{d^U(x) - d^L(x)\}^2 d\mathbb{P}_0(x)} \leq 2\epsilon,
\end{aligned}$$

where the last inequality follows from the definition of ϵ-bracket with respect to (w.r.t.) the $L_2(\mathbb{P}_0)$ norm. As a result, for a constant $\tilde{A} > 0$, it holds

$$H_B(\epsilon, \mathcal{H}_{RCv}, \|\cdot\|_{\mathbb{P}_0}) \leq \frac{\tilde{A}C}{\epsilon}. \tag{A35}$$

Combining the point (v) above and (A35), there exists $C_3 > 0$, such that

$$H_B(\epsilon, \mathcal{F}_a, \|\cdot\|_{\mathbb{P}_0}) \leq \frac{C_3}{\epsilon} \tag{A36}$$

holds w.p.a.1. For $f_a \in \mathcal{F}_a$, points (iii) and (iv) imply that

$$\|f_a\|_{\mathbb{P}_0} \leq C_1 \left(\frac{\log N}{N}\right)^{1/3} \tag{A37}$$

holds w.p.a.1.

We have defined $\mathbb{P}_0$ to be the joint probability measure of $Z = (Y, W, X)$. With some abuse of notation, in the following, $\mathbb{P}$ will be used to denote $\mathbb{P}_0$ whenever the context permits without risk of confusion. Further, we use $\mathscr{E}$ to denote the event that both (A36) and (A37) happen. Note that we can select sufficiently large constants to ensure that $\mathbb{P}(\mathscr{E})$ approaches as close to one as desired.

In the following, we define $\|\mathbb{G}_N\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\sqrt{N}(\mathbb{P}_N - \mathbb{P}_0)f|$, and we use $J_B(\delta)$ to denote $J_B(\delta, \mathcal{F}_a, \|\cdot\|_{B,\mathbb{P}_0})$. Let $\eta_N := C_1\left(\frac{\log N}{N}\right)^{1/3}$. For any positive constants B and ν, there exist positive constants B_1, B_2, and C_2, such that for all N large enough,

$$\begin{aligned}
&\mathbb{P}(|II_1| > BN^{-1/2}) \leq \mathbb{P}(|II_1| > BN^{-1/2}, \mathscr{E}) + \mathbb{P}(\mathscr{E}^c) \\
\leq &\mathbb{P}\left(\|\mathbb{G}_N\|_{\mathcal{F}_a} > B, \mathscr{E}\right) + \frac{\nu}{2} \leq \frac{\mathbb{E}\left(\|\mathbb{G}_N\|_{\mathcal{F}_a} | \mathscr{E}\right)}{B} + \frac{\nu}{2}
\end{aligned}$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

$$\lesssim \frac{1}{B} J_B(\eta_N)\left(1+\frac{J_B(\eta_N)}{\sqrt{N}\eta_N^2}\right)+\frac{\nu}{2} \lesssim \frac{1}{B}(\eta_N+2B_1^{1/2}\eta_N^{1/2})\left(1+\frac{\eta_N+2B_1^{1/2}\eta_N^{1/2}}{\sqrt{N}\eta_N^2}\right)+\frac{\nu}{2}$$

$$\lesssim C_2\left(\frac{\log N}{N}\right)^{1/6}\times\left(1+\frac{B_2}{(\log N)^{\frac{1}{2}}}\right)+\frac{\nu}{2}\leq \nu, \tag{A38}$$

where the third inequality follows from the Markov inequality; the first wave inequality ($\lesssim$) follows from Lemma 3.4.2 of van der Vaart and Wellner (1996); the second wave inequality follows from (A36) and equation (.2) in BGH-supp[9]; and the third wave inequality follows from $\eta_N \lesssim \eta_N^{1/2}$ and the definition of η_N. Since ν can be chosen arbitrarily small, we obtain

$$II_1 = o_p(N^{-1/2}). \tag{A39}$$

A.6.4. *The Rate of II_2.* By the law of iterated expectations and $p(x)=\mathbb{E}[W|X=x]$,

$$II_2 = \int \{\delta(x)-\bar{\delta}_N(x)\}\{w-p(x)\}d\mathbb{P}_0(z)$$
$$= \int \{\delta(x)-\bar{\delta}_N(x)\}\mathbb{E}[\{W-p(X)\}|X=x]d\mathbb{P}_0(x)=0.$$

A.6.5. *The Rate of II_3.* The inequality (A32) implies:

$$II_3 = \int \{\delta(x)-\bar{\delta}_N(x)\}\{p(x)-\tilde{p}(x)\}d\mathbb{P}_0(z)$$
$$\lesssim \int \{p(x)-\tilde{p}(x)\}^2 d\mathbb{P}_0(z) = O_p\left(\left(\frac{\log N}{N}\right)^{2/3}\right) = o_p(N^{-1/2}),$$

where the wave inequality follows from (A32), and the second equality follows from Theorem 1. Combining the rates for II_1, II_2, and II_3, we have

$$II = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[D(Z_i)|X_i]\{W_i-\tilde{p}(X_i)\} = o_p(N^{-1/2}). \tag{A40}$$

A.6.6. *The Rate of III.* By defining $V(\cdot)=D(\cdot)-\mathbb{E}[D(Z)|X=\cdot]$, we have

$$III = \int V(z)\{\tilde{p}(x)-p(x)\}d\mathbb{P}_N(z)$$
$$= \int V(z)\{\tilde{p}(x)-p(x)\}d(\mathbb{P}_n(z)-\mathbb{P}_0(z)) + \int V(z)\left[\tilde{p}(x)-p(x)\right]d\mathbb{P}_0(z)$$
$$=: III_1 + III_2.$$

[9]Equation (.2) in BGH-supp states that $J_B(\delta)\leq \delta + 2C^{1/2}\delta^{1/2}$ holds for some positive constant C.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

By definition, $\mathbb{E}[V(Z)|X=x]=0$ holds for all $x\in\mathcal{X}$. Then, $III_2=0$ follows from the law of iterated expectations.

The derivation of the rate of III_1 is similar to that of II_1. The main difference is $D(z)$ not being assumed to be uniformly bounded over $\mathcal{Z}$. Consequently, we shall use Lemma 3.4.3 from van der Vaart and Wellner (1996), instead of Lemma 3.4.2. The former is formulated w.r.t. the Bernstein norm and is suitable for unbounded function classes. To simplify our discussion and avoid cumbersome notation, we will reuse some notations previously introduced in Appendix A.6.2 (e.g., those notations denoting various constants), provided it does not lead to confusion in the context.

Recall the function classes $\mathcal{M}_{RK}$, $\mathcal{G}_{RK}$, and $\mathcal{D}_{RKv}$ defined in (A33). Here, we define additionally

$$\mathcal{H}^{(2)}_{RKv}=\{h: h(z)=V(z)d(x),\ d(\cdot)\in\mathcal{D}_{RKv},\ z\in\mathcal{Z}\},$$
$$\mathcal{F}_b=\left\{f: f(z)=V(z)\{\tilde{p}(x)-p(x)\},\ z\in\mathcal{Z}\right\}.$$

Let (d^L,d^U) be any ϵ-bracket within the function class $\mathcal{D}_{RCv}$. Define

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

Note that (h^L,h^U) is a bracket within $\mathcal{H}^{(2)}_{RKv}$, and its size is

$$\sqrt{\int\{h^U(z)-h^L(z)\}^2d\mathbb{P}_0(z)}=\sqrt{\int D(z)^2\{d^U(x)-d^L(x)\}^2d\mathbb{P}_0(z)}$$
$$=\sqrt{\int\mathbb{E}\left[D(Z)^2|X=x\right]\{d^U(x)-d^L(x)\}^2d\mathbb{P}_0(x)}\leq A_1\epsilon,$$

for some $A_1>0$. The last inequality follows from Assumption 4 and the definition of ϵ-bracket w.r.t. the $L_2(\mathbb{P}_0)$ norm. As a result, for some $\tilde{A}>0$, we have

$$H_B(\epsilon,\mathcal{H}^{(2)}_{RCv},\|\cdot\|_{\mathbb{P}_0})\leq\frac{\tilde{A}C}{\epsilon}. \tag{A41}$$

We now switch to the Bernstein norm because we prefer not to impose a bound on $D(z)$. Let $\|\cdot\|_{B,\mathbb{P}}$ be the Bernstein norm under a measure $\mathbb{P}$. By the definition,

$$\|h\|^2_{B,\mathbb{P}}=2\mathbb{P}(\exp(|h|)-|h|-1)=2\int\sum_{k=2}^{\infty}\frac{1}{k!}|h|^kd\mathbb{P}(z),$$

where the second equality follows by the extension of the natural exponential function. Next, we attempt to bound the Bernstein norm of $H^{-1}h(\cdot)$, where H is a positive number

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

that we will select in subsequent steps to establish a finite upper bound. For a constant C and any $h \in \mathcal{H}^{(2)}_{RCv}$, it holds

$$\begin{aligned}
\left\|H^{-1}h^U - H^{-1}h^L\right\|^2_{B,\mathbb{P}_0} &= 2\int \sum_{k=2}^{\infty} \frac{1}{H^k}\frac{1}{k!}|D(z)\{d^U(x) - d^L(x)\}|^k d\mathbb{P}_0(z) \\
&\leq 2\int \sum_{k=2}^{\infty} \frac{1}{H^k}\frac{1}{k!}|D(z)|^k |d^U(x) - d^L(x)|^k d\mathbb{P}_0(z) \\
&\leq 2\sum_{k=2}^{\infty} \frac{1}{H^k}\frac{(4C)^{k-2}}{k!} k! M_0^{k-2} c_0 \int |d^U(x) - d^L(x)|^2 d\mathbb{P}_0(x) \\
&= \frac{2}{H^2}\sum_{k=2}^{\infty} \frac{(4M_0C)^{k-2}}{H^{k-2}} c_0 \int |d^U(x) - d^L(x)|^2 d\mathbb{P}_0(x) \\
&= \frac{2}{H^2}\sum_{k=2}^{\infty} \left(\frac{4M_0C}{H}\right)^{k-2} c_0\epsilon^2 = 2c_0\left(\frac{\epsilon}{H}\right)^2,
\end{aligned}$$

where the second inequality follows from Assumption 4 and $d(\cdot) \leq 2C$ (implied by $d(\cdot) \in \mathcal{D}_{RCv}$); c_0 and M_0 are the same constants defined in Assumption 4 (iii); the third equality follows from the definition of the ϵ-bracket; and the last equality follows from choosing $H = 8M_0C$. As a result, we have for some positive constant C_1,

$$H_B(\epsilon, \mathcal{H}^{(2)}_{RCv}, \|\cdot\|_{B,\mathbb{P}_0}) \leq \frac{C_1}{\epsilon}. \tag{A42}$$

Further, by similar arguments, for any $h \in \mathcal{H}^{(2)}_{RCv}$, it holds

$$\begin{aligned}
\left\|H^{-1}h\right\|^2_{B,\mathbb{P}_0} &\leq 2\int \sum_{k=2}^{\infty} \frac{1}{H^k}\frac{1}{k!}|D(z)|^k |d(x)|^k d\mathbb{P}_0(z) \\
&= \frac{2}{H^2}\sum_{k=2}^{\infty} \frac{(2M_0C)^{k-2}}{H^{k-2}} c_0 \int |d(x)|^2 d\mathbb{P}_0(x) \\
&= \frac{2}{H^2}\sum_{k=2}^{\infty} \left(\frac{2M_0C}{H}\right)^{k-2} c_1 v^2 = \left(\frac{2}{H}\right)^2 c_1 v^2,
\end{aligned}$$

where the third equality follows from $d(\cdot) \in \mathcal{D}_{RCv}$. Thus,

$$\left\|H^{-1}h\right\|_{B,\mathbb{P}_0} \lesssim \frac{v}{H}. \tag{A43}$$

Based on these results, we now study the function class $\mathcal{F}_b$. Set $C \geq 1$ and $v := C_2\left(\frac{\log N}{N}\right)^{1/3}$ for some $C_2 > 0$. Then, by Theorem 1 and the definitions of $\mathcal{H}^{(2)}_{RCv}$ and $\mathcal{F}_b$,

$$\mathcal{F}_b \subseteq \mathcal{H}^{(2)}_{RCv} \tag{A44}$$

holds w.p.a.1. Furthermore, we define

$$\tilde{\mathcal{F}}_b = H^{-1}\mathcal{F}_b. \tag{A45}$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

By combining (A42), (A44), and (A45), there exists $C_2 > 0$, such that

$$H_B(\epsilon, \tilde{\mathcal{F}}_b, \|\cdot\|_{B,\mathbb{P}_0}) \leq \frac{C_2}{\epsilon} \tag{A46}$$

holds w.p.a.1. Further, for $f_b \in \mathcal{F}_a$, we define $\tilde{f}_b = H^{-1} f_b$. By (A43) and $\nu = C_2 \left(\frac{\log N}{N}\right)^{1/3}$, there exists some $C_3 > 0$, such that

$$\left\|\tilde{f}_b\right\|_{B,\mathbb{P}_0} \leq C_3 \left(\frac{\log N}{N}\right)^{1/3} \tag{A47}$$

holds w.p.a.1.

Again, we use $\mathscr{E}$ to denote the event that both (A46) and (A47) happen. With sufficiently large constants selected, $\mathbb{P}(\mathscr{E})$ can approach as close to one as desired. For $\eta_N := C_3 \left(\frac{\log N}{N}\right)^{1/3}$ and any positive constants B and ν, there exist constants B_1, B_2, and C_4, such that for all N large enough, it holds that

$$\begin{aligned}
&\mathbb{P}(|III_1| > BN^{-1/2}) \leq \mathbb{P}(|II_1| > BN^{-1/2}, \mathscr{E}) + \mathbb{P}(\mathscr{E}^c) \\
\leq &\mathbb{P}\left(\|\mathbb{G}_N\|_{\mathcal{F}_b} > B, \mathscr{E}\right) + \frac{\nu}{2} \leq \frac{H}{B}\mathbb{E}\left(\|\mathbb{G}_N\|_{\tilde{\mathcal{F}}_b} | \mathscr{E}\right) + \frac{\nu}{2} \\
\lesssim &\frac{H}{B} J_B(\eta_N)\left(1 + \frac{J_B(\eta_N)}{\sqrt{N}\eta_N^2}\right) + \frac{\nu}{2} \lesssim \frac{H}{B}(\eta_N + 2B_1^{1/2}\eta_N^{1/2})\left(1 + \frac{\eta_N + 2B_1^{1/2}\eta_N^{1/2}}{\sqrt{N}\eta_N^2}\right) + \frac{\nu}{2} \\
\lesssim &C_4 \left(\frac{\log N}{N}\right)^{1/6} \times \left(1 + \frac{B_2}{(\log N)^{\frac{1}{2}}}\right) + \frac{\nu}{2} \leq \nu,
\end{aligned} \tag{A48}$$

where the third inequality follows from the Markov inequality and the definition of $\tilde{\mathcal{F}}_b$ in (A45); the first wave inequality ($\lesssim$) comes from Lemma 3.4.3 of van der Vaart and Wellner (1996); the second wave inequality comes from (A46) and equation (.2) in BGH-supp; and the third wave inequality follows from $\eta_N \lesssim \eta_N^{1/2}$ and the definition of η_N. Since ν can be chosen arbitrarily small, we obtain $III_1 = o_p(N^{-1/2})$. Combined with $III_2 = 0$ yields

$$III = o_p(N^{-1/2}). \tag{A49}$$

A.6.7. *The Rate of IV.* Note that

$$\begin{aligned}
IV =& \frac{2}{N}\sum_{i=1}^{N}\left(\frac{W_i}{\check{p}_i^3} - \frac{1-W_i}{\{1-\check{p}_i\}^3}\right) Y_i\{\tilde{p}(X_i) - p(X_i)\}^2 \\
=& \frac{2}{N}\sum_{i=1}^{N}\left(\frac{W_i}{\check{p}_i^3} - \frac{1-W_i}{\{1-\check{p}_i\}^3}\right) \{Y_i \mathbb{I}(Y_i > 0)\}\{\tilde{p}(X_i) - p(X_i)\}^2 \\
&+ \frac{2}{N}\sum_{i=1}^{N}\left(\frac{W_i}{\check{p}_i^3} - \frac{1-W_i}{\{1-\check{p}_i\}^3}\right) \{Y_i \mathbb{I}(Y_i \leq 0)\}\{\tilde{p}(X_i) - p(X_i)\}^2 \\
=:& IV_+ + IV_-.
\end{aligned}$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

For IV_+, we have

$$\begin{aligned}
|IV_+| \leq & 2 \sup_{i\in\{1:N\}} \left| \frac{W_i}{\check{p}_i^3} - \frac{1-W_i}{\{1-\check{p}_i\}^3} \right| \frac{1}{N} \sum_{i=1}^{N} \{Y_i \mathbb{I}(Y_i > 0)\} \{\tilde{p}(X_i) - p(X_i)\}^2 \\
\leq & 2 \sup_{i\in\{1:N\}} \left| \frac{1}{\check{p}_i^3} + \frac{1}{\{1-\check{p}_i\}^3} \right| \sup_{i\in\{1:N\}} \{\tilde{p}(X_i) - p(X_i)\}^2 \frac{1}{N} \sum_{i=1}^{N} \{Y_i \mathbb{I}(Y_i > 0)\} \\
\leq & 2 \cdot \left| \frac{1}{\underline{p}^3} + \frac{1}{\{1-\bar{p}\}^3} + o_p(1) \right| \cdot O_p\left(\frac{\log N}{N} \right)^{2/3} \cdot O_p(1) \\
= & o_p(N^{-1/2}),
\end{aligned}$$

where the second inequality follows from $W_i \in \{0, 1\}$, and $\check{p}_i$ being nonnegative (because both $\tilde{p}(X_i)$ and $p(X_i)$ are nonnegative); the third inequality follows from Assumptions 3 and 4 and Theorem 1.

By a similar argument, we have

$$\begin{aligned}
|IV_-| \leq & 2 \max_{1\leq i\leq N} \left| \frac{1}{\check{p}_i^3} + \frac{1}{\{1-\check{p}_i\}^3} \right| \max_{1\leq i\leq N} \{\tilde{p}(X_i) - p(X_i)\}^2 \frac{1}{N} \sum_{i=1}^{N} \{-Y_i \mathbb{I}(Y_i \leq 0)\} \\
= & o_p(N^{-1/2}),
\end{aligned}$$

and thus,

$$IV = o_p(N^{-1/2}). \tag{A50}$$

Remark A1. [The role of the UC-isotonic estimator] We observe the critical role played by the UC-isotonic estimator in establishing the rate of IV: it allows us to uniformly bound $\frac{W_i}{\check{p}_i^3} - \frac{1-W_i}{\{1-\check{p}_i\}^3}$ from above (with probability approaching one). However, if we were to use a standard isotonic estimator instead, there is no guarantee that those $\frac{W_i}{\check{p}_i^3} - \frac{1-W_i}{\{1-\check{p}_i\}^3}$ at boundaries are bounded. This lack of boundedness occurs because the bias inherent in the standard isotonic estimator at boundaries is not mitigated by increasing the sample size. Although this bias affects only the summands at the two shrinking boundaries, the bias is toward zero and will be disproportionately amplified by the reciprocal structure, considerably impacting the overall moment estimator. The consequence is partially exemplified by column (d) of Table S8 in the Supplementary Material. This column presents the case with a conservative per-averaging (only averaging the first and the last $\lfloor N^{1/3} \rfloor$), which closely approximates the scenario without pre-processing the data.

A.6.8. *Summary of Appendix A.6.* Combining (A28), (A40), (A49), and (A50) yields

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

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

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

A.7. Proof of Theorem 4

The proof is similar to that of Theorem 1. Since we have $\tilde{\alpha}-\alpha_0 = O_p(N^{-1/2}) = o_p(N^{-1/3})$, the estimation of α_0 does not affect the uniform convergence rate. All the steps in Appendix 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), we have

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}(2W_i-1)\left(Y_i - \frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{\tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha})} - \frac{(1-W_i)Y_i}{1-\tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha})}\right). \quad \textbf{(A51)}$$

It remains to derive the asymptotic properties of (A51). Recall that $\mathbb{E}_\alpha[D(Z)|u] := \mathbb{E}[D(Z)|X'\alpha = u]$. Similarly to (A27),

$$\begin{aligned}
\tilde{\tau}-\tau =& \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{W_iY_i}{p_0(X_i'\alpha_0)} - \frac{(1-W_i)Y_i}{1-p_0(X_i'\alpha_0)} - \tau\right) - \mathbb{E}_{\tilde{\alpha}}[D(Z)|X_i'\tilde{\alpha}]\{W_i - p_0(X_i'\alpha_0)\}\right] \\
&+ \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\tilde{\alpha}}[D(Z)|X_i'\tilde{\alpha}]\{W_i - \tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha})\} \\
&- \frac{1}{N}\sum_{i=1}^{N}\{D(Z_i) - \mathbb{E}_{\tilde{\alpha}}[D(Z)|X_i'\tilde{\alpha}]\}\{\tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha}) - p_0(X_i'\alpha_0)\} \\
&+ \frac{2}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{\check{p}_i^3} - \frac{Y_i(1-W_i)}{\{1-\check{p}_i\}^3}\right)\{\tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha}) - p_0(X_i'\alpha_0)\}^2 \\
=:& I_m + II_m - III_m + IV_m.
\end{aligned}$$

Among these terms, the convergence rates of I_m, II_m, and IV_m can be derived similarly as that in Section A.6, while III_m behaves differently than III. We will discuss these rates in the following subsections.

A.8.1. *The Limit of I_m.* It holds that

$$\begin{aligned}
I_m =& \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{W_iY_i}{p_0(X_i'\alpha_0)} - \frac{(1-W_i)Y_i}{1-p_0(X_i'\alpha_0)} - \tau\right) - \mathbb{E}_{\alpha_0}[D(Z_i)|X_i'\alpha_0]\{W_i - p_0(X_i'\alpha_0)\}\right] \\
&+ \frac{1}{N}\sum_{i=1}^{N}\{\mathbb{E}_{\alpha_0}[D(Z_i)|X_i'\alpha_0] - \mathbb{E}_{\tilde{\alpha}}[D(Z_i)|X_i'\tilde{\alpha}]\}\{W_i - p_0(X_i'\alpha_0)\} \\
=:& I_{m,1} + I_{m,2}.
\end{aligned}$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

Similarly to (A30), we define $\delta_\alpha(u) = \mathbb{E}[D(Z)|X'\alpha = u]$. Then,

$$\begin{aligned} I_{m,2} &= \int \{\delta_{\alpha_0}(x'\alpha_0) - \delta_{\tilde{\alpha}}(x'\tilde{\alpha})\}\{w - p_0(x'\alpha_0)\} d(\mathbb{P}_n(z) - \mathbb{P}_0(z)) \\ &\quad + \int \{\delta_{\alpha_0}(x'\alpha_0) - \delta_{\tilde{\alpha}}(x'\tilde{\alpha})\}\{w - p_0(x'\alpha_0)\} d\mathbb{P}_0(z) \\ &=: I_{m,2,1} + I_{m,2,2}. \end{aligned}$$

Let us first study the rate of $I_{m,2,1}$. Given that the function class of $\left\{\delta_\alpha(u) = \mathbb{E}_\alpha[D(Z)|X'\alpha = u] : \alpha \in \mathcal{B}(\alpha_0, \delta_0), u \in I_\alpha = \{x'\alpha : x \in \mathcal{X}\}\right\}$ is parameterized by a k-dimensional parameter α, the ϵ-bracket number of this class is of the order of $\frac{1}{\epsilon}$ (see, e.g., van der Vaart and Wellner, 1996, Ex. 19.7). Its corresponding entropy is smaller than that presented in (A36).

Furthermore, BGH shows that $\tilde{\alpha}$ is $\sqrt{N}$-consistent of α_0 (recall that we have defined $\tilde{\alpha} = \hat{\alpha}$), and we know that $\delta_\alpha(u)$ is by construction differentiable w.r.t. α for all $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$ and $u \in I_\alpha = \{x'\alpha : x \in \mathcal{X}\}$. As a result, we have $\left\|\delta_{\alpha_0}(\cdot'\alpha_0) - \delta_{\tilde{\alpha}}(\cdot'\tilde{\alpha})\right\|_{\mathbb{P}_0} = O_p(N^{-1/2})$. Therefore, we can apply similar arguments for the term II_1 in Appendix A.6.2 to show that $I_{m,2,1} = o_p(N^{-1/2})$.

Finally, $I_{m,2,2} = 0$ by the law of iterated expectations. Thus, we have shown that $I_{m,2} = o_p(N^{-1/2})$. Consequently,

$$\begin{aligned} I_m &= \frac{1}{N}\sum_{i=1}^N \left[\left(\frac{W_i Y_i}{p_0(X_i'\alpha_0)} - \frac{(1-W_i)Y_i}{1-p_0(X_i'\alpha_0)} - \tau\right) - \mathbb{E}_{\alpha_0}[D(Z_i)|X_i'\alpha_0]\{W_i - p_0(X_i'\alpha_0)\}\right] \\ &\quad + o_p(N^{-1/2}) \\ &= \frac{1}{N}\sum_{i=1}^N \{m(Z_i) + M(Z_i)\} + o_p(N^{-1/2}), \end{aligned} \tag{A52}$$

where functions $m(\cdot)$ and $M(\cdot)$ are defined in Theorem 5.

A.8.2. *The Rates of II_m and IV_m.* By the consistency of $\tilde{\alpha}$ and Assumption 2', it holds that for all large N, the function $p_{\tilde{\alpha}}(u) = \mathbb{E}[W|X'\tilde{\alpha} = u]$ is monotone increasing, w.p.a.1. As a result, we can apply Theorem 4 and the same arguments presented in Appendix A.6.2 to show

$$II_m = o_p(N^{-1/2}). \tag{A53}$$

See pp. 17–20 of BGH-supp for a similar case concerning the monotone single-index model.

By Theorem 4 and the same arguments as presented in Appendix A.6.7, it holds that

$$IV_m = o_p(N^{-1/2}). \tag{A54}$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

A.8.3. *The Rates of III_m.* The term III_m can be decomposed as

$$\begin{aligned} III_m &= \frac{1}{N}\sum_{i=1}^{N}\{D(Z_i) - \mathbb{E}_{\tilde{\alpha}}[D(Z_i)|X_i'\tilde{\alpha}]\}\{\tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha}) - p_0(X_i'\alpha_0)\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\{D(Z_i) - \mathbb{E}_{\tilde{\alpha}}[D(Z_i)|X_i'\tilde{\alpha}]\}\{\tilde{p}_{\tilde{\alpha}}(X_i'\tilde{\alpha}) - p_{\tilde{\alpha}}(X_i'\tilde{\alpha})\} \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\{D(Z_i) - \mathbb{E}_{\tilde{\alpha}}[D(Z_i)|X_i'\tilde{\alpha}]\}\{p_{\tilde{\alpha}}(X_i'\tilde{\alpha}) - p_0(X_i'\alpha_0)\} \\ &=: III_{m,1} + III_{m,2}. \end{aligned} \tag{A55}$$

By the consistency of $\tilde{\alpha}$, Assumption 2', it holds that for all large N, the function $p_{\tilde{\alpha}}(u) = \mathbb{E}[W|X'\tilde{\alpha} = u]$ is monotone increasing, w.p.a.1. Therefore, we can apply Theorem 4 and the same arguments presented in Appendix A.6.6 to show

$$III_{m,1} = o_p(N^{-1/2}). \tag{A56}$$

For $III_{m,2}$, by Lemma 17 of BGH-supp, it holds that

$$\frac{\partial}{\partial \alpha^{(j)}} p_\alpha(x'\alpha)\bigg|_{\alpha=\alpha_0} = \{x^{(j)} - \mathbb{E}_{\alpha_0}[X^{(j)}|X'\alpha_0 = x'\alpha_0]\}p_0^{(1)}(x'\alpha_0),$$

where $\alpha^{(j)}$ and $x^{(j)}$ are j-th elements of vectors α and x. Extending $III_{m,2}$ around α_0 yields

$$\begin{aligned} III_{m,2} &= \frac{1}{N}\sum_{i=1}^{N}\{D(Z_i) - \mathbb{E}_{\tilde{\alpha}}[D(Z_i)|X_i'\tilde{\alpha}]\}\{p_{\tilde{\alpha}}(X_i'\tilde{\alpha}) - p_0(X_i'\alpha_0)\} \\ &= \frac{1}{N}\sum_{i=1}^{N}[D(Z_i) - \mathbb{E}_{\tilde{\alpha}}[D(Z_i)|X_i'\tilde{\alpha}]][X_i - \mathbb{E}_{\alpha_0}[X_i|X_i'\alpha_0]]'p_0^{(1)}(X_i'\alpha_0)(\tilde{\alpha} - \alpha_0) \\ &\quad + o_p(\tilde{\alpha} - \alpha_0) \\ &= (\tilde{\alpha} - \alpha_0)\frac{1}{N}\sum_{i=1}^{N}\{D(Z_i) - \mathbb{E}_{\alpha_0}[D(Z_i)|X_i'\alpha_0]\}[X_i - \mathbb{E}_{\alpha_0}[X_i|X_i'\alpha_0]]'p_0^{(1)}(X_i'\alpha_0) \\ &\quad + o_p(\tilde{\alpha} - \alpha_0), \end{aligned} \tag{A57}$$

where the last equality follows from $\mathbb{E}_{\alpha_0}[D(Z_i)|X_i'\alpha_0] - \mathbb{E}_{\tilde{\alpha}}[D(Z_i)|X_i'\tilde{\alpha}] = o_p(1)$. By the law of large numbers and the law of iterated expectations,

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N}[D(Z_i) - \mathbb{E}_{\alpha_0}[D(Z_i)|X_i'\alpha_0]][X_i - \mathbb{E}_{\alpha_0}[X_i|X_i'\alpha_0]]'p_0^{(1)}(X_i'\alpha_0) \\ &\to_p \mathbb{E}[\mathrm{Cov}_{\alpha_0}(D(Z), X|X'\alpha_0)p_0^{(1)}(X'\alpha_0)], \end{aligned} \tag{A58}$$

where $\mathrm{Cov}_{\alpha_0}(D(Z), X|\cdot) := \mathbb{E}\left\{[D(Z) - \mathbb{E}_{\alpha_0}[D(Z)|X'\alpha_0]][X_i - \mathbb{E}_{\alpha_0}[X|X'\alpha_0]]'|X'\alpha_0 = \cdot\right\}$.

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

Furthermore, by Theorem 5 of BGH and $\tilde{\alpha} \equiv \hat{\alpha}$, we have

$$\tilde{\alpha} - \alpha_0 = \mathbb{E}[p_0^{(1)}(X'\alpha_0)\mathrm{Cov}_{\alpha_0}(X|X'\alpha_0)]^- \frac{1}{N}\sum_{i=1}^{N}\{X_i - \mathbb{E}_{\alpha_0}[X_i|X_i'\alpha_0]\}\{W_i - p_0(X_i'\alpha_0)\} + o_p(\tilde{\alpha} - \alpha_0), \tag{A59}$$

where $\mathbf{B}^-$ represents the Moore–Penrose inverse of a matrix $\mathbf{B}$.

For each $i = 1, \ldots, N$, define

$$A(Z_i) = -\mathbb{E}[\mathrm{Cov}_{\alpha_0}(D(Z), X|X'\alpha_0)p_0^{(1)}(X'\alpha_0)]\mathbb{E}[p_0^{(1)}(X'\alpha_0)\mathrm{Cov}_{\alpha_0}(X|X'\alpha_0)]^- \times \{X_i - \mathbb{E}_{\alpha_0}[X_i|X_i'\alpha_0]\}\{W_i - p_0(X_i'\alpha_0)\}.$$

Then, combining (A55)–(A59) and $\tilde{\alpha} - \alpha_0 = O_p(N^{-1/2})$ yields

$$III_m = \frac{1}{N}\sum_{i=1}^{N} -A(Z_i) + o_p(N^{-1/2}). \tag{A60}$$

A.8.4. *Summary of Appendix A.8.* Combining (A52)–(A54) and (A60), we obtain

$$\sqrt{N}(\tilde{\tau} - \tau) = \sqrt{N}(I_m + II_m - III_m + IV_m) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{m(Z_i) + M(Z_i) + A(Z_i)\} + o_p(1).$$

A.9. Proof of Theorem 6

The proof is adapted from Groeneboom and Hendrickx (2017). By Theorem 2, it is sufficient to show the validity of the bootstrap approximation for $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{\tilde{p}(X_i)} - \frac{(1-W_i)Y_i}{1-\tilde{p}(X_i)}\right)$.

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 with the resample $\{Z_i^*\}_{i=1}^N$. By the same arguments for (A27), we have

$$\begin{aligned}\hat{\tau}^* - \tau &= \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{W_i^*Y_i^*}{p(X_i^*)} - \frac{(1-W_i^*)Y_i^*}{1-p(X_i^*)} - \tau\right) - \mathbb{E}[D(Z_i)|X_i^*]\{W_i^* - p(X_i^*)\}\right] \\ &+ \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[D(Z_i)|X_i^*]\left[W_i^* - \tilde{p}(X_i^*)\right] - \frac{1}{N}\sum_{i=1}^{N}\{D(Z_i) - \mathbb{E}[D(Z_i)|X_i^*]\}\{\tilde{p}(X_i^*) - p(X_i^*)\} \\ &+ \frac{2}{N}\sum_{i=1}^{N}\left(\frac{W_i^*Y_i^*}{\check{p}_i^3} - \frac{(1-W_i^*)Y_i^*}{\{1-\check{p}_i\}^3}\right)\{\tilde{p}(X_i^*) - p(X_i^*)\}^2 := I^* + II^* - III^* + IV^*. \end{aligned} \tag{A61}$$

For the term IV^*, with some abuse of notation, we use the same $\check{p}_i^3$ to denote a random value between $\tilde{p}(X_i^*)$ and $p(X_i^*)$. By similar arguments in Appendices A.6.2, A.6.6, and A.6.7, we

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

have

$$II^* = o_{p_M}(N^{-1/2}), \quad III^* = o_{p_M}(N^{-1/2}), \quad IV^* = o_{p_M}(N^{-1/2}),$$

where P_M is the probability measure in the bootstrap world defined in p. 3450 of Groeneboom and Hendrickx (2017). As a result, we have

$$\hat{\tau}^* - \tau = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{W_i^* Y_i^*}{p(X_i^*)} - \frac{(1-W_i^*)Y_i^*}{1-p(X_i^*)} - \tau\right) - \mathbb{E}[D(Z)|X_i^*]\{W_i^* - p(X_i^*)\}\right] + o_{p_M}(N^{-1/2}).$$

Define

$$m(Z,\tau,p(\cdot)) = \frac{YW}{p(X)} - \frac{Y(1-W)}{1-p(X)} - \tau, \qquad M(Z) = -\mathbb{E}[D(Z)|X]\{W - p(X)\}.$$

Then, we can write

$$\begin{aligned}\hat{\tau}^* - \tau = &\left[\frac{1}{N}\sum_{i=1}^{N} m(Z_i^*,\tau,p(X_i^*)) - \frac{1}{N}\sum_{i=1}^{N} m(Z_i,\tau,p(X_i))\right] \\ &+ \left[\frac{1}{N}\sum_{i=1}^{N} M(Z_i^*) - \frac{1}{N}\sum_{i=1}^{N} M(Z_i)\right] \\ &+ \frac{1}{N}\sum_{i=1}^{N}\{m(Z_i,\tau,p(X_i)) + M(Z_i)\} + o_{p_M}(N^{-1/2}). \end{aligned} \tag{A62}$$

From Appendix A.6.1, we have

$$\hat{\tau} - \tau = \frac{1}{N}\sum_{i=1}^{N}\{m(Z_i,\tau,p(X_i)) + M(Z_i)\} + o_p(N^{-1/2}). \tag{A63}$$

Subtracting (A63) from (A62),

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

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 the expectation under P_M. Consequently, a central limit theorem yields $\sqrt{N}(\hat{\tau}^* - \hat{\tau}) \xrightarrow{d} N(0,\Omega)$, where Ω is defined in Theorem 3. This proves the conclusion (i) of Theorem 6, i.e.,

$$\sup_{t\in\mathbb{R}} |\mathbb{P}^*\{\sqrt{N}(\hat{\tau}^* - \hat{\tau}) \le t\} - \mathbb{P}\{\sqrt{N}(\hat{\tau} - \tau) \le t\}| \xrightarrow{p} 0. \tag{A64}$$

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

For (ii), note that by the definition of $c^*_{1-\alpha}$, it holds that $\mathbb{P}^*\{\sqrt{N}(\hat{\tau}^* - \hat{\tau}) \le c^*_{1-\alpha}\} = 1 - \alpha + o_p(1)$. Combining this result with (A64) yields

$$
\begin{aligned}
&|\mathbb{P}\{\sqrt{N}(\hat{\tau} - \tau) \le c^*_{1-\alpha}\} - (1-\alpha)| \\
&\quad = |\mathbb{P}\{\sqrt{N}(\hat{\tau} - \tau) \le c^*_{1-\alpha}\} - \mathbb{P}^*\{\sqrt{N}(\hat{\tau}^* - \hat{\tau}) \le c^*_{1-\alpha}\}| + o_p(1) \\
&\quad \le \sup_{t \in \mathbb{R}} |\mathbb{P}\{\sqrt{N}(\hat{\tau} - \tau) \le t\} - \mathbb{P}^*\{\sqrt{N}(\hat{\tau}^* - \hat{\tau}) \le t\}| + o_p(1) \\
&\quad \xrightarrow{p} 0,
\end{aligned}
$$

and thus the conclusion (ii) of Theorem 6 follows.

SUPPLEMENTARY MATERIAL

Mengshan Xu and Taisuke Otsu (October 6, 2025): Supplement to "Isotonic propensity score matching," Econometric Theory Supplementary Material. To view, please visit: https://doi.org/10.1017/S0266466625100133.

REFERENCES

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74, 235–267.

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76, 1537–1557.

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

Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84, 781–807.

Adusumilli, K. (2020). Bootstrap inference for propensity score matching. Working Paper.

Ai, C., & Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71, 1795–1843.

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

Armstrong, T. B., & Kolesár, M. (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89, 1141–1177.

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

Babii, A., & Kumar, R. (2023). Isotonic regression discontinuity designs. *Journal of Econometrics*, 234(2), 371–393.

Balabdaoui, F., Durot, C., & Jankowski, H. (2019). Least squares estimation in the monotone single index model. *Bernoulli*, 25, 3276–3310.

Balabdaoui, F., Groeneboom, P., & Hendrickx, K. (2019). Score estimation in the monotone single index model. *Scandinavian Journal of Statistics*, 46, 517–544.

Balabdaoui, F., & Groeneboom, P. (2021). Profile least squares estimators in the monotone single index model. In A. Daouia and A. Ruiz-Gazen (Eds.), *Advances in contemporary statistics and econometrics* (pp. 3–22). Springer.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions: The theory and application of isotonic regression*. John Wiley & Sons.

Barlow, R., & Brunk, H. (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67, 140–147.

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

Bodory, H., Camponovo, L., Huber, M., & Lechner, M. (2016). A wild bootstrap algorithm for propensity score matching estimators. Working Paper.

Brunk, H. D. (1958). On the estimation of parameters restricted by inequalities. *Annals of Mathematical Statistics*, 29, 437–454.

Cattaneo, M. D., & Farrell, M. H. (2013). Optimal convergence rates, bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics*, 174, 127–143.

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

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

Chen, X., & Santos, A. (2018). Overidentification in regular models. *Econometrica*, 86, 1771–1817.

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107, 261–265.

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

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, 90, 1501–1535.

Chernozhukov, V., Newey, W. K., & Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90, 967–1027.

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

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

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

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84, 151–161.

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

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

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

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

Groeneboom, P., & Hendrickx, K. (2017). The nonparametric bootstrap for the current status model. *Electronic Journal of Statistics*, 11, 3446–3484.

Groeneboom, P., & Hendrickx, K. (2018). Current status linear regression. *Annals of Statistics*, 46, 1415–1444.

Groeneboom, P., & Jongbloed, G. (2014). *Nonparametric estimation under shape constraints*. Cambridge University Press.

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

https://doi.org/10.1017/S0266466625100133 Published online by Cambridge University Press

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315–331.
Han, A. K. (1987). Non-parametric analysis of a generalized regression model. *Journal of Econometrics*, 35, 303–316.
Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64, 605–654.
Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65, 261–294.
Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66, 1017–1098.
Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71, 1161–1189.
Huang, J. (2002). A note on estimating a partly linear model under monotonicity constraints. *Journal of Statistical Planning and Inference*, 107, 343–351.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86, 4–29.
Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78, 2021–2042.
Klein, R. W., & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61, 387–421.
Lin, Z., Ding, P., & Han, F. (2023). Estimation based on nearest neighbor matching: From density ratio to average treatment effect. *Econometrica*, 91, 2187–2217.
Liu, Y., & Qin, J. (2022). Tuning-parameter-free optimal propensity score matching approach for causal inference. Preprint, arXiv:2205.13200.
Liu, Y., & Qin, J. (2024). Tuning-parameter-free propensity score matching approach for causal inference under shape restriction. *Journal of Econometrics*, 244, 105829.
Matzkin, R. L. (1992). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica*, 60, 239–270.
Meyer, M. C. (2006). Consistency and power in tests with shape-restricted alternatives. *Journal of Statistical Planning and Inference*, 136, 3931–3947.
Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5, 99–135.
Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62, 1349–1382.
Newey, W. K., Hsieh, F. & Robins, J. M. (1998). Undersmoothing and bias corrected functional estimation. Working Paper 98-17, MIT.
Newey, W. K., Hsieh, F., & Robins, J. M. (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72, 947–962.
Otsu, T., & Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112, 1720–1732.
Qin, J., Yu, T., Li, P., Liu, H., & Chen, B. (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.
Rao, B. P. (1969) Estimation of a unimodal density. *Sankhyā, A*, 31, 23–36.
Rao, B. P. (1970). Estimation for distributions with monotone failure rate. *Annals of Mathematical Statistics*, 41, 507–519.
Ritov, J. M., & Ritov, Y. A. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16, 285–319.
Robins, J., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122–129.
Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56, 931–954.
Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84, 1024–1032.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.
Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39, 33–38.
Rothe, C. (2017). Robust confidence intervals for average treatment effects under limited overlap. *Econometrica*, 85, 645–660.
Rothe, C., & Firpo, S. (2019). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. *Econometric Theory*, 35, 1048–1087.
Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 1096–1120.
Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, 61, 123–137.
van de Geer, S. (2000). *Empirical processes in M-estimation*. Cambridge University Press.
van der Vaart, A. (1991). On differentiable functionals. *Annals of Statistics*, 19, 178–204.
van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.
Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Annals of Statistics*, 9, 443–448.
Xu, M. (2021). *Essays in semiparametric estimation and inference with monotonicity constraints* [Doctoral dissertation]. London School of Economics and Political Science.
Yu, K. (2014). On partial linear additive isotonic regression. *Journal of the Korean Statistical Society*, 43, 11–17.
Yuan, A., Yin, A., & Tan, M. T. (2021). Enhanced doubly robust procedure for causal inference. *Statistics in Biosciences*, 13, 454–478.