Submitted to *Econometrica*

ESTIMATION BASED ON NEAREST NEIGHBOR MATCHING: FROM DENSITY RATIO TO AVERAGE TREATMENT EFFECT

ZHEXIAO LIN
Department of Statistics, University of California, Berkeley

PENG DING
Department of Statistics, University of California, Berkeley

FANG HAN
Department of Statistics, University of Washington

Nearest neighbor (NN) matching is widely used in observational studies for causal effects. Abadie and Imbens (2006) provided the first large-sample analysis of NN matching. Their theory focuses on the case with the number of NNs, M fixed. We reveal something new out of their study and show that, once allowing M to diverge with the sample size, an intrinsic statistic in their analysis constitutes a consistent estimator of the density ratio with regard to covariates across the treated and control groups. Consequently, with a diverging M, the NN matching with Abadie and Imbens (2011)'s bias correction yields a doubly robust estimator of the average treatment effect and is semiparametrically efficient if the density functions are sufficiently smooth and the outcome model is consistently estimated. It can thus be viewed as a precursor of the double machine learning estimators.

KEYWORDS: graph-based statistics, stochastic geometry, double robustness, double machine learning, propensity score.

Zhexiao Lin: zhexiaolin@berkeley.edu
Peng Ding: pengdingpku@berkeley.edu
Fang Han: fanghan@uw.edu

1. INTRODUCTION

Matching methods (Greenwood, 1945, Chapin, 1947, Cochran and Rubin, 1973, Rubin, 2006, Rosenbaum, 2010) aim to balance observations from different groups through minimizing group differences in observed covariates. Such methods have proven their usefulness for causal inference in various disciplines, including economics (Imbens, 2004), epidemiology (Brookhart et al., 2006), political science (Ho et al., 2007, Sekhon, 2008) and sociology (Morgan and Harding, 2006).

Among all the matching methods, nearest neighbor (NN) matching (Rubin, 1973) is likely the most frequently used and easiest to implement approach. In the simplest treatment-control study, NN matching assigns each treatment (control) individual to M control (treatment) individuals with the smallest distance to it. In this regard, two questions arise. First, how do we select the number of matches, M? This is referred to in the literature as ratio matching, and is both important and delicate, well-known to be related to the bias-variance trade-off in nonparametic statistics (Smith, 1997, Rubin and Thomas, 2000, Imbens and Rubin, 2015). Second, how do we perform large-sample statistical inference for NN matching estimators? Such an analysis is usually nonstandard and technically challenging. Indeed, it was long-lacking in the literature until Abadie and Imbens (2006).

To answer the above two questions, a series of papers (Abadie and Imbens, 2006, 2008, 2011, 2012) established large-sample properties of M-NN matching for estimating the average treatment effect (ATE). These results are, however, only valid when, in ratio matching, M is fixed. The according message is then mixed. As a matter of fact, the ATE estimator based on M-NN matching with a fixed M is asymptotically biased and inefficient. While bias correction is now feasible to alleviate the first issue (Abadie and Imbens, 2011), the lack of efficiency seems fundamental.

This manuscript revisits the study of Abadie and Imbens (2006) from a new perspective, bridging M-NN matching to density ratio estimation (Nguyen et al., 2010, Sugiyama et al., 2012) as well as double robustness (Scharfstein et al., 1999, Bang and Robins, 2005). To this end, our analysis stresses, in ratio matching, the benefits of forcing M to diverge with the sample size n in order to achieve statistical efficiency. Our claim is thus aligned with

observations in the random graph-based inference literature (Wald and Wolfowitz, 1940, Friedman and Rafsky, 1979, Henze, 1988, Liu and Singh, 1993, Henze and Penrose, 1999, Berrett et al., 2019, Bhattacharya, 2019, Shi et al., 2022a,b, Lin and Han, 2023).

The contributions of this manuscript are two-fold. First, we show that a statistic that plays a pivotal role in the analysis of Abadie and Imbens (2006), $K_M(x)$ (Abadie and Imbens (2006, Page 240); to be defined in (2.2) of Section 2), which measures the number of matched times of the covariate value x, actually gives rise to a consistent density ratio estimator in the two-sample setting. Furthermore, from the angle of density ratio estimation, this NN matching-based estimator is to our knowledge the first one that simultaneously satisfies being conceptually one-step, computationally efficient, and statistically rate-optimal. This estimator itself is thus an appealing alternative to existing density ratio estimators.

Getting back to the original ATE estimation problem, our second contribution is to use the above insights to bridge the bias-corrected matching estimator (Abadie and Imbens, 2011), doubly robust estimators (Scharfstein et al., 1999, Bang and Robins, 2005, Farrell, 2015), and double machine learning estimators (Chernozhukov et al., 2018). In fact, Abadie and Imbens (2011)'s bias-corrected estimator can be formulated as

$$\widehat{\tau}_M^{\mathrm{bc}} = \widehat{\tau}^{\mathrm{reg}} + \frac{1}{n}\Big[\sum_{i=1,D_i=1}^{n}\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i - \sum_{i=1,D_i=0}^{n}\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i\Big]$$

(see Lemma 3.1, with notation introduced in Section 3 and $K_M(i)$ representing the number of times the unit i is matched), and then $1 + K_M(i)/M$ converges to the inverses of the propensity scores $1-e(X_i)$ and $e(X_i)$ for units with $D_i = 0$ and 1, respectively. One could then leverage the general double robustness and double machine learning theory to validate the following two properties of $\widehat{\tau}_M^{\mathrm{bc}}$.

(1) *Consistency:* $\widehat{\tau}_M^{\mathrm{bc}}$ converges in probability to the population ATE, if either the density (propensity score) functions satisfy certain conditions or the outcome (regression) model is consistently estimated, with $M \log n/n \to 0$ and $M \to \infty$ as $n \to \infty$.

(2) *Semiparametric efficiency:* $\widehat{\tau}_M^{\mathrm{bc}}$ is an asymptotically Normal estimator of τ with the asymptotic variance attaining the semiparametric efficiency lower bound (Hahn, 1998), if the density functions are sufficiently smooth, the outcome model is consis-

tently estimated, and M scales with n at a proper rate. Furthermore, a simple consistent estimator of the asymptotic variance is available.

Although Abadie and Imbens (2006, Theorem 5) hints at the necessity of allowing M to diverge for gaining efficiency, we provide rigorous theory for their conjecture. Our results thus complement those made in Abadie and Imbens (2006, 2011) and provide additional theoretical justifications for practitioners to use NN matching for inferring the ATE.

Technically, our analysis hinges on a diverging M that grows with n. In contrast, existing results on NN matching for causal effects (Abadie and Imbens, 2006, 2008, 2011, 2012) all focused on a fixed M. Instead, we take a different route to establish nonasymptotic moment bounds on $K_M(x)$ with more flexibility in specifying the rate of M with respect to n (see Lin and Han (2023) for a similar idea in analyzing rank-based statistics).

Paper organization. The rest of this manuscript proceeds as follows. Section 2 gives a brief overview of the NN matching-based density ratio estimator. Section 3 revisits Abadie and Imbens (2011)'s bias-corrected NN matching-based estimator of the ATE, $\widehat{\tau}_M^{\mathrm{bc}}$. Section 4 elaborates on the double robustness and semiparametric efficiency of $\widehat{\tau}_M^{\mathrm{bc}}$ as well as its double machine learning version. Section 5 presents simulation studies to complement the theory. Section 6 includes some final remarks. We relegate technical details to the appendix as well as an online appendix. Appendices A and B introduce the algorithms and theory for the NN matching-based density ratio estimator. Appendix C and the online appendix present the proofs of results in the paper and in the appendix, respectively.

Notation. For any integers $n, d \geqslant 1$, let $[\![n]\!] = \{1, 2, \ldots, n\}$, $n!$ be the factorial of n, and $\mathbb{R}^d$ be the d-dimensional real space. A set consisting of distinct elements $x_1, \ldots, x_n$ is written as either $\{x_1, \ldots, x_n\}$ or $\{x_i\}_{i=1}^n$, and its cardinality is written by $|\{x_i\}_{i=1}^n|$. The corresponding sequence is denoted by $[x_1, \ldots, x_n]$ or $[x_i]_{i=1}^n$. Let $\mathbb{1}(\cdot)$ denote the indicator function. For any $a, b \in \mathbb{R}$, write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We use $\xrightarrow{\mathsf{d}}$ and $\xrightarrow{\mathsf{p}}$ to denote convergence in distribution and in probability, respectively. For any sequence of random variables $\{X_n\}$, write $X_n = o_{\mathrm{P}}(1)$ if $X_n \xrightarrow{\mathsf{p}} 0$ and $X_n = O_{\mathrm{P}}(1)$ if X_n is bounded in probability. Let P_Z represent the law of a random variable Z.

2. DENSITY RATIO ESTIMATION VIA NN-MATCHING

Consider two general random vectors X, Z in $\mathbb{R}^d$ that are defined on the same probability space, with d to be a fixed positive integer. Let ν_0 and ν_1 represent the probability measures of X and Z, respectively. Assume ν_0 and ν_1 are absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^d$ equipped with the Euclidean norm $\|\cdot\|$; denote the corresponding densities (Radon–Nikodym derivatives) by f_0 and f_1. Assume further that ν_1 is absolutely continuous with respect to ν_0 and write the corresponding density ratio, f_1/f_0, as r; set $0/0 = 0$ by default.

Assume $X_1, \ldots, X_{N_0}$ are N_0 independent copies of X, $Z_1, \ldots, Z_{N_1}$ are N_1 independent copies of Z, and $[X_i]_{i=1}^{N_0}$ and $[Z_j]_{j=1}^{N_1}$ are mutually independent. The problem of estimating the density ratio r based on $\{X_1, \ldots, X_{N_0}, Z_1, \ldots, Z_{N_1}\}$ is fundamental in economics (Cunningham, 2021), information theory (Cover and Thomas, 2006), machine learning (Sugiyama et al., 2012), statistics (Imbens and Rubin, 2015), and other fields.

In density ratio estimation, NN-based estimators are advocated before due to its computational efficiency; cf. Lima et al. (2008), Póczos and Schneider (2011), Kremer et al. (2015), Noshad et al. (2017), Berrett et al. (2019), Zhao and Lai (2020), among many others. Based on Abadie and Imbens (2006, 2008, 2011, 2012)'s NN matching framework, we propose a new density ratio estimator based on NN matching. To this end, some necessary notation is introduced first.

DEFINITION 2.1—NN matching: For any $x, z \in \mathbb{R}^d$ and $M \in [\![N_0]\!]$,

(i) let $\mathcal{X}_{(M)}(\cdot) : \mathbb{R}^d \to \{X_i\}_{i=1}^{N_0}$ be the mapping that returns the value of the input z's M-th NN in $\{X_i\}_{i=1}^{N_0}$, i.e., the value of $x \in \{X_i\}_{i=1}^{N_0}$ such that

$$\sum_{i=1}^{N_0} \mathbb{1}\Big(\|X_i - z\| \leqslant \|x - z\|\Big) = M; \tag{2.1}$$

(ii) let $K_M(\cdot) : \mathbb{R}^d \to \{0\} \bigcup [\![N_1]\!]$ be the mapping that returns the number of matched times of x, i.e.,

$$K_M(x) = K_M\big(x; \{X_i\}_{i=1}^{N_0}, \{Z_j\}_{j=1}^{N_1}\big) = \sum_{j=1}^{N_1} \mathbb{1}\Big(\|x - Z_j\| \leqslant \|\mathcal{X}_{(M)}(Z_j) - Z_j\|\Big); \tag{2.2}$$

(iii) let $A_M(\cdot) : \mathbb{R}^d \to \mathcal{B}(\mathbb{R}^d)$ be the corresponding mapping from $\mathbb{R}^d$ to the class of all Borel sets in $\mathbb{R}^d$ so that

$$A_M(x) = A_M\big(x; \{X_i\}_{i=1}^{N_0}\big) = \Big\{z \in \mathbb{R}^d : \|x - z\| \leqslant \|\mathcal{X}_{(M)}(z) - z\|\Big\} \tag{2.3}$$

returns the catchment area of x in the setting of (ii);

Because ν_0 is absolutely continuous with respect to the Lebesgue measure, (2.1) has a unique solution. Abadie and Imbens (2006, Pages 240 and 260) introduced the terms $K_M(\cdot)$ and $A_M(\cdot)$ to analyze the asymptotic behavior of their NN matching-based ATE estimator. We also adopt their terminology "catchment area" in Definition 2.1(iii). Proposition 2.1 below formally links $K_M(\cdot)$ to $A_M(\cdot)$. It was established in the proof of Abadie and Imbens (2006, Lemma 3), and is stated here to aid understanding.

PROPOSITION 2.1: *For any $x \in \mathbb{R}^d$, we have $K_M(x) = \sum_{j=1}^{N_1} \mathbb{1}\Big(Z_j \in A_M(x)\Big)$.*

REMARK 2.1—Relation between $A_M(X_i)$'s and Voronoi tessellation when $M = 1$: We can verify that, due to the absolute continuity of ν_0, $[A_1(X_i)]_{i=1}^{N_0}$ are almost surely disjoint except for a Lebesgue measure zero area, and partition $\mathbb{R}^d$ into N_0 polygons. Furthermore, we can also verify that $\{A_1(X_i)\}_{i=1}^{N_0}$ are exactly the Voronoi tessellation defined in Voronoi (1908), which plays a vital role in stochastic and computational geometry. In this case, each element $A_1(X_i)$ is a Voronoi cell from the definition of (2.3).

With these notation and concepts, we are now ready to introduce the following density ratio estimator based on NN matching.

DEFINITION 2.2—NN matching-based density ratio estimator: For any $M \in [\![N_0]\!]$ and $x \in \mathbb{R}^d$, we define the following estimator for $r(x)$:

$$\widehat{r}_M(x) = \widehat{r}_M\Big(x; \{X_i\}_{i=1}^{N_0}, \{Z_j\}_{j=1}^{N_1}\Big) = \frac{N_0}{N_1}\frac{K_M(x)}{M}. \tag{2.4}$$

The estimator $\widehat{r}_M(\cdot)$ is by construction a one-step estimator, and satisfies the following two properties simultaneously.

(P1) Computationally *of low complexity*: it is of a sub-quadratic (and nearly linear when M is small) time complexity via a careful algorithmic formulation based on k-d trees (see Algorithms 1–2 and Theorem A.1 in Appendix A), and thus in many scientific applications is computationally more attractive than its optimization-based alternatives (Lima et al., 2008, Kremer et al., 2015, Borgeaud et al., 2021).

(P2) Statistically *rate-optimal*: it is information-theoretically efficient in terms of achieving an upper bound of estimation accuracy that matches the corresponding minimax lower bound over a class of Lipschitz density functions (see Appendix B).

3. REVISITING THE BIAS-CORRECTED MATCHING ESTIMATOR OF THE ATE

This section studies the bias-corrected NN matching-based estimator of the ATE, proposed in Abadie and Imbens (2011) to correct the asymptotic bias of the original matching-based estimator derived by Abadie and Imbens (2006). To this end, we leverage the new insights in Section 2 as well as the technical results in Appendices A–B, and bridge the study to both the classic double robustness and the modern double machine learning frameworks.

We first review the setup for the NN matching-based estimator and its bias-corrected version. Following Abadie and Imbens (2006), let $[(X_i, D_i, Y_i)]_{i=1}^n$ be n independent copies of (X, D, Y), where $D \in \{0, 1\}$ is a binary variable, let $X \in \mathbb{R}^d$ represent the individual covariates, assumed to be absolute continuous admitting a density f_X, and let $Y \in \mathbb{R}$ stand for the outcome variable.

For each unit $i \in [\![n]\!]$, we observe $D_i = 1$ if in the treated group and $D_i = 0$ if in the control group. Let $n_0 = \sum_{i=1}^n (1 - D_i)$ and $n_1 = \sum_{i=1}^n D_i$ be the numbers of control and treated units, respectively. Under the potential outcomes framework (Rubin, 1974), the unit i has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, but we observe only one of them: $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. The goal is to estimate the population ATE, $\tau = \mathrm{E}[Y_i(1) - Y_i(0)]$, based on the observations $\{(X_i, D_i, Y_i)\}_{i=1}^n$. To estimate ATE, we consider its empirical counterpart $\widehat{\tau}_M = n^{-1} \sum_{i=1}^n [\widehat{Y}_i(1) - \widehat{Y}_i(0)]$, where $\widehat{Y}_i(0)$ and $\widehat{Y}_i(1)$ are the imputed outcomes of $Y_i(0)$ and $Y_i(1)$. Following Abadie and Imbens (2006), we focus on the matching-

based estimator by imputing missing potential outcomes as

$$\widehat{Y}_i(0) = \begin{cases} Y_i, & \text{if } D_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j, & \text{if } D_i = 1 \end{cases} \quad \text{and} \quad \widehat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j, & \text{if } D_i = 0, \\ Y_i, & \text{if } D_i = 1. \end{cases}$$

Here $\mathcal{J}_M(i)$ represents the index set of M-NNs of X_i in $\{X_j : D_j = 1 - D_i\}_{j=1}^n$, i.e., the set of all indices $j \in [\![n]\!]$ such that $D_j = 1 - D_i$ and $\sum_{\ell=1, D_\ell = 1-D_i}^n \mathbb{1}(\|X_\ell - X_i\| \leqslant \|X_j - X_i\|) \leqslant M$. With a slight abuse of notation, let $K_M(i)$ represent the number of matched times for unit i, i.e., $K_M(i) = \sum_{j=1, D_j = 1-D_i}^n \mathbb{1}(i \in \mathcal{J}_M(j))$. We can then rewrite the above matching-based estimator as

$$\widehat{\tau}_M = \frac{1}{n}\Big[\sum_{i=1, D_i=1}^n \Big(1 + \frac{K_M(i)}{M}\Big) Y_i - \sum_{i=1, D_i=0}^n \Big(1 + \frac{K_M(i)}{M}\Big) Y_i\Big]. \tag{3.1}$$

However, when $d > 1$, the bias of $\widehat{\tau}_M$ is asymptotically non-negligible (Abadie and Imbens, 2006). To fix this, Abadie and Imbens (2011) proposed the following bias-corrected version for $\widehat{\tau}_M$. In detail, let $\widehat{\mu}_0(x)$ and $\widehat{\mu}_1(x)$ be mappings from $\mathbb{R}^d$ to $\mathbb{R}$ that estimate the conditional means of the outcomes $\mu_0(x) = \mathrm{E}[Y \,|\, X = x, D = 0]$ and $\mu_1(x) = \mathrm{E}[Y \,|\, X = x, D = 1]$, respectively, with the corresponding residuals $\widehat{R}_i = Y_i - \widehat{\mu}_{D_i}(X_i),\ i \in [\![n]\!]$. Define the estimator based on outcome regressions as $\widehat{\tau}^{\mathrm{reg}} = n^{-1}\sum_{i=1}^n [\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)]$. Consider the bias-corrected matching-based estimator in Abadie and Imbens (2011):

$$\widehat{\tau}_M^{\mathrm{bc}} = \frac{1}{n}\sum_{i=1}^n \Big[\widehat{Y}_i^{\mathrm{bc}}(1) - \widehat{Y}_i^{\mathrm{bc}}(0)\Big], \tag{3.2}$$

with

$$\widehat{Y}_i^{\mathrm{bc}}(0) = \begin{cases} Y_i, & \text{if } D_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} (Y_j + \widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_j)), & \text{if } D_i = 1, \end{cases}$$

and

$$\widehat{Y}_i^{\mathrm{bc}}(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} (Y_j + \widehat{\mu}_1(X_i) - \widehat{\mu}_1(X_j)), & \text{if } D_i = 0, \\ Y_i, & \text{if } D_i = 1. \end{cases}$$

Lemma 3.1 below shows an equivalent form of $\widehat{\tau}_M^{\mathrm{bc}}$.

LEMMA 3.1: *The bias-corrected matching-based estimator in (3.2) can be rewritten in terms of $\widehat{\tau}^{\mathrm{reg}}$ and the residuals $\widehat{R}_i$'s as:*

$$\widehat{\tau}_M^{\mathrm{bc}} = \widehat{\tau}^{\mathrm{reg}} + \frac{1}{n}\Big[\sum_{i=1,D_i=1}^{n}\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i - \sum_{i=1,D_i=0}^{n}\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i\Big]. \tag{3.3}$$

Otsu and Rai (2017) derived another linear form of $\widehat{\tau}_M^{\mathrm{bc}}$ to motivate a bootstrap procedure for variance estimation. The form in (3.3) is related to doubly robust estimators reviewed shortly. In detail, we first have some outcome models and residuals defined in the same way as above, and then let $\widehat{e}(x):\mathbb{R}^d\to\mathbb{R}$ be a generic estimator of the propensity score (Rosenbaum and Rubin, 1983), $e(x)=\mathrm{P}(D=1\,|\,X=x)$. The doubly robust estimator in Scharfstein et al. (1999) and Bang and Robins (2005) could then be formulated as

$$\widehat{\tau}^{\mathrm{dr}} = \widehat{\tau}^{\mathrm{reg}} + \frac{1}{n}\Big[\sum_{i=1,D_i=1}^{n}\frac{1}{\widehat{e}(X_i)}\widehat{R}_i - \sum_{i=1,D_i=0}^{n}\frac{1}{1-\widehat{e}(X_i)}\widehat{R}_i\Big]. \tag{3.4}$$

Conditional on $(D_1,\ldots,D_n)$, $[X_i:D_i=\omega]_{i=1}^n$ are n_ω i.i.d. random variables sampled from the distribution of $X\,|\,D=\omega$, and the two groups of sample points, $[X_i:D_i=0]_{i=1}^n$ and $[X_i:D_i=1]_{i=1}^n$, are mutually independent. Let $f_{X\,|\,D=\omega}$ denote the density of $X\,|\,D=\omega$. From the construction of $K_M(i)$ and results in Appendix B, once allowing M to diverge to infinity, conditional on $(D_1,\ldots,D_n)$, $n_0/n_1\cdot K_M(i)/M$ and $n_1/n_0\cdot K_M(i)/M$ are consistent estimators of $f_{X\,|\,D=1}(X_i)/f_{X\,|\,D=0}(X_i)$ and $f_{X\,|\,D=0}(X_i)/f_{X\,|\,D=1}(X_i)$ for units with $D_i=0$ and $D_i=1$, respectively. Because n_1/n_0 converges almost surely to $\mathrm{P}(D=1)/\mathrm{P}(D=0)$ by the law of large numbers, the statistic $1+K_M(i)/M$ is then a consistent estimator of $1/(1-e(X_i))$ and $1/e(X_i)$ for units with $D_i=0$ and $D_i=1$, respectively. Thus, in view of (3.4), the bias-corrected matching-based estimator $\widehat{\tau}_M^{\mathrm{bc}}$ in (3.3) is actually a doubly robust estimator of τ, and accordingly, should also enjoy all the corresponding desirable properties. This novel insight into $\widehat{\tau}_M^{\mathrm{bc}}$ allows us to establish its asymptotic properties with a diverging M.

4. ASYMPTOTIC ANALYSIS WITH DIVERGING M

The theory for matching with a diverging M has been an important gap in the literature. With a univariate covariate, Abadie and Imbens (2006) provided a heuristic argument about the additional efficiency gain for $\widehat{\tau}_M$ with larger M. With a general covariate, Abadie and Imbens (2011) used simulation to evaluate the finite-sample properties of $\widehat{\tau}_M^{\mathrm{bc}}$ and highlighted the importance of bias correction with large M. Nevertheless, existing theoretical results for NN matching estimators all focused on fixed M (Abadie and Imbens, 2006, 2008, 2011, 2016, Kallus, 2020, Armstrong and Kolesár, 2021, Ferman, 2021). In this section, we will present the corresponding theory with a diverging M and also make connections between $\widehat{\tau}_M^{\mathrm{bc}}$ and double robustness/double machine learning estimators.

4.1. *The original matching-based estimator*

We first analyze the original bias-corrected matching-based estimator $\widehat{\tau}_M^{\mathrm{bc}}$. Let $U_\omega = Y(\omega) - \mu_\omega(X)$ for $\omega \in \{0,1\}$ and $\mathbb{X}$ be the support of X. Let $\|\cdot\|_\infty$ denote the L_∞ norm of a function.

We need following assumptions to prove the consistency of $\widehat{\tau}_M^{\mathrm{bc}}$.

ASSUMPTION 4.1: *(i) For almost all $x \in \mathbb{X}$, D is independent of $(Y(0), Y(1))$ conditional on $X = x$, and there exist some constants $\eta > 0$ such that $\eta < \mathrm{P}(D = 1 \,|\, X = x) < 1 - \eta$.*

(ii) $[(X_i, D_i, Y_i)]_{i=1}^n$ *are i.i.d. following the joint distribution of* (X, D, Y).

(iii) $\mathrm{E}[U_\omega^2 \,|\, X = x]$ *is uniformly bounded for almost all* $x \in \mathbb{X}$ *and* $\omega \in \{0,1\}$.

(iv) $\mathrm{E}[\mu_\omega^2(X)]$ *is bounded for* $\omega \in \{0,1\}$.

ASSUMPTION 4.2: *For $\omega \in \{0,1\}$, there exists a deterministic function $\bar{\mu}_\omega(\cdot) : \mathbb{R}^d \to \mathbb{R}$ such that $\mathrm{E}[\bar{\mu}_\omega^2(X)]$ is bounded and the estimator $\widehat{\mu}_\omega(x)$ satisfies $\|\widehat{\mu}_\omega - \bar{\mu}_\omega\|_\infty = o_{\mathrm{P}}(1)$.*

ASSUMPTION 4.3: *For $\omega \in \{0,1\}$, the estimator $\widehat{\mu}_\omega(x)$ satisfies $\|\widehat{\mu}_\omega - \mu_\omega\|_\infty = o_{\mathrm{P}}(1)$.*

Assumption 4.1(i) is the unconfoundedness and overlap assumptions, and is often referred to as the strong ignorability condition (Rosenbaum and Rubin, 1983). Assumption

4.2 allows for outcome model misspecification; for example, if $\widehat{\mu}_\omega = \bar{\mu}_\omega = 0$, $\widehat{\tau}_M^{\mathrm{bc}}$ then reduces to $\widehat{\tau}_M$. Assumption 4.3 assumes that the outcome models are consistently estimated.

We need the following assumptions to prove the efficiency of $\widehat{\tau}_M^{\mathrm{bc}}$.

ASSUMPTION 4.4: *(i)* $\mathrm{E}[U_\omega^2 \,|\, X = x]$ *is uniformly bounded away from zero for almost all* $x \in \mathbb{X}$ *and* $\omega \in \{0,1\}$.

(ii) There exist some constants $\kappa > 0$ *such that* $\mathrm{E}[|U_\omega|^{2+\kappa} \,|\, X = x]$ *is uniformly bounded for almost all* $x \in \mathbb{X}$ *and* $\omega \in \{0,1\}$.

(iii) $\max_{t \in \Lambda_{\lfloor d/2 \rfloor + 1}} \|\partial^t \mu_\omega\|_\infty$ *is bounded, where for any positive integer* k, Λ_k *is the set of all* d*-dimensional vectors of nonnegative integers* $t = (t_1, \ldots, t_d)$ *such that* $\sum_{i=1}^d t_i = k$ *and* $\lfloor \cdot \rfloor$ *stands for the floor function.*

ASSUMPTION 4.5: *For* $\omega \in \{0,1\}$, *the estimator* $\widehat{\mu}_\omega(x)$ *satisfies*

$$\max_{t \in \Lambda_{\lfloor d/2 \rfloor + 1}} \|\partial^t \widehat{\mu}_\omega\|_\infty = O_{\mathrm{P}}(1) \quad \text{and} \quad \max_{t \in \Lambda_\ell} \|\partial^t \widehat{\mu}_\omega - \partial^t \mu_\omega\|_\infty = O_{\mathrm{P}}(n^{-\gamma_\ell}) \ \text{ for all } \ \ell \in [\![\lfloor d/2 \rfloor]\!],$$

with some constants γ_ℓ*'s satisfying* $\gamma_\ell > \frac{1}{2} - \frac{\ell}{d}$ *for* $\ell = 1, 2, \ldots, \lfloor d/2 \rfloor$.

Assumption 4.4 is comparable to Assumption A.4 and the assumptions in Abadie and Imbens (2011, Theorem 2). Compared with the assumptions in Abadie and Imbens (2011, Theorem 2), Assumption 4.4(iii) is weaker in the sense that it only requires a finite order of smoothness. Assumption 4.5 again assumes the approximation accuracy of the outcome models, with lower convergence rates required for higher order derivatives of the outcome models. Under some smoothness conditions on the outcome model as made in Abadie and Imbens (2011), Assumption 4.5 holds using power series approximation (Abadie and Imbens, 2011, Lemma A.1). Lastly, compared with Chernozhukov et al. (2018), we need approximation accuracy concerning derivatives of the outcome model estimator, which is not required in Chernozhukov et al. (2018); see Section 4.2 for more discussions.

Theorem 4.1 below presents the double robustness and semiparametric efficiency properties of $\widehat{\tau}_M^{\mathrm{bc}}$. Recall the semiparametric efficiency lower bound for estimating ATE (see

Hahn (1998)):

$$\sigma^2 = \mathrm{E}\Big[\mu_1(X) - \mu_0(X) + \frac{D(Y-\mu_1(X))}{e(X)} - \frac{(1-D)(Y-\mu_0(X))}{1-e(X)} - \tau\Big]^2,$$

and introduce an estimator for σ^2 based on NN matching:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) + (2D_i-1)\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i - \widehat{\tau}_M^{\mathrm{bc}}\Big]^2.$$

THEOREM 4.1: *(i) (Double robustness of $\widehat{\tau}_M^{\mathrm{bc}}$) On one hand, if the distribution of (X,D,Y) satisfies Assumptions 4.1, 4.2, either $(\mathrm{P}_{X\,|\,D=0}, \mathrm{P}_{X\,|\,D=1})$ or $(\mathrm{P}_{X\,|\,D=1}, \mathrm{P}_{X\,|\,D=0})$ satisfies Assumption B.1 in the appendix, and $M\log n/n \to 0$ and $M \to \infty$ as $n \to \infty$, then $\widehat{\tau}_M^{\mathrm{bc}} - \tau \xrightarrow{\mathsf{p}} 0$.*

On the other hand, if the distribution of (X,D,Y) satisfies Assumptions 4.1 and 4.3, then $\widehat{\tau}_M^{\mathrm{bc}} - \tau \xrightarrow{\mathsf{p}} 0$.

(ii) (Semiparametric efficiency of $\widehat{\tau}_M^{\mathrm{bc}}$) Assume the distribution of (X,D,Y) satisfies Assumptions 4.1, 4.4, 4.5 and either $(\mathrm{P}_{X\,|\,D=0}, \mathrm{P}_{X\,|\,D=1})$ or $(\mathrm{P}_{X\,|\,D=1}, \mathrm{P}_{X\,|\,D=0})$ satisfies Assumption B.1 in the appendix. Define

$$\gamma = \Big\{\min_{\ell\in[\![\lfloor d/2\rfloor]\!]}\Big[1-\Big(\frac{1}{2}-\gamma_\ell\Big)\frac{d}{\ell}\Big]\Big\} \wedge \Big[1-\frac{1}{2}\frac{d}{\lfloor d/2\rfloor+1}\Big],$$

recalling that γ_ℓ's were introduced in Assumption 4.5. Then, if $M \to \infty$ and $M/n^\gamma \to 0$ as $n \to \infty$, we have $\sqrt{n}(\widehat{\tau}_M^{\mathrm{bc}} - \tau) \xrightarrow{\mathsf{d}} N(0,\sigma^2)$.

If in addition Assumption 4.3 holds, then $\widehat{\sigma}^2 \xrightarrow{\mathsf{p}} \sigma^2$.

REMARK 4.1: To be in line with the double robustness terminology, we can call Assumptions B.1 used in Theorem 4.1 the "density (or propensity) model assumptions" and Assumptions 4.3–4.5 the "outcome (or regression) model assumptions".

REMARK 4.2: The first part of Theorem 4.1(i) requires $M \to \infty$ for achieving the consistency of the propensity score model. When M is fixed and the outcome model is misspecified, $\widehat{\tau}_M^{\mathrm{bc}}$ is no longer doubly robust in the sense of Theorem 4.1. However, it does not imply that $\widehat{\tau}_M^{\mathrm{bc}}$ is inconsistent for estimating τ. In fact, Abadie and Imbens (2006, Theorem 3) showed that $\widehat{\tau}_M^{\mathrm{bc}}$ with a fixed M is still consistent even if we choose $\widehat{\mu}_w = 0$ for $w = 0, 1$.

One can further show that $\widehat{\tau}_M^{\text{bc}}$ with a fixed M is consistent as long as the outcome models are smooth but misspecified since the matching discrepancy then converges to zero.

REMARK 4.3: Theorem 4.1 has implications for practical data analysis. We discuss two. First, it highlights the importance of allowing M to diverge in asymptotic analysis. Nevertheless, it is a challenging problem to choose M in finite samples. We use simulation to illustrate the choice of M. Second, it gives an alternative variance estimator $\widehat{\sigma}^2$ for the bias-corrected matching estimator when M diverges. Abadie and Imbens (2006) gave another variance estimator for fixed M. While it is challenging to compare the two variance estimators in theory, we use simulation to compare them in finite samples. See Section 5 for the details of simulation.

If $d = 1$ and we pick $\widehat{\mu}_\omega = 0$ for $\omega \in \{0, 1\}$, then Assumption 4.5 automatically holds and the bias-corrected estimator $\widehat{\tau}_M^{\text{bc}}$ reduces to the original estimator $\widehat{\tau}_M$ studied in Abadie and Imbens (2006). Theorem 4.1(ii) then directly implies the following corollary that corresponds to Abadie and Imbens (2006, Corollary 1) with one key difference that M goes to infinity here.

COROLLARY 4.1—Semiparametric efficiency of $\widehat{\tau}_M$ when $d = 1$: *Assume $d = 1$, the distribution of (X, D, Y) satisfies Assumptions 4.1, 4.4, and either $(\mathrm{P}_{X \mid D=0}, \mathrm{P}_{X \mid D=1})$ or $(\mathrm{P}_{X \mid D=1}, \mathrm{P}_{X \mid D=0})$ satisfies Assumption B.1 in the appendix. If $M \to \infty$ and $M/n^{\frac{1}{2}} \to 0$ as $n \to \infty$, then $\sqrt{n}(\widehat{\tau}_M - \tau) \xrightarrow{\text{d}} N(0, \sigma^2)$.*

REMARK 4.4: By picking $\widehat{\mu}_\omega = 0$ for $\widehat{\tau}_M$, Assumption 4.3 is in general no longer satisfied. Accordingly, in Corollary 4.1, $\widehat{\sigma}^2$ may not be a consistent estimator of σ^2 without additional assumptions. However, by decomposing σ^2 into the form of Theorem 1 in Hahn (1998), one could still estimate σ^2 via a similar and direct way as what is outlined in Section 4 in Abadie and Imbens (2006). We do not pursue this track in detail here as the case of $d = 1$ without Assumption 4.3 is beyond the main scope of this manuscript.

4.2. *A double machine learning version of the matching*

Assumptions 4.4 and 4.5 enforce arguably strong requirements on the smoothness of the outcome model. To weaken such assumptions, Chernozhukov et al. (2018) introduced the idea of double machine learning. In this section, we consider the option to combine NN matching with double machine learning.

Assume n is divisible by K for simplicity. Let $[I_k]_{k=1}^K$ be a K-fold random partition of $[\![n]\!]$, with each of size equal to $n' = n/K$. For each $k \in [\![K]\!]$ and $\omega \in \{0,1\}$, construct $\widehat{\mu}_{\omega,k}(\cdot)$ using data $[(X_j, D_j, Y_j)]_{j=1, j\notin I_k}^n$, and let $K_{M,k}(i)$ be the number of matched times for unit i by adding (X_i, D_i, Y_i) into $[(X_j, D_j, Y_j)]_{j=1, j\notin I_k}^n$. Define

$$\breve{\tau}_{M,k}^{\mathrm{bc}} = \frac{1}{n'} \sum_{i=1, i\in I_k}^{n} \Big[\widehat{\mu}_{1,k}(X_i) - \widehat{\mu}_{0,k}(X_i)\Big] + \frac{1}{n'}\Big[\sum_{i=1, i\in I_k, D_i=1}^{n} \Big(1 + \frac{K_{M,k}(i)}{M}\Big) \Big(Y_i - \widehat{\mu}_{1,k}(X_i)\Big) - \sum_{i=1, i\in I_k, D_i=0}^{n} \Big(1 + \frac{K_{M,k}(i)}{M}\Big)\Big(Y_i - \widehat{\mu}_{0,k}(X_i)\Big)\Big]$$

for $k = 1, \ldots, K$, and then define $\widetilde{\tau}_{M,K}^{\mathrm{bc}} = K^{-1} \sum_{k=1}^K \breve{\tau}_{M,k}^{\mathrm{bc}}$. We can use the same variance estimator $\widehat{\sigma}^2$ for $\widetilde{\tau}_{M,K}^{\mathrm{bc}}$.

To analyze $\widetilde{\tau}_{M,K}^{\mathrm{bc}}$ instead of $\widehat{\tau}_M^{\mathrm{bc}}$, we replace Assumptions 4.4 and 4.5 with the following two assumptions.

ASSUMPTION 4.6: *(i)* $\mathrm{E}[U_\omega^2]$ *is bounded away from zero for* $\omega \in \{0,1\}$.
(ii) There exist some constants $\kappa > 0$ *such that* $\mathrm{E}[|Y|^{2+\kappa}]$ *is bounded.*

ASSUMPTION 4.7: *For* $\omega \in \{0,1\}$, *the estimator* $\widehat{\mu}_\omega(x)$ *satisfies* $\|\widehat{\mu}_\omega - \mu_\omega\|_\infty = o_{\mathrm{P}}(n^{-d/(4+2d)})$.

REMARK 4.5: Assumption 4.6 corresponds to Assumption 5.1 in Chernozhukov et al. (2018), and is similar to Assumption 4 in Abadie and Imbens (2006). Assumption 4.7 assumes approximation accuracy of the outcome model under the L_∞ norm. Abadie and Imbens (2011) used the power series approximation (Newey, 1997) to estimate the outcome model, which under some classic nonparametric statistics assumptions automatically satis-

fies Assumption 4.7 (cf. Lemma A.1 in Abadie and Imbens (2011)). The same conclusion also applies to spline and wavelet regression estimators; cf. Chen and Christensen (2015).

REMARK 4.6: Assumption 4.7 assumes an approximation rate under L_∞ norm. This is different from the L_2 norm put in Chernozhukov et al. (2018, Assumption 5.1), but can be handled with some trivial modifications to the proof of Chernozhukov et al. (2018, Theorem 5.1) since one can replace the Cauchy–Schwarz inequality by the L_1-L_∞ Hölder's inequality. An L_1-norm bound on $K_M(i)/M$, to be established in Theorem B.4 in the appendix, can then be applied directly.

THEOREM 4.2: *(i) (Double robustness of $\widetilde{\tau}^{\mathrm{bc}}_{M,K}$) On one hand, if the distribution of (X, D, Y) satisfies Assumptions 4.1, 4.2, either $(\mathrm{P}_{X\,|\,D=0}, \mathrm{P}_{X\,|\,D=1})$ or $(\mathrm{P}_{X\,|\,D=1}, \mathrm{P}_{X\,|\,D=0})$ satisfies Assumption B.1 in the appendix, and $M \log n/n \to 0$ and $M \to \infty$ as $n \to \infty$, then $\widetilde{\tau}^{\mathrm{bc}}_{M,K} - \tau \xrightarrow{\mathsf{p}} 0$.*

On the other hand, if the distribution of (X, D, Y) satisfies Assumptions 4.1 and 4.3, then $\widetilde{\tau}^{\mathrm{bc}}_{M,K} - \tau \xrightarrow{\mathsf{p}} 0$.

(ii) (Semiparametric efficiency of $\widetilde{\tau}^{\mathrm{bc}}_{M,K}$) Assume the distribution of (X, D, Y) satisfies Assumptions 4.1, 4.6, 4.7 and either $(\mathrm{P}_{X\,|\,D=0}, \mathrm{P}_{X\,|\,D=1})$ or $(\mathrm{P}_{X\,|\,D=1}, \mathrm{P}_{X\,|\,D=0})$ satisfies Assumption B.3 in the appendix. Then if we pick $M = \alpha n^{\frac{2}{2+d}}$ for some constant $\alpha > 0$, then $\sqrt{n}(\widetilde{\tau}^{\mathrm{bc}}_{M,K} - \tau) \xrightarrow{\mathsf{d}} N(0, \sigma^2)$.

In addition, we have $\widehat{\sigma}^2 \xrightarrow{\mathsf{p}} \sigma^2$.

REMARK 4.7: There are two parts where Theorem 4.2(ii) requires stronger conditions than Theorem 4.1(ii). First, Theorem 4.2(ii) requires M to grow polynomially fast with n, whereas Theorem 4.1(ii) only requires M to (i) diverge not so fast for controlling the difference of matching units; and (ii) diverge to infinity (no matter how slowly it is) for achieving semiparametric efficiency. The assumptions in Theorems 4.1(ii) and 4.2(ii) both ensure semiparametric efficiency for bias-corrected matching-based estimators. Second, Theorem 4.1(ii) only requires Assumption B.1 for the density model. This is again weaker than the Lipschitz-type conditions (Assumption B.3) assumed in Theorem 4.2(ii) but is in line with the observations made in Abadie and Imbens (2006) and Abadie and Imbens

(2011). Of note, these relaxations are possible due to adding more smoothness assumptions on the outcome model (Assumptions 4.4–4.5 versus Assumptions 4.6–4.7).

REMARK 4.8: Technically, to use Chernozhukov et al. (2018)'s Theorem 5.1 to establish Theorem 4.2, we need some modifications. This is because Chernozhukov et al. (2018) considered estimating $1/e(X)$ and $1/(1-e(X))$ via plugging in an estimate of $e(X)$, whereas $\widetilde{\tau}^{\text{bc}}_{M,K}$ directly uses $1 + K_M(X)/M$ to estimate $1/e(X)$ and $1/(1-e(X))$ for units with $D = 1$ and $D = 0$, respectively. We elaborate the modifications in the proof of Theorem 4.2(ii).

5. SIMULATION

This section uses simulation to complement the theory. We consider bias-corrected matching estimators with either a fixed or diverging M, with the asymptotic variance estimated by either $\widehat{\sigma}^2$ or the estimator introduced in Abadie and Imbens (2006, Section 4).

The first data are from the National Supported Work (LaLonde, 1986). We use the specific sample studied in Dehejia and Wahba (1999). The data contain 185 treated and 260 control units. To simulate data from this study, we follow the Monte Carlo simulation design of Athey et al. (2023), and use the same pre-treatment variables that include "age", "education", "black", "hispanic", "married", "nodegree", "re74", and "re75". By using the conditional Wasserstein Generative Adversarial Networks (WGAN), one could then create a large population of observations similar to the real data, and have access to both potential outcomes for evaluating the treatment effect. Specifically, we directly use the conditional WGAN generated data available on the repository of Athey et al. (2023). There the population size is 1,000,000. For a given sample size n, we set $n_1 = n * 185/(185 + 260)$ and $n_0 = n * 260/(185 + 260)$, and draw samples from the generated data separately for treated and control groups. We consider $n \in \{600, 1200, 4800, 9600\}$.

The second data are from Shadish et al. (2008), which evaluated the effects of mathematical training on mathematics test performance. We use the data from the nonrandomized arm. The data contain 79 treated and 131 control units. We use nine pre-treatment covariates including "vocabulary pretest", "mathematics pretest", "number of prior mathe-

matics courses", "caucasian", "age", "male", "mother education", "father education", and "high school GPA". We follow the Monte Carlo simulation design of Athey et al. (2023) to generate new data with population size 1,000,000. For a given sample size n, we set $n_1 = n * 79/(79+131)$ and $n_0 = n * 131/(79+131)$. Other settings are the same as those of the first data.

We consider the estimator $\widehat{\tau}_M^{\text{bc}}$ with both fixed $M \in \{1, 4, 16\}$ and diverging $M = \lfloor \alpha n^{2/(2+d)} \rfloor$ of $d = 4$ for the first data and $d = 7$ for the second data; here the diverging rate is suggested by Theorem B.4. In this study, we pick $\alpha \in \{0.5, 1, 2, 5, 10\}$. Notice that here we choose $d = 4$ for the first data since in the eight pre-treatment variables there are only four continuous variables; it is straightforward to check that the rest four binary variables won't affect the asymptotic properties established in this manuscript as well as those in Abadie and Imbens (2006). We choose $d = 7$ for the second data based on the same reason. For the outcome models, we consider the second order power series. The estimator's asymptotic variance is estimated using either $\widehat{\sigma}$ in Theorem 4.1(ii) (SE) or Abadie and Imbens (2006)'s (AISE). We implement 2,000 repetitions and Tables I and II report the calculated root-mean-squared-error (RMSE), bias, standard deviation (SD), mean-absolute-error (MAE), and the empirical coverage rate for nominal 95% and 90% confidence intervals. Tables I and II also provide, inside the parentheses, the root-n scaled RMSE, bias, SD, and MAE divided by σ^*, with σ^* computed as the sample size-scaled standard deviation of $\widehat{\tau}_M^{\text{bc}}$ with the sample size chosen to be $100,000$, $\alpha = 1$, and $2,000$ Monte Carlo repetitions. Here we use the value $(\sigma^*)^2$ to approximate the semiparametric efficiency lower bound if the assumptions in Theorem 4.1 hold.

For the first data, two observations are in-line.

1. Regardless of which n is chosen, picking $M = \lfloor \alpha n^{2/(2+d)} \rfloor$ with α set to be 1 nearly always achieves the smallest SD, RMSE, and MAE. The simulation results thus support our recommendation to increase M for achieving better statistical performance.
2. Although consistency is established under different requirements for M, the two considered asymptotic variance estimators (SE and AISE) both yield good empirical cov-

erage rates. The coverage rates are both very close to the nominal ones when n is large and there is not much difference between the two.

Some similar observations can be found for the second data. Notably, although picking $M = \lfloor \alpha n^{2/(2+d)} \rfloor$ with $\alpha = 1$ is not achieving the smallest RMSE this time, its RMSE is very close to the smallest. However, for the second data, AISE yields generally better coverage rates than SE, although SE's coverage rates are improving as n increases.

To conclude, the simulation results generally support (a) increasing M with the sample size n for minimizing the RMSE; and (b) exploiting Abadie and Imbens (2006)'s approach to estimating the asymptotic variance of $\widehat{\tau}_M^{\mathrm{bc}}$. For choosing the M, the simulation results favor $M = \lfloor \alpha n^{2/(2+d)} \rfloor$ with α selected to be 1, while calculating the theoretically optimal α is believed to be difficult and also is beyond the scope of this manuscript.

The codes to reproduce all the simulation results presented in this section are available on https://github.com/zhexiaolin/Estimation-based-on-nearest-neighbor-matching.

6. SOME FINAL REMARKS

Some alternative matching estimators can also achieve double robustness or semiparametric efficiency. Yang and Zhang (2023) proposed to use the NN matching based on the propensity score (Rosenbaum and Rubin, 1983, Abadie and Imbens, 2016) and the prognostic score (Hansen, 2008) simultaneously, and established the double robustness of the resulting matching estimator. They focused on fixed M, and consequently, their estimator did not achieve semiparametric efficiency. Wang and Zubizarreta (2023) proposed a matching method based on integer programming to ensure global balance of the covariates, and established the efficiency of the resulting difference-in-means estimator. They focused on fixed M, and even with fixed M, their integer programming problem was computationally challenging compared with NN matching.

There are three additional questions addressed in Abadie and Imbens (2006, 2012). First, estimation of the average treatment effect on the treated (ATT) can be incorporated in the double robustness and double machine learning framework (Theorem 4.2) and matching framework (Theorem 4.1(ii)) in a similar way. Second, asymptotic normality (with an additional asymptotic bias term) of $\widehat{\tau}_M$ in general d can be established as Theorem 4.1(ii).

TABLE I

SIMULATION RESULTS, LALONDE (1986), $\sigma^* = 9.55$.

n	M	RMSE	Bias	SD	MAE	95% Coverage		90% Coverage	
						SE	AISE	SE	AISE
600	$M=1$	1.055(2.71)	-0.039(-0.10)	1.054(2.70)	0.469(1.20)	0.930	0.913	0.868	0.858
	$M=4$	1.043(2.68)	-0.038(-0.10)	1.042(2.67)	0.442(1.13)	0.926	0.911	0.862	0.847
	$M=16$	1.037(2.66)	-0.027(-0.07)	1.036(2.66)	0.435(1.12)	0.931	0.913	0.873	0.858
	$\alpha=0.5$	1.043(2.68)	-0.038(-0.10)	1.042(2.67)	0.442(1.13)	0.926	0.911	0.862	0.847
	$\alpha=1$	1.039(2.67)	-0.034(-0.09)	1.039(2.67)	0.437(1.12)	0.928	0.911	0.864	0.845
	$\alpha=2$	1.037(2.66)	-0.027(-0.07)	1.036(2.66)	0.435(1.12)	0.931	0.913	0.873	0.858
	$\alpha=5$	1.037(2.66)	-0.022(-0.06)	1.037(2.66)	0.434(1.11)	0.948	0.927	0.891	0.869
	$\alpha=10$	1.037(2.66)	-0.058(-0.15)	1.036(2.66)	0.433(1.11)	0.947	0.926	0.901	0.882
1200	$M=1$	0.341(1.24)	-0.001(-0.00)	0.341(1.24)	0.272(0.99)	0.939	0.941	0.886	0.890
	$M=4$	0.310(1.12)	0.002(0.01)	0.310(1.12)	0.248(0.90)	0.934	0.936	0.884	0.887
	$M=16$	0.305(1.11)	0.014(0.05)	0.305(1.11)	0.244(0.88)	0.939	0.940	0.887	0.889
	$\alpha=0.5$	0.309(1.12)	0.003(0.01)	0.309(1.12)	0.247(0.90)	0.940	0.942	0.882	0.883
	$\alpha=1$	0.305(1.11)	0.008(0.03)	0.305(1.11)	0.244(0.89)	0.941	0.943	0.882	0.884
	$\alpha=2$	0.306(1.11)	0.018(0.06)	0.305(1.11)	0.244(0.89)	0.939	0.939	0.886	0.887
	$\alpha=5$	0.307(1.11)	0.029(0.10)	0.306(1.11)	0.246(0.89)	0.950	0.950	0.898	0.895
	$\alpha=10$	0.307(1.11)	0.020(0.07)	0.306(1.11)	0.245(0.89)	0.955	0.956	0.908	0.907
4800	$M=1$	0.163(1.18)	-0.001(-0.01)	0.163(1.18)	0.129(0.94)	0.949	0.948	0.900	0.901
	$M=4$	0.149(1.08)	-0.000(-0.00)	0.149(1.08)	0.118(0.86)	0.951	0.951	0.897	0.897
	$M=16$	0.145(1.05)	0.001(0.01)	0.145(1.05)	0.116(0.84)	0.952	0.950	0.903	0.903
	$\alpha=0.5$	0.146(1.06)	-0.000(-0.00)	0.146(1.06)	0.117(0.85)	0.949	0.948	0.899	0.899
	$\alpha=1$	0.145(1.05)	0.001(0.01)	0.145(1.05)	0.116(0.84)	0.952	0.950	0.903	0.903
	$\alpha=2$	0.145(1.05)	0.006(0.04)	0.144(1.05)	0.116(0.84)	0.953	0.953	0.906	0.907
	$\alpha=5$	0.145(1.05)	0.017(0.12)	0.144(1.05)	0.116(0.84)	0.957	0.957	0.906	0.903
	$\alpha=10$	0.147(1.06)	0.027(0.19)	0.144(1.05)	0.117(0.85)	0.958	0.958	0.909	0.910
9600	$M=1$	0.115(1.18)	-0.003(-0.03)	0.115(1.18)	0.092(0.94)	0.951	0.952	0.897	0.896
	$M=4$	0.106(1.08)	-0.002(-0.02)	0.105(1.08)	0.084(0.86)	0.950	0.950	0.901	0.901
	$M=16$	0.103(1.06)	-0.001(-0.01)	0.103(1.06)	0.082(0.84)	0.948	0.948	0.902	0.902
	$\alpha=0.5$	0.104(1.07)	-0.001(-0.01)	0.104(1.07)	0.082(0.85)	0.953	0.951	0.904	0.905
	$\alpha=1$	0.103(1.06)	-0.001(-0.01)	0.103(1.06)	0.082(0.84)	0.950	0.950	0.904	0.903
	$\alpha=2$	0.103(1.05)	0.001(0.02)	0.103(1.05)	0.082(0.84)	0.954	0.953	0.906	0.906
	$\alpha=5$	0.104(1.07)	0.010(0.11)	0.103(1.06)	0.083(0.85)	0.951	0.950	0.899	0.898
	$\alpha=10$	0.106(1.09)	0.020(0.20)	0.104(1.07)	0.084(0.87)	0.951	0.951	0.899	0.900

TABLE II

SIMULATION RESULTS, SHADISH ET AL. (2008), $\sigma^* = 3.85$.

n	M	RMSE	Bias	SD	MAE	95% Coverage		90% Coverage	
						SE	AISE	SE	AISE
600	$M=1$	0.190(1.21)	0.005(0.03)	0.190(1.21)	0.152(0.97)	0.879	0.942	0.806	0.888
	$M=4$	0.182(1.16)	0.007(0.04)	0.182(1.16)	0.144(0.92)	0.879	0.926	0.799	0.875
	$M=16$	0.180(1.14)	0.011(0.07)	0.179(1.14)	0.143(0.91)	0.865	0.921	0.785	0.866
	$\alpha=0.5$	0.185(1.18)	0.006(0.04)	0.185(1.18)	0.147(0.94)	0.875	0.932	0.799	0.880
	$\alpha=1$	0.182(1.16)	0.007(0.04)	0.182(1.16)	0.144(0.92)	0.879	0.926	0.799	0.875
	$\alpha=2$	0.180(1.15)	0.009(0.06)	0.180(1.14)	0.143(0.91)	0.870	0.922	0.798	0.866
	$\alpha=5$	0.179(1.14)	0.012(0.07)	0.179(1.14)	0.143(0.91)	0.863	0.922	0.780	0.863
	$\alpha=10$	0.179(1.14)	0.012(0.08)	0.179(1.14)	0.142(0.91)	0.856	0.924	0.780	0.864
1200	$M=1$	0.130(1.17)	0.006(0.05)	0.129(1.16)	0.103(0.93)	0.905	0.948	0.840	0.893
	$M=4$	0.123(1.11)	0.008(0.07)	0.123(1.11)	0.098(0.88)	0.898	0.934	0.822	0.879
	$M=16$	0.122(1.09)	0.013(0.12)	0.121(0.19)	0.097(0.87)	0.892	0.931	0.818	0.878
	$\alpha=0.5$	0.125(1.13)	0.008(0.07)	0.125(1.13)	0.100(0.90)	0.898	0.943	0.836	0.885
	$\alpha=1$	0.123(1.11)	0.008(0.07)	0.123(1.11)	0.098(0.88)	0.898	0.934	0.822	0.879
	$\alpha=2$	0.122(1.10)	0.011(0.10)	0.121(1.09)	0.097(0.87)	0.896	0.934	0.819	0.877
	$\alpha=5$	0.122(1.10)	0.015(0.13)	0.121(1.09)	0.097(0.87)	0.889	0.932	0.816	0.875
	$\alpha=10$	0.121(1.09)	0.017(0.16)	0.120(1.08)	0.097(0.87)	0.880	0.930	0.799	0.876
4800	$M=1$	0.064(1.15)	0.006(0.11)	0.063(1.14)	0.051(0.91)	0.918	0.943	0.858	0.890
	$M=4$	0.060(1.09)	0.007(0.12)	0.060(1.08)	0.048(0.87)	0.912	0.939	0.839	0.877
	$M=16$	0.060(1.08)	0.009(0.17)	0.059(1.07)	0.048(0.86)	0.902	0.926	0.825	0.865
	$\alpha=0.5$	0.061(1.09)	0.006(0.12)	0.060(1.09)	0.048(0.87)	0.918	0.941	0.844	0.876
	$\alpha=1$	0.060(1.08)	0.007(0.13)	0.060(1.07)	0.048(0.86)	0.908	0.933	0.838	0.870
	$\alpha=2$	0.060(1.08)	0.009(0.16)	0.059(1.07)	0.048(0.86)	0.902	0.930	0.829	0.865
	$\alpha=5$	0.060(1.08)	0.012(0.21)	0.059(1.06)	0.048(0.86)	0.895	0.920	0.824	0.861
	$\alpha=10$	0.060(1.09)	0.015(0.26)	0.059(1.05)	0.048(0.87)	0.891	0.916	0.819	0.858
9600	$M=1$	0.045(1.14)	0.005(0.14)	0.044(1.13)	0.036(0.91)	0.923	0.940	0.864	0.886
	$M=4$	0.042(1.07)	0.006(0.15)	0.042(1.06)	0.034(0.86)	0.920	0.933	0.853	0.881
	$M=16$	0.042(1.06)	0.008(0.20)	0.041(1.04)	0.033(0.85)	0.910	0.928	0.847	0.869
	$\alpha=0.5$	0.042(1.08)	0.006(0.15)	0.042(1.06)	0.034(0.86)	0.922	0.938	0.856	0.882
	$\alpha=1$	0.042(1.06)	0.006(0.16)	0.041(1.05)	0.033(0.85)	0.916	0.934	0.851	0.872
	$\alpha=2$	0.042(1.06)	0.008(0.20)	0.041(1.04)	0.033(0.85)	0.912	0.929	0.847	0.869
	$\alpha=5$	0.042(1.07)	0.010(0.25)	0.041(1.04)	0.034(0.86)	0.902	0.922	0.842	0.862
	$\alpha=10$	0.042(1.08)	0.012(0.31)	0.041(1.03)	0.034(0.86)	0.897	0.916	0.829	0.860

Third, unbalanced designs with n_0 much larger than n_1 cannot be incorporated in the double robustness and double machine learning framework, but can be studied in the same way as Theorem 4.1(ii).

APPENDIX A: DENSITY RATIO ESTIMATION I: COMPUTATION

Additional notation. For any two real sequences $\{a_n\}_{n=1}^{\infty}$and $\{b_n\}_{n=1}^{\infty}$, write $a_n \lesssim b_n$ (or equivalently, $b_n \gtrsim a_n$) if there exists a universal constant $C > 0$ such that $a_n/b_n \leqslant C$ for all sufficiently large n, and write $a_n \prec b_n$ (or equivalently, $b_n \succ a_n$) if $a_n/b_n \to 0$ as n goes to infinity. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. We write $a_n = O(b_n)$ if $|a_n| \lesssim b_n$ and $a_n = o(b_n)$ if $|a_n| \prec b_n$. Denote the closed ball in $\mathbb{R}^d$ centered at x with radius δ by $B_{x,\delta}$. In the sequel, let $c, C, C', C'', C''', ...$ be generic positive constants whose actual values may change at different locations.

This section discusses implementation and establishes Property (P1) for the proposed estimator $\widehat{r}_M(\cdot)$. To this end, we separately discuss two cases:

Case I: estimating only the values of $\widehat{r}_M(\cdot)$ at the observed data points $X_1, \ldots, X_{N_0}$;

Case II: estimating the values of $\widehat{r}_M(\cdot)$ at both the observed data points $X_1, \ldots, X_{N_0}$ and n new points $x_1, \ldots, x_n \in \mathbb{R}^d$.

Case I. In many applications, we are only interested in a functional of density ratios at observed sample points, i.e., the values of $\Phi\big(\{r(X_i)\}_{i=1}^{N_0}\big)$ for some given functions Φ defined on $\mathbb{R}^{N_0}$. Check, e.g., in a slightly different but symmetric form – (3.3) for such an example on ATE estimation. To this end, it is natural to consider the plug-in estimator $\Phi\big(\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}\big)$, for which it suffices to compute the values of $\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}$.

Built on the k-d tree structure (Bentley, 1975) for tracking NNs, Algorithm 1 below outlines an easy to implement algorithm to simultaneously compute all the values of $\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}$. This algorithm could be regarded as a direct extension of the celebrated Friedman-Bentley-Finkel algorithm (Friedman et al., 1977) to the NN matching setting.

Case II. Suppose we are interested in estimating density ratios at both the observed and n new points in $\mathbb{R}^d$. A naive algorithm is then to insert each new point into observed points and perform Algorithm 1 in order. However, this algorithm is not ideal as the correspond-

Algorithm 1: Density ratio estimators at sample points.

Input: $\{X_i\}_{i=1}^{N_0}$, $\{Z_j\}_{j=1}^{N_1}$, and M.

Output: $\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}$.

Build a k-d tree using $\{X_i\}_{i=1}^{N_0}$;

for $j = 1 : N_1$ **do**

  Search the M-NNs of Z_j in $\{X_i\}_{i=1}^{N_0}$ using the k-d tree;

  Store the indices of the M-NNs of Z_j as S_j;

Count and store the number of occurrence in $\bigcup_{j=1}^{N_1} S_j$ for each element in $[\![N_0]\!]$, which is then $\{K_M(X_i)\}_{i=1}^{N_0}$;

Obtain $\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}$ based on (2.4).

ing time complexity would be n times the complexity of Algorithm 1, which could be computationally heavy with a large number of new points.

Instead, we develop a more sophisticated implementation. Let the new points be $\{x_i\}_{i=1}^{n}$. Algorithm 2 computes all the values of $\{\widehat{r}_M(x_i)\}_{i=1}^{n}$ as well as $\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}$. The key message delivered here is that, compared with the aforementioned naive implementation, in Algorithm 2 we only need to construct one single k-d tree; the matching elements are then categorized to two different sets, corresponding to those with regard to X_i's and x_i's, separately. Such an implementation is thus intuitively much more efficient.

Theorem A.1 below elaborates on the computational advantage of the proposed estimator.

THEOREM A.1: *(1) The average time complexity of Algorithm 1 to compute all the values of* $\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}$ *is* $O\Big((d + N_1M/N_0)N_0 \log N_0\Big)$.

(2) Assume $[x_i]_{i=1}^{n}$ *are independent and identically distributed (i.i.d.) following* ν_0 *and are independent of* $[X_i]_{i=1}^{N_0}$. *Then the average time complexity of Algorithm 2 to compute all the values of* $\{\widehat{r}_M(x_i)\}_{i=1}^{n}$ *and* $\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}$ *is* $O\Big((d + N_1M/N_0)(N_0 + n)\log(N_0 + n)\Big)$.

Algorithm 2: Density ratio estimators at both sample and new points.

Input: $\{X_i\}_{i=1}^{N_0}$, $\{Z_j\}_{j=1}^{N_1}$, M, and new points $\{x_i\}_{i=1}^{n}$.

Output: $\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}$ and $\{\widehat{r}_M(x_i)\}_{i=1}^{n}$.

Build a k-d tree using $\{X_i\}_{i=1}^{N_0} \cup \{x_i\}_{i=1}^{n}$;

for $j = 1 : N_1$ **do**

  Set S_j and S_j' be two empty sets;

  $m \leftarrow 1$;

  while $|S_j| < M$ **do**

    Search the m-th NN of Z_j in $\{X_i\}_{i=1}^{N_0} \cup \{x_i\}_{i=1}^{n}$;

    if *the m-th NN of Z_j is in* $\{X_i\}_{i=1}^{N_0}$ **then**

      add the index into S_j;

    else

      add the index into S_j';

    $m \leftarrow m + 1$;

  Store the indices sets S_j and S_j';

Count and store the number of occurrence in $\bigcup_{j=1}^{N_1} S_j$ for each element in $[\![N_0]\!]$, which is then $\{K_M(X_i)\}_{i=1}^{N_0}$. Count and store the number of occurrence in $\bigcup_{j=1}^{N_1} S_j'$ for each element in $[\![n]\!]$, which is then $\{K_M(x_i)\}_{i=1}^{n}$;

Obtain $\{\widehat{r}_M(X_i)\}_{i=1}^{N_0}$ and $\{\widehat{r}_M(x_i)\}_{i=1}^{n}$ based on (2.4).

REMARK A.1—Comparison to non-NN-based estimators: Assuming $N_0 \asymp N_1 \asymp N$, it is worth noting that optimization-based methods are commonly of a time complexity $O(N^2)$ if not worse (Noshad et al., 2017). They are thus less appealing in terms of handling gigantic data as was argued in, e.g., astronomy (Lima et al., 2008, Kremer et al., 2015) and big text analysis (Borgeaud et al., 2021) applications.

REMARK A.2—Comparison to the two-step NN-based density ratio estimator: Regarding Case I, a direct calculation yields that the time complexity of the simple two-step NN-based method, which separately estimates f_1 and f_0 based on individual M-NN den-

sity estimators, is $O(dN_0 \log N_0 + dN_1 \log N_1 + N_0 M \log N_0 + N_0 M \log N_1)$. It is thus of the same order as Algorithm 1 when $N_1 \asymp N_0$, while computationally heavier when $N_1 \prec N_0$. Regarding Case II, the time complexity of the simple two-step NN-based method is $O(dN_0 \log N_0 + dN_1 \log N_1 + (N_0 + n)M \log N_0 + (N_0 + n)M \log N_1)$. Thus, if n is of less or equal order of N_0, it is of the same order when $N_1 \asymp N_0$, while computationally heavier than Algorithm 2 when $N_1 \prec N_0$.

REMARK A.3—Comparison to the one-step NN-based density ratio estimator in Noshad et al. (2017): To estimate f-divergence measures, Noshad et al. (2017) constructed another one-step NN-based estimator admitting the simple form: $\widehat{r}'_M(x) = (N_0/N_1)(\mathcal{M}_i/(\mathcal{N}_i+1))$, where $\mathcal{N}_i$ and $\mathcal{M}_i$ are the numbers of points in $\{X_i\}_{i=1}^{N_0}$ and $\{Z_i\}_{i=1}^{N_1}$ among the M NNs of x; cf. Noshad et al. (2017, Equ. (20)). For Case I, its time complexity is $O(d(N_0 + N_1)\log(N_0 + N_1) + N_0 M \log(N_0 + N_1))$; while for Case II it is $O(d(N_0 + N_1)\log(N_0 + N_1) + (N_0 + n)M \log(N_0 + N_1))$. Both are at the same order as the naive NN-based one, but unlike the naive approach, this estimator is indeed one-step. However, it is still theoretically unclear if this estimator is statistically efficient; see Remark B.4 ahead for more details.

APPENDIX B: DENSITY RATIO ESTIMATION II: THEORY

This section introduces the theory for density ratio estimation based on NN matching. To this end, before establishing detailed theoretical properties (e.g., consistency and the rate of convergence) for $\widehat{r}_M(\cdot)$, we first exhibit a lemma elaborating on the (asymptotic) L^p moments of $\nu_1(A_M(x))$, the ν_1-measure of the catchment area. This novel result did not appear in Abadie and Imbens's analysis. It is also of independent interest in stochastic and computational geometry in light of Remark 2.1.

LEMMA B.1—Asymptotic L^p moments of catchment areas's ν_1-measure: *Assuming* $M \log N_0/N_0 \to 0$ *as* $N_0 \to \infty$, *we have* $\lim_{N_0\to\infty}(N_0/M)\mathrm{E}[\nu_1(A_M(x))] = r(x)$ *holds for* ν_0*-almost all* x. *If we further assume* $M \to \infty$, *then for any positive integer* p, *we have* $\lim_{N_0\to\infty}(N_0/M)^p\mathrm{E}[\nu_1^p(A_M(x))] = [r(x)]^p$ *holds for* ν_0*-almost all* x.

REMARK B.1—Relation to the measure of Voronoi cells: When $M = 1$ and $\nu_0 = \nu_1$, the measure of catchment areas reduces to the measure of Voronoi cells as pointed out in Remark 2.1. Interestingly, in the stochastic geometry literature, Devroye et al. (2017) studied a related problem of bounding the moments of the measure of Voronoi cells (cf. Theorem 2.1 therein). Setting $M = 1$ and $\nu_0 = \nu_1$ in the first part of Lemma B.1 and recalling Remark 2.1, we can derive their Theorem 2.1(i). On the other hand, Devroye et al. (2017, Theorem 2.1(ii)) showed that when $\nu_0 = \nu_1$, $p = 2$, and $d \leqslant 3$, $(M^{-1}N_0)^2\mathrm{E}[\nu_1^2(A_M(x))]$ converges to 1 whereas $N_0^2\mathrm{E}[\nu_1^2(A_1(x))]$ does not; cf. Devroye et al. (2017, Section 4.2). This supports the necessity of forcing $M \to \infty$ for stabilizing the moments of $\widehat{r}_M(\cdot)$.

B.1. *Consistency*

We first establish the pointwise consistency of the estimator $\widehat{r}_M(x)$ for $r(x)$. This requires nearly no assumption on ν_0, ν_1 except for those made at the beginning of Section 2, in line with similar observations made in NN-based density estimation (Biau and Devroye, 2015, Theorem 3.1).

THEOREM B.1—Pointwise consistency: *Assume* $M \log N_0/N_0 \to 0$ *as* $N_0 \to \infty$.

(i) (Asymptotic unbiasedness) For ν_0*-almost all* x*, we have* $\lim_{N_0\to\infty} \mathrm{E}\big[\widehat{r}_M(x)\big] = r(x)$.

(ii) (Pointwise L_p *consistency) Let* p *be any positive integer and assume further that* $MN_1/N_0 \to \infty$ *and* $M \to \infty$ *as* $N_0 \to \infty$*. Then for* ν_0*-almost all* x*, we have* $\lim_{N_0\to\infty} \mathrm{E}\big[|\widehat{r}_M(x) - r(x)|^p\big] = 0$.

For evaluating the global consistency of the estimator, on the other hand, it is necessary to introduce the following (global) L_p risk:

$$L_p \text{ risk} = \mathrm{E}\Big[\Big|\widehat{r}_M(X) - r(X)\Big|^p \;\Big|\; X_1, \ldots, X_{N_0}, Z_1, \ldots, Z_{N_1}\Big] = \int_{\mathbb{R}^d} \Big|\widehat{r}_M(x) - r(x)\Big|^p f_0(x)\mathrm{d}x,$$

where X is a copy drawn from ν_0 that is independent of the data. For the L_p risk consistency of the estimator, we impose conditions on ν_0 and ν_1 further as follows.

Define the supports of ν_0 and ν_1 as S_0 and S_1, respectively. For any set $S \subset \mathbb{R}^d$, define the diameter of S as $\mathrm{diam}(S) = \sup_{x,z\in S}\|x - z\|$.

ASSUMPTION B.1: *(i)* ν_0, ν_1 *are two probability measures on* $\mathbb{R}^d$*, both are absolutely continuous with respect to* λ*, and* ν_1 *is absolutely continuous with respect to* ν_0*.*

(ii) There exists a constant $R > 0$ *such that* $\operatorname{diam}(S_0) \leqslant R$*.*

(iii) There exist two constants $f_L, f_U > 0$ *such that for any* $x \in S_0$ *and* $z \in S_1$*,* $f_L \leqslant f_0(x) \leqslant f_U$ *and* $f_1(z) \leqslant f_U$*.*

(iv) There exists a constant $a \in (0, 1)$ *such that for any* $\delta \in (0, \operatorname{diam}(S_0)]$ *and* $z \in S_1$*,* $\lambda(B_{z,\delta} \cap S_0) \geqslant a\lambda(B_{z,\delta})$*, recalling that* $B_{z,\delta}$ *represents the closed ball in* $\mathbb{R}^d$ *with center at* z *and radius* δ*.*

REMARK B.2: Assumption B.1 is standard in the literature for establishing the global consistency of density ratio estimators. The regularity conditions on the support ensure that the angle of the support is not too sharp, which trivially hold for any d-dimensional cube. These conditions were also enforced in Nguyen et al. (2010, Theorem 1), Sugiyama et al. (2008, Assumption 1), Kpotufe (2017, Definition 1), among many others.

We then establish the L_p risk consistency of the estimator via the Hardy–Littlewood maximal inequality (Stein, 2016); cf. Lemma S3.2 in the online appendix. Of note, this inequality was used in Han et al. (2020) in a relative manner in order to study the information-theoretic limit of entropy estimation.

THEOREM B.2—L_p risk consistency: *Assume the pair of* ν_0, ν_1 *satisfies Assumption B.1. Let* p *be any positive integer. Assume further that* $M \log N_0/N_0 \to 0$*,* $MN_1/N_0 \to \infty$*, and* $M \to \infty$ *as* $N_0 \to \infty$*. We then have*

$$\lim_{N_0 \to \infty} \mathrm{E}\Big[\int_{\mathbb{R}^d} \Big| \widehat{r}_M(x) - r(x) \Big|^p f_0(x) \mathrm{d}x \Big] = 0.$$

As a direct corollary of Theorem B.2, one can obtain the limit of any finite moment of $\nu_1(A_M(\cdot))$ with a random center. This can be regarded as a global extension to Lemma B.1.

COROLLARY B.1: *Assume the same conditions as in Theorem B.2. We then have* $\lim_{N_0 \to \infty} (N_0/M)^p \mathrm{E}[\nu_1^p(A_M(W))] = \mathrm{E}([r(W)]^p)$*, where* W *follows an arbitrary distribution that is absolutely continuous with respect to* ν_0 *and has density bounded above and below by two positive constants. In particular, it holds when* W *is drawn from* ν_0*.*

B.2. *Rates of Convergence*

In this section, we establish the rates of convergence for $\widehat{r}(x)$ under both pointwise and global measures. We first consider the pointwise mean square error (MSE) convergence rate and show that $\widehat{r}_M(\cdot)$ is minimax optimal in that regard. In the sequel, we fix an $x \in \mathbb{R}^d$ and consider the following local assumption on (ν_0, ν_1).

ASSUMPTION B.2—Local assumption: *(i)* *ν_0, ν_1 are two probability measures on $\mathbb{R}^d$, both are absolutely continuous with respect to λ, and ν_1 is absolutely continuous with respect to ν_0.*

(ii) *There exist two constants $f_L, f_U > 0$ such that $f_0(x) \geqslant f_L$ and $f_1(x) \leqslant f_U$.*

(iii) *There exists a constant $\delta > 0$ such that for any $z \in B_{x,\delta}$, $|f_0(x) - f_0(z)| \vee |f_1(x) - f_1(z)| \leqslant L\|x - z\|$ for some constants $L > 0$.*

Define the following probability class

$$\mathcal{P}_{x,\mathrm{p}}(f_L, f_U, L, d, \delta) = \Big\{(\nu_0, \nu_1) : \text{Assumption B.2 holds}\Big\}.$$

The following theorem establishes the uniform pointwise convergence rate of $\widehat{r}_M(\cdot)$.

THEOREM B.3—Pointwise rates of convergence: *Assume $M \log N_0 / N_0 \to 0$ and $M / \log N_0 \to \infty$ as $N_0 \to \infty$. Consider a sufficiently large N_0.*

(i) *Asymptotic bias:*

$$\sup_{(\nu_0,\nu_1) \in \mathcal{P}_{x,\mathrm{p}}(f_L, f_U, L, d, \delta)} \Big| \mathrm{E}\big[\widehat{r}_M(x)\big] - r(x) \Big| \leqslant C \Big(\frac{M}{N_0}\Big)^{1/d},$$

where $C > 0$ is a constant only depending on f_L, f_U, L, d.

Further assume $M N_1 / N_0 \to \infty$ as $N_0 \to \infty$.

(ii) *Asymptotic variance:*

$$\sup_{(\nu_0,\nu_1) \in \mathcal{P}_{x,\mathrm{p}}(f_L, f_U, L, d, \delta)} \mathrm{Var}\big[\widehat{r}_M(x)\big] \leqslant C' \Big[\Big(\frac{1}{M}\Big) + \Big(\frac{N_0}{M N_1}\Big) \Big],$$

where $C' > 0$ is a constant only depending on f_L, f_U.

(iii) Asymptotic MSE:

$$\sup_{(\nu_0,\nu_1)\in\mathcal{P}_{x,\mathrm{p}}(f_L,f_U,L,d,\delta)} \mathrm{E}\big[\widehat{r}_M(x)-r(x)\big]^2 \leqslant C''\Big[\Big(\frac{M}{N_0}\Big)^{2/d}+\Big(\frac{1}{M}\Big)+\Big(\frac{N_0}{MN_1}\Big)\Big],$$

where $C''>0$ is a constant only depending on f_L, f_U, L, d.

Further assume $N_1^{-\frac{d}{2+d}}\log N_0 \to 0$ as $N_0\to\infty$.

(iv) Fix $\alpha>0$ and take $M=\alpha\cdot\{N_0^{\frac{2}{2+d}}\vee(N_0N_1^{-\frac{d}{2+d}})\}$. We have

$$\sup_{(\nu_0,\nu_1)\in\mathcal{P}_{x,\mathrm{p}}(f_L,f_U,L,d,\delta)} \mathrm{E}\big[\widehat{r}_M(x)-r(x)\big]^2 \leqslant C'''(N_0\wedge N_1)^{-\frac{2}{2+d}}, \tag{B.1}$$

where $C'''>0$ is a constant only depending on f_L, f_U, L, d, α.

The final rate of convergence in (B.1) matches the established minimax lower bound in Lipschitz density function estimation (Tsybakov, 2009, Section 2). By some simple manipulation, the argument in Tsybakov (2009, Exercise 2.8) directly extends to density ratio as the latter is a harder statistical problem (Kpotufe, 2017, Remark 3). This is formally stated in the following proposition.

PROPOSITION B.1—Pointwise MSE minimax lower bound: *For sufficiently large N_0 and N_1,*

$$\inf_{\widetilde{r}}\sup_{(\nu_0,\nu_1)\in\mathcal{P}_{x,\mathrm{p}}(f_L,f_U,L,d,\delta)} \mathrm{E}\big[\widetilde{r}(x)-r(x)\big]^2 \geqslant c(N_0\wedge N_1)^{-\frac{2}{2+d}},$$

where $c>0$ is a constant only depending on f_L, f_U, L, d and the infimum is taken over all measurable functions.

We then move on to the global risk and study the rates of convergence in this regard. To this end, a global assumption on (ν_0,ν_1) is given below.

ASSUMPTION B.3—Global assumption: *(i) ν_0,ν_1 are two probability measures on $\mathbb{R}^d$, both are absolutely continuous with respect to λ, and ν_1 is absolutely continuous with respect to ν_0.*

(ii) There exists a constant $R>0$ such that $\mathrm{diam}(S_0)\leqslant R$.

(iii) *There exist two constants* $f_L, f_U > 0$ *such that for any* $x \in S_0$ *and* $z \in S_1$, $f_L \leqslant f_0(x) \leqslant f_U$ *and* $f_1(z) \leqslant f_U$.

(iv) *There exists a constant* $a \in (0,1)$ *such that for any* $\delta \in (0, \mathrm{diam}(S_0)]$ *and any* $z \in S_1$, $\lambda(B_{z,\delta} \cap S_0) \geqslant a\lambda(B_{z,\delta})$.

(v) *There exists a constant* $H > 0$ *such that the surface area (Hausdorff measure, Evans and Garzepy (2018, Section 3.3)) of* S_1 *is bounded by* H.

(vi) *There exists a constant* $L > 0$ *such that for any* $x, z \in S_1$, $|f_0(x) - f_0(z)| \vee |f_1(x) - f_1(z)| \leqslant L\|x - z\|$.

REMARK B.3: Assumption B.3 is standard in the literature for establishing the global risk of density ratio estimators; similar assumptions were made in Zhao and Lai (2022, Assumption 1) and Zhao and Lai (2020, Assumption 1). Note that the regularity conditions on the support automatically hold for d-dimensional cubes, and the restriction on the surface area is added to control the boundary effect on NN-based methods.

Define the following probability class

$$\mathcal{P}_{\mathrm{g}}(f_L, f_U, L, d, a, H, R) = \Big\{(\nu_0, \nu_1) : \text{Assumption B.3 holds}\Big\}. \tag{B.2}$$

The next theorem establishes the uniform rate of convergence of $\widehat{r}(\cdot)$ within the above probability class under the L_1 risk. This rate is further matched by a minimax lower bound derived in Theorem 1 of Zhao and Lai (2022) using similar arguments as in the pointwise case.

THEOREM B.4—Global rates of convergence under the L_1 risk: *Assume* $M \log N_0/N_0 \to 0$, $M/\log N_0 \to \infty$, $MN_1/N_0 \to \infty$ *as* $N_0 \to \infty$. *Consider a sufficiently large* N_0.

(i) *We have the following uniform upper bound,*

$$\sup_{(\nu_0,\nu_1)\in\mathcal{P}_{\mathrm{g}}(f_L,f_U,L,d,a,H,R)} \mathrm{E}\Big[\int_{\mathbb{R}^d} \big|\widehat{r}_M(x) - r(x)\big| f_0(x)\mathrm{d}x\Big]$$
$$\leqslant C\Big[\Big(\frac{M}{N_0}\Big)^{1/d} + \Big(\frac{1}{M}\Big)^{1/2} + \Big(\frac{N_0}{MN_1}\Big)^{1/2}\Big],$$

where $C > 0$ *is a constant only depending on* f_L, f_U, a, H, L, d.

(ii) Further assume $N_1^{-\frac{d}{2+d}} \log N_0 \to 0$ *as* $N_0 \to \infty$, *fix* $\alpha > 0$, *and take* $M = \alpha \cdot \{N_0^{\frac{2}{2+d}} \vee (N_0 N_1^{-\frac{d}{2+d}})\}$. *We then have*

$$\sup_{(\nu_0,\nu_1)\in\mathcal{P}_{\mathrm{g}}(f_L,f_U,L,d,a,H,R)} \mathrm{E}\Big[\int_{\mathbb{R}^d} \Big|\widehat{r}_M(x) - r(x)\Big| f_0(x)\mathrm{d}x\Big] \leqslant C'(N_0 \wedge N_1)^{-\frac{1}{2+d}},$$

where $C' > 0$ *is a constant only depending on* $f_L, f_U, a, H, L, d, \alpha$.

PROPOSITION B.2—Global minimax lower bound under the L_1 risk: *If* a *is sufficiently small and* H, R *are sufficiently large, then for sufficiently large* N_0 *and* N_1,

$$\inf_{\widetilde{r}} \sup_{(\nu_0,\nu_1)\in\mathcal{P}_{\mathrm{g}}(f_L,f_U,L,d,a,H,R)} \mathrm{E}\Big[\int_{\mathbb{R}^d} \Big|\widetilde{r}(x) - r(x)\Big| f_0(x)\mathrm{d}x\Big] \geqslant c(N_0 \wedge N_1)^{-\frac{1}{2+d}},$$

where $c > 0$ *is a constant only depending on* f_L, f_U, L, d *and the infimum is taken over all measurable functions.*

REMARK B.4—Comparison to the one-step estimator in Noshad et al. (2017): The estimator introduced in Remark A.3 by Noshad et al. (2017) is, to our knowledge, the only alternative density ratio estimator in the literature that is able to attain both the property (P1) and being one-step. However, the arguments in Noshad et al. (2017, Section III) can only yield the bound $\mathrm{E}\big[\widehat{r}'_M(x) - r(x)\big]^2 \lesssim (M/N_0)^{1/d} + M^{-1}$ for $(\nu_0,\nu_1) \in \mathcal{P}_{x,\mathrm{p}}(f_L, f_U, L, d, \delta)$. This is via Equ. (21) therein, de-poissonizing the estimator, and further assuming N_1/N_0 converges to a positive constant. The above bound is strictly looser than the bound $(M/N_0)^{2/d} + M^{-1}$ for $\widehat{r}_M(\cdot)$ shown in Theorem B.3. However, it seems mathematically challenging to improve their analysis and accordingly, unlike $\widehat{r}_M(\cdot)$, it is still theoretically unclear if the estimator $\widehat{r}'_M(x)$ is a statistically efficient density ratio estimator.

APPENDIX C: PROOFS OF THE RESULTS IN SECTIONS 3 AND 4

C.1. *Proof of Lemma 3.1*

PROOF OF LEMMA 3.1: By simple algebra, we have

$$\widehat{\tau}_M^{\mathrm{bc}} = \frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{Y}_i^{\mathrm{bc}}(1) - \widehat{Y}_i^{\mathrm{bc}}(0)\Big]$$

$$
\begin{aligned}
=&\frac{1}{n}\sum_{i=1,D_i=1}^{n}\Big[Y_i-\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}(Y_j+\widehat{\mu}_0(X_i)-\widehat{\mu}_0(X_j))\Big]\\
&+\frac{1}{n}\sum_{i=1,D_i=0}^{n}\Big[\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}(Y_j+\widehat{\mu}_1(X_i)-\widehat{\mu}_1(X_j))-Y_i\Big]\\
=&\frac{1}{n}\sum_{i=1,D_i=1}^{n}\Big[\widehat{R}_i+\widehat{\mu}_1(X_i)-\widehat{\mu}_0(X_i)-\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\widehat{R}_j\Big]\\
&+\frac{1}{n}\sum_{i=1,D_i=0}^{n}\Big[\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\widehat{R}_j-\widehat{R}_i+\widehat{\mu}_1(X_i)-\widehat{\mu}_0(X_i)\Big]\\
=&\frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{\mu}_1(X_i)-\widehat{\mu}_0(X_i)\Big]+\frac{1}{n}\Big[\sum_{i=1,D_i=1}^{n}\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i-\sum_{i=1,D_i=0}^{n}\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i\Big].
\end{aligned}
$$

This completes the proof. *Q.E.D.*

C.2. *Proof of Theorem 4.1*

PROOF OF THEOREM 4.1(I): **Part I.** Suppose the density function is sufficiently smooth. For any $i\in[\![n]\!]$, let $\bar{R}_i=Y_i-\bar{\mu}_{D_i}(X_i)$. From (3.3),

$$
\begin{aligned}
&\widehat{\tau}^{\mathrm{bc}}_M=\widehat{\tau}^{\mathrm{reg}}+\frac{1}{n}\Big[\sum_{i=1}^{n}D_i\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i-\sum_{i=1}^{n}(1-D_i)\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i\Big]\\
=&\frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{\mu}_1(X_i)-\bar{\mu}_1(X_i)\Big]-\frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{\mu}_0(X_i)-\bar{\mu}_0(X_i)\Big]\\
&+\frac{1}{n}\Big[\sum_{i=1}^{n}(2D_i-1)\Big(1+\frac{K_M(i)}{M}\Big)\Big(\bar{\mu}_{D_i}(X_i)-\widehat{\mu}_{D_i}(X_i)\Big)\Big]\\
&+\frac{1}{n}\Big[\sum_{i=1}^{n}D_i\Big(1+\frac{K_M(i)}{M}-\frac{1}{e(X_i)}\Big)\bar{R}_i-\sum_{i=1}^{n}(1-D_i)\Big(1+\frac{K_M(i)}{M}-\frac{1}{1-e(X_i)}\Big)\bar{R}_i\Big]\\
&+\frac{1}{n}\Big[\sum_{i=1}^{n}\Big(1-\frac{D_i}{e(X_i)}\Big)\bar{\mu}_1(X_i)-\sum_{i=1}^{n}\Big(1-\frac{1-D_i}{1-e(X_i)}\Big)\bar{\mu}_0(X_i)\Big]\\
&+\frac{1}{n}\Big[\sum_{i=1}^{n}\frac{D_i}{e(X_i)}Y_i-\sum_{i=1}^{n}\frac{1-D_i}{1-e(X_i)}Y_i\Big]. \qquad \text{(C.1)}
\end{aligned}
$$

For each pair of terms, we only establish the first half part under treatment, and the second half under control can be established in the same way.

For the first term in (C.1),

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{\mu}_1(X_i)-\bar{\mu}_1(X_i)\Big]\Big| \leqslant \|\widehat{\mu}_1-\bar{\mu}_1\|_\infty = o_{\mathrm{P}}(1). \tag{C.2}$$

For the second term in (C.1),

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^{n} D_i\Big(1+\frac{K_M(i)}{M}\Big)\Big(\bar{\mu}_1(X_i)-\widehat{\mu}_1(X_i)\Big)\Big| \\ \leqslant &\|\widehat{\mu}_1-\bar{\mu}_1\|_\infty \frac{1}{n}\sum_{i=1}^{n} D_i\Big(1+\frac{K_M(i)}{M}\Big) = \|\widehat{\mu}_1-\bar{\mu}_1\|_\infty = o_{\mathrm{P}}(1),\end{aligned} \tag{C.3}$$

where the last step is due to $\sum_{i=1}^{n} D_i K_M(i) = n_0 M$.

Notice that from Assumption 4.1 (i), $\mathrm{P}_{X\,|\,D=0}$ and $\mathrm{P}_{X\,|\,D=1}$ share the same support, and their densities are both bounded and bounded away from zero as long as one is. Then $(\mathrm{P}_{X\,|\,D=0}, \mathrm{P}_{X\,|\,D=1})$ and $(\mathrm{P}_{X\,|\,D=1}, \mathrm{P}_{X\,|\,D=0})$ both satisfy Assumption B.1 as long as one satisfies. For the third term in (C.1), by Theorem B.2,

$$\begin{aligned}&\Big\{\mathrm{E}\Big[\Big|\frac{1}{n}\sum_{i=1}^{n} D_i\Big(1+\frac{K_M(i)}{M}-\frac{1}{e(X_i)}\Big)\bar{R}_i\Big|\Big]\Big\}^2 \leqslant \Big\{\mathrm{E}\Big[\Big|D_i\Big(1+\frac{K_M(i)}{M}-\frac{1}{e(X_i)}\Big)\bar{R}_i\Big|\Big]\Big\}^2 \\ \leqslant &\mathrm{E}\Big[1+\frac{K_M(i)}{M}-\frac{1}{e(X_i)}\Big]^2 \mathrm{E}[D_i\bar{R}_i]^2 = \mathrm{E}\Big[1+\frac{K_M(i)}{M}-\frac{1}{e(X_i)}\Big]^2 \mathrm{E}[D_i(Y_i(1)-\bar{\mu}_1(X_i))^2] \\ \leqslant &\mathrm{E}\Big[1+\frac{K_M(i)}{M}-\frac{1}{e(X_i)}\Big]^2 \mathrm{E}\Big[\sigma_1^2(X_i)+(\mu_1(X_i)-\bar{\mu}_1(X_i))^2\Big] = o(1),\end{aligned} \tag{C.4}$$

where $\sigma_1^2(x) = \mathrm{E}[U_1^2 \,|\, X = x]$ for $x \in \mathbb{X}$.

For the fourth term in (C.1), notice that

$$\mathrm{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(1-\frac{D_i}{e(X_i)}\Big)\bar{\mu}_1(X_i)\,\Big|\,X_1,\ldots,X_n\Big] = 0,$$

and

$$\mathrm{Var}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(1-\frac{D_i}{e(X_i)}\Big)\bar{\mu}_1(X_i)\Big] = \mathrm{E}\Big[\mathrm{Var}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(1-\frac{D_i}{e(X_i)}\Big)\bar{\mu}_1(X_i)\,\Big|\,X_1,\ldots,X_n\Big]\Big]$$

$$=\frac{1}{n}\mathrm{E}\Big[\bar{\mu}_1^2(X_i)\Big(\frac{1}{e(X_i)}-1\Big)\Big]=O(n^{-1}).$$

Then

$$\frac{1}{n}\sum_{i=1}^{n}\Big(1-\frac{D_i}{e(X_i)}\Big)\bar{\mu}_1(X_i)=o_{\mathrm{P}}(1). \tag{C.5}$$

For the fifth term in (C.1), notice that $\mathrm{E}[Y^2]$ are bounded and $[(X_i,D_i,Y_i)]_{i=1}^n$ are i.i.d.. Using the weak law of large numbers yields

$$\frac{1}{n}\Big[\sum_{i=1}^{n}\frac{D_i}{e(X_i)}Y_i-\sum_{i=1}^{n}\frac{1-D_i}{1-e(X_i)}Y_i\Big]\xrightarrow{\mathsf{p}}\mathrm{E}\Big[Y_i(1)-Y_i(0)\Big]=\tau. \tag{C.6}$$

Plugging (C.2), (C.3), (C.4), (C.5), (C.6) into (C.1) completes the proof.

Part II. Suppose the outcome model is correct. By (3.3),

$$\begin{aligned}
\widehat{\tau}_M^{\mathrm{bc}}&=\widehat{\tau}^{\mathrm{reg}}+\frac{1}{n}\Big[\sum_{i=1}^{n}D_i\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i-\sum_{i=1}^{n}(1-D_i)\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i\Big]\\
=&\frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{\mu}_1(X_i)-\mu_1(X_i)\Big]-\frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{\mu}_0(X_i)-\mu_0(X_i)\Big]\\
&+\frac{1}{n}\Big[\sum_{i=1}^{n}(2D_i-1)\Big(1+\frac{K_M(i)}{M}\Big)\Big(\mu_{D_i}(X_i)-\widehat{\mu}_{D_i}(X_i)\Big)\Big]\\
&+\frac{1}{n}\Big[\sum_{i=1}^{n}D_i\Big(1+\frac{K_M(i)}{M}\Big)\Big(Y_i-\mu_1(X_i)\Big)-\sum_{i=1}^{n}(1-D_i)\Big(1+\frac{K_M(i)}{M}\Big)\Big(Y_i-\mu_0(X_i)\Big)\Big]\\
&+\frac{1}{n}\sum_{i=1}^{n}\Big[\mu_1(X_i)-\mu_0(X_i)\Big].
\end{aligned} \tag{C.7}$$

For the first term in (C.7),

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{\mu}_1(X_i)-\mu_1(X_i)\Big]\Big|\leqslant\|\widehat{\mu}_1-\mu_1\|_\infty=o_{\mathrm{P}}(1). \tag{C.8}$$

For the second term in (C.7),

$$\Big|\frac{1}{n}\sum_{i=1}^{n}D_i\Big(1+\frac{K_M(i)}{M}\Big)\Big(\mu_1(X_i)-\widehat{\mu}_1(X_i)\Big)\Big|$$

$$\leqslant \|\widehat{\mu}_1 - \mu_1\|_\infty \frac{1}{n}\sum_{i=1}^{n} D_i\Big(1 + \frac{K_M(i)}{M}\Big) = \|\widehat{\mu}_1 - \mu_1\|_\infty = o_{\mathrm{P}}(1). \tag{C.9}$$

For the third term in (C.7), noticing that $K_M(\cdot)$ is a function of $(X_1, \ldots, X_n)$ and $(D_1, \ldots, D_n)$, we can obtain

$$\mathrm{E}\Big[\frac{1}{n}\sum_{i=1}^{n} D_i\Big(1 + \frac{K_M(i)}{M}\Big)\Big(Y_i - \mu_1(X_i)\Big)\,\Big|\, X_1, \ldots, X_n, D_1, \ldots, D_n\Big] = 0.$$

By a martingale representation (Abadie and Imbens, 2012) and then the martingale convergence theorem (which holds for both fixed and diverging M), we obtain

$$\frac{1}{n}\sum_{i=1}^{n} D_i\Big(1 + \frac{K_M(i)}{M}\Big)\Big(Y_i - \mu_1(X_i)\Big) = o_{\mathrm{P}}(1). \tag{C.10}$$

For the fourth term in (C.7), notice that $\mathrm{E}[\mu_\omega^2(X)]$ is bounded for $\omega \in \{0, 1\}$. Using the weak law of large number, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\mu_1(X_i) - \mu_0(X_i)\Big] \xrightarrow{\mathsf{p}} \mathrm{E}\Big[\mu_1(X_i) - \mu_0(X_i)\Big] = \tau. \tag{C.11}$$

Plugging (C.8), (C.9), (C.10), (C.11) into (C.7) completes the proof. *Q.E.D.*

PROOF OF THEOREM 4.1 (II): For $\omega \in \{0, 1\}$ and $m \in [\![M]\!]$, let $j_m(i)$ represent the index of mth-NN of X_i in $\{X_j : D_j = 1 - D_i\}_{j=1}^n$, i.e., the index $j \in [\![n]\!]$ such that $D_j = 1 - D_i$ and $\sum_{\ell=1, D_\ell = 1-D_i}^{n} \mathbb{1}(\|X_\ell - X_i\| \leqslant \|X_j - X_i\|) = m$. With a little abuse of notation, let $\epsilon_i = Y_i - \mu_{D_i}(X_i)$ for any $i \in [\![n]\!]$. By the definition of $\widehat{\tau}_M^{\mathrm{bc}}$ in (3.3), we can verify the decomposition $\widehat{\tau}_M^{\mathrm{bc}} = \bar{\tau}(\boldsymbol{X}) + E_M + B_M - \widehat{B}_M$, where

$$\bar{\tau}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}\Big[\mu_1(X_i) - \mu_0(X_i)\Big],$$

$$E_M = \frac{1}{n}\sum_{i=1}^{n}(2D_i - 1)\Big(1 + \frac{K_M(i)}{M}\Big)\epsilon_i,$$

$$B_M = \frac{1}{n}\sum_{i=1}^{n}(2D_i - 1)\Big[\frac{1}{M}\sum_{m=1}^{M}\Big(\mu_{1-D_i}(X_i) - \mu_{1-D_i}(X_{j_m(i)})\Big)\Big],$$

$$\widehat{B}_M = \frac{1}{n}\sum_{i=1}^{n}(2D_i-1)\Big[\frac{1}{M}\sum_{m=1}^{M}\Big(\widehat{\mu}_{1-D_i}(X_i)-\widehat{\mu}_{1-D_i}(X_{j_m(i)})\Big)\Big].$$

We have the following central limit theorem on $\bar{\tau}(\boldsymbol{X})+E_M$.

LEMMA C.1: $\sqrt{n}\sigma^{-1/2}(\bar{\tau}(\boldsymbol{X})+E_M-\tau)\xrightarrow{\mathsf{d}}N(0,1)$.

For the bias term $B_M-\widehat{B}_M$, define $U_{m,i}=X_{j_m(i)}-X_i$ for any $i\in[\![n]\!]$ and $m\in[\![M]\!]$. We then have the following lemma bounding the moments of $U_{M,i}$.

LEMMA C.2: *Let p be any positive integer. Then there exists a constant $C_p>0$ only depending on p such that for any $i\in[\![n]\!]$ and $M\in[\![n_{1-D_i}]\!]$, $\mathrm{E}[\|U_{M,i}\|^p\,|\,D_1,\ldots,D_n]\leqslant C_p(M/n_{1-D_i})^{p/d}$.*

In light of the smoothness conditions on μ_ω and approximation conditions on $\widehat{\mu}_\omega$ for $\omega\in\{0,1\}$, we can establish the following lemma using Lemma C.2.

LEMMA C.3: $\sqrt{n}(B_M-\widehat{B}_M)\xrightarrow{\mathsf{p}}0$.

Combining Lemma C.1 and Lemma C.3 completes the proof. *Q.E.D.*

PROOF OF THEOREM 4.1(II), CONSISTENCY OF $\widehat{\sigma}^2$: By definition, we can verify the decomposition $\widehat{\sigma}^2-\sigma^2=T_1+T_2+T_3+T_4$, where

$$\begin{aligned}
T_1 = \frac{1}{n}\sum_{i=1}^{n}\Big[&\widehat{\mu}_1(X_i)-\widehat{\mu}_0(X_i)+D_i\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i-(1-D_i)\Big(1+\frac{K_M(i)}{M}\Big)\widehat{R}_i-\widehat{\tau}_M^{\mathrm{bc}}\Big]^2\\
&-\frac{1}{n}\sum_{i=1}^{n}\Big[\mu_1(X_i)-\mu_0(X_i)+D_i\Big(1+\frac{K_M(i)}{M}\Big)(Y_i-\mu_1(X_i))\\
&-(1-D_i)\Big(1+\frac{K_M(i)}{M}\Big)(Y_i-\mu_0(X_i))-\widehat{\tau}_M^{\mathrm{bc}}\Big]^2,\\
T_2 = \frac{1}{n}\sum_{i=1}^{n}\Big[&\mu_1(X_i)-\mu_0(X_i)+D_i\Big(1+\frac{K_M(i)}{M}\Big)(Y_i-\mu_1(X_i))\\
&-(1-D_i)\Big(1+\frac{K_M(i)}{M}\Big)(Y_i-\mu_0(X_i))-\widehat{\tau}_M^{\mathrm{bc}}\Big]^2
\end{aligned}$$

$$-\frac{1}{n}\sum_{i=1}^{n}\left[\mu_1(X_i)-\mu_0(X_i)+\frac{D_i}{e(X_i)}(Y_i-\mu_1(X_i))-\frac{1-D_i}{1-e(X_i)}(Y_i-\mu_0(X_i))-\widehat{\tau}_M^{\mathrm{bc}}\right]^2,$$

$$T_3=\frac{1}{n}\sum_{i=1}^{n}\left[\mu_1(X_i)-\mu_0(X_i)+\frac{D_i}{e(X_i)}(Y_i-\mu_1(X_i))-\frac{1-D_i}{1-e(X_i)}(Y_i-\mu_0(X_i))-\widehat{\tau}_M^{\mathrm{bc}}\right]^2$$

$$-\frac{1}{n}\sum_{i=1}^{n}\left[\mu_1(X_i)-\mu_0(X_i)+\frac{D_i}{e(X_i)}(Y_i-\mu_1(X_i))-\frac{1-D_i}{1-e(X_i)}(Y_i-\mu_0(X_i))-\tau\right]^2,$$

$$T_4=\frac{1}{n}\sum_{i=1}^{n}\left[\mu_1(X_i)-\mu_0(X_i)+\frac{D_i}{e(X_i)}(Y_i-\mu_1(X_i))-\frac{1-D_i}{1-e(X_i)}(Y_i-\mu_0(X_i))-\tau\right]^2-\sigma^2.$$

By Assumption 4.3, Assumption 4.1, Theorem B.2, and the fact that $\widehat{\tau}_M^{\mathrm{bc}}=O_{\mathrm{P}}(1)$, we have $T_1=o_{\mathrm{P}}(1)$. By Assumption 4.1, Theorem B.2, and $\widehat{\tau}_M^{\mathrm{bc}}=O_{\mathrm{P}}(1)$, we have $T_2=o_{\mathrm{P}}(1)$. By Assumption 4.1 and $\widehat{\tau}_M^{\mathrm{bc}}-\tau=o_{\mathrm{P}}(1)$, we have $T_3=o_{\mathrm{P}}(1)$. By the fact that $[(X_i,D_i,Y_i)]_{i=1}^n$ are i.i.d., Assumption 4.1 and the weak law of large numbers, we have $T_4=o_{\mathrm{P}}(1)$. Combining the above four facts together then completes the proof. *Q.E.D.*

C.3. *Proof of Theorem 4.2*

PROOF OF THEOREM 4.2: For Theorem 4.2 (i), analysis analogous to the proof of Theorem 4.1(i) can be performed on $\breve{\tau}_{M,k}^{\mathrm{bc}}$ for any $k\in[\![K]\!]$. Then the results apply to $\widetilde{\tau}_{M,K}^{\mathrm{bc}}$ automatically since K is fixed.

For Theorem 4.2 (ii), from Definition 3.1 in Chernozhukov et al. (2018), $\widetilde{\tau}_{M,K}^{\mathrm{bc}}$ is the double machine learning estimator. We then follow the proof of Theorem 5.1 (recalling Remark 4.8) and use the notations in Chernozhukov et al. (2018), essentially checking Assumption 3.1 and 3.2 therein. In the following the notation in Chernozhukov et al. (2018) is adopted.

For estimating the ATE, from Equ. (5.3) in Chernozhukov et al. (2018), the score (or the efficient influence function (Tsiatis, 2006, Section 3.4)) is

$$\psi(X,D,Y;\widetilde{\tau},\widetilde{\zeta})=\widetilde{\mu}_1(X)-\widetilde{\mu}_0(X)+\frac{D(Y-\widetilde{\mu}_1(X))}{\widetilde{e}(X)}-\frac{(1-D)(Y-\widetilde{\mu}_0(X))}{1-\widetilde{e}(X)}-\widetilde{\tau},$$

where $\widetilde{\zeta}(x) = (\widetilde{\mu}_0(x), \widetilde{\mu}_1(x), \widetilde{\rho}_0(x), \widetilde{\rho}_1(x))$ are the nuisance parameters by letting $\widetilde{\rho}_0(x) = 1/(1-\widetilde{e}(x))$ and $\widetilde{\rho}_1(x) = 1/\widetilde{e}(x)$. Let $\rho_0(x) = 1/(1-e(x))$ and $\rho_1(x) = 1/e(x)$. Then the true value is $\zeta(x) = (\mu_0(x), \mu_1(x), \rho_0(x), \rho_1(x))$.

We can then write the score as

$$\psi(X, D, Y; \widetilde{\tau}, \widetilde{\zeta}) = \widetilde{\mu}_1(X) - \widetilde{\mu}_0(X) + D(Y - \widetilde{\mu}_1(X))\widetilde{\rho}_1(X) - (1-D)(Y - \widetilde{\mu}_0(X))\widetilde{\rho}_0(X) - \widetilde{\tau}.$$

For any $p > 0$, let $\|f\|_p = \|f(X, D, Y)\|_p = (\int |f(\omega)|^p \mathrm{dP}_{(X,D,Y)}(\omega))^{1/p}$. For the κ in Assumption 4.1, let $q = 2 + \kappa/2$, $q_1 = 2 + \kappa$ and q_2 such that $q^{-1} = q_1^{-1} + q_2^{-1}$. Let $\mathcal{T}_n$ be the set consisting of all $\widetilde{\zeta}$ such that for $\omega \in \{0, 1\}$,

$$\|\widetilde{\mu}_\omega - \mu_\omega\|_\infty = o(n^{-d/(4+2d)}), \quad \|\widetilde{\rho}_\omega - \rho_\omega\|_1 = O(n^{-1/(d+2)}), \quad \|\widetilde{\rho}_\omega - \rho_\omega\|_{q_2} = o(1).$$

Then the selection of $\mathcal{T}_n$ satisfies Assumption 3.2(a) in Chernozhukov et al. (2018) from Assumption 4.7, Theorem B.4, and Theorem B.2, respectively.

For step 1 in the proof of Theorem 5.1 in Chernozhukov et al. (2018), we verify the Neyman orthogonality. We can show that $\mathrm{E}\psi(X, D, Y; \tau, \zeta) = 0$. For any $\widetilde{\zeta} \in \mathcal{T}_n$, the Gateaux derivative in the direction $\widetilde{\zeta} - \zeta$ is

$$\begin{aligned}
&\partial_{\widetilde{\zeta}} \mathrm{E}\psi(X, D, Y; \tau, \zeta)[\widetilde{\zeta} - \zeta] = \mathrm{E}[\widetilde{\mu}_1(X) - \mu_1(X)] - \mathrm{E}[\widetilde{\mu}_0(X) - \mu_0(X)] \\
&- \mathrm{E}[D(\widetilde{\mu}_1(X) - \mu_1(X))\rho_1(X)] + \mathrm{E}[(1-D)(\widetilde{\mu}_0(X) - \mu_0(X))\rho_0(X)] \\
&+ \mathrm{E}[D(Y - \mu_1(X))(\widetilde{\rho}_1(X) - \rho_1(X))] - \mathrm{E}[(1-D)(Y - \mu_0(X))(\widetilde{\rho}_0(X) - \rho_0(X))].
\end{aligned}$$

We can check that the above quantity is zero, which completes this step.

Step 2 and step 3 therein can be directly applied.

For step 4 therein, we can establish in the same way that for $\omega \in \{0, 1\}$, $\|\mu_\omega\|_{q_1} = O(1)$ from $\|Y\|_{q_1} = O(1)$, and $\tau = O(1)$. Then from Hölder's inequality and $\|\rho_\omega\|_\infty$ is bounded for $\omega \in \{0, 1\}$, for any $\widetilde{\zeta} \in \mathcal{T}_n$,

$$\begin{aligned}
&\|\psi(X, D, Y; \tau, \widetilde{\zeta})\|_q = \|\widetilde{\mu}_1(X) - \widetilde{\mu}_0(X) + (2D-1)(Y - \widetilde{\mu}_D(X))\widetilde{\rho}_D(X) - \tau\|_q \\
\leqslant &\|\widetilde{\mu}_1(X)\|_q + \|\widetilde{\mu}_0(X)\|_q + \|(Y - \widetilde{\mu}_1(X))\widetilde{\rho}_1(X)\|_q + \|(Y - \widetilde{\mu}_0(X))\widetilde{\rho}_0(X)\|_q + \tau
\end{aligned}$$

$$\leqslant \|\mu_1\|_q + \|\widetilde{\mu}_1 - \mu_1\|_\infty + \|\mu_0\|_q + \|\widetilde{\mu}_0 - \mu_0\|_\infty + (\|Y\|_{q_1} + \|\mu_1\|_{q_1} + \|\widetilde{\mu}_1 - \mu_1\|_\infty)\|\widetilde{\rho}_1\|_{q_2}$$
$$+ (\|Y\|_{q_1} + \|\mu_0\|_{q_1} + \|\widetilde{\mu}_0 - \mu_0\|_\infty)\|\widetilde{\rho}_0\|_{q_2} + \tau = O(1).$$

The last step is from the definition of $\mathcal{T}_n$. Then we complete this step.

For step 5 therein, by Hölder's inequality, for any $\widetilde{\zeta} \in \mathcal{T}_n$,

$$\|\psi(X, D, Y; \tau, \widetilde{\zeta}) - \psi(X, D, Y; \tau, \zeta)\|_2$$
$$\leqslant \|\widetilde{\mu}_1 - \mu_1\|_2 + \|\widetilde{\mu}_0 - \mu_0\|_2 + \|D(Y - \widetilde{\mu}_1(X))\widetilde{\rho}_1(X) - D(Y - \mu_1(X))\rho_1(X)\|_2$$
$$+ \|(1 - D)(Y - \widetilde{\mu}_0(X))\widetilde{\rho}_0(X) - (1 - D)(Y - \mu_0(X))\rho_0(X)\|_2$$
$$\leqslant \|\widetilde{\mu}_1 - \mu_1\|_\infty + \|\widetilde{\mu}_0 - \mu_0\|_\infty + (\|Y\|_{q_1} + \|\mu_1(X)\|_{q_1})\|\widetilde{\rho}_1 - \rho_1\|_{q_2} + \|\widetilde{\mu}_1 - \mu_1\|_\infty\|\widetilde{\rho}_1\|_2$$
$$+ (\|Y\|_{q_1} + \|\mu_0(X)\|_{q_1})\|\widetilde{\rho}_0 - \rho_0\|_{q_2} + \|\widetilde{\mu}_0 - \mu_0\|_\infty\|\widetilde{\rho}_0\|_2 = o(1).$$

The last step is due to the definition of $\mathcal{T}_n$.

Notice that for any $t \in (0, 1)$,

$$\partial_t^2 \mathrm{E}\psi(X, D, Y; \tau, \zeta + t(\widetilde{\zeta} - \zeta)) = -2\mathrm{E}[(2D - 1)(\widetilde{\mu}_D(X) - \mu_D(X))(\widetilde{\rho}_D(X) - \rho_D(X))].$$

Then by the definition of $\mathcal{T}_n$, for any $\widetilde{\zeta} \in \mathcal{T}_n$,

$$|\partial_t^2 \mathrm{E}\psi(X, D, Y; \tau, \zeta + t(\widetilde{\zeta} - \zeta))| \leqslant 2 \sum_{w \in \{0,1\}} \|\widetilde{\mu}_w - \mu_w\|_\infty \|\widetilde{\rho}_w - \rho_w\|_1 = o(n^{-1/2}).$$

We then complete this step and thus complete the proof.

The consistency of the variance estimator can be established in the same way as Theorem 4.1(ii). *Q.E.D.*

REFERENCES

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267. [1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 13, 14, 15, 16, 17, 18]

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557. [2, 4, 5, 10]

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, 29:1–11. [1, 2, 3, 4, 5, 7, 8, 10, 11, 14, 15]

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498):833–843. [2, 4, 5, 18, 34]

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84:781–807. [10, 18]

Armstrong, T. B. and Kolesár, M. (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89:1141–1177. [10]

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2023). Using Wasserstein Generative Adversarial Networks for the design of Monte Carlo simulations. *Journal of Econometrics*, (in press). [16, 17]

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973. [2, 3, 9]

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517. [21]

Berrett, T. B., Samworth, R. J., and Yuan, M. (2019). Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318. [3, 5]

Bhattacharya, B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society. Series B*, 81(3):575–602. [3]

Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer. [25]

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Driessche, G. v. d., Lespiau, J.-B., Damoc, B., Clark, A., et al. (2021). Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 2206–2240. Proceedings of Machine Learning Research. [7, 23]

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156. [2]

Chapin, F. S. (1947). *Experimental Designs in Sociological Research*. Harper and Brothers. [2]

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465. [15]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68. [3, 11, 14, 15, 16, 36, 37]

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā, Series A*, 35(4):417–446. [2]

Cover, T. M. and Thomas, J. (2006). *Elements of Information Theory (2nd Edition)*. John Wiley and Sons. [5]

Cunningham, S. (2021). *Causal Inference: the Mixtape*. Yale University Press. [5]

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062. [16]

Devroye, L., Györfi, L., Lugosi, G., and Walk, H. (2017). On the measure of Voronoi cells. *Journal of Applied Probability*, 54(2):394–408. [25]

Evans, L. C. and Garzepy, R. F. (2018). *Measure Theory and Fine Properties of Functions*. Routledge. [29]

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23. [3]

Ferman, B. (2021). Matching estimators with few treated and many control observations. *Journal of Econometrics*, 225:295–307. [10]

Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3):209–226. [21]

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717. [3]

Greenwood, E. (1945). *Experimental Sociology*. Columbia University Press. [2]

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331. [3, 12, 13]

Han, Y., Jiao, J., Weissman, T., and Wu, Y. (2020). Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48(6):3228–3250. [26]

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488. [18]

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783. [3]

Henze, N. and Penrose, M. D. (1999). On the multivariate runs test. *The Annals of Statistics*, 27(1):290–298. [3]

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236. [2]

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29. [2]

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press. [2, 5]

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21:1–54. [10]

Kpotufe, S. (2017). Lipschitz density-ratios, structured data, and data-driven tuning. In *2017 International Conference on Artificial Intelligence and Statistics*, pages 1320–1328. PMLR. [26, 28]

Kremer, J., Gieseke, F., Pedersen, K. S., and Igel, C. (2015). Nearest neighbor density ratio estimation for large-scale applications in astronomy. *Astronomy and Computing*, 12:67–72. [5, 7, 23]

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620. [16, 19]

Lima, M., Cunha, C. E., Oyaizu, H., Frieman, J., Lin, H., and Sheldon, E. S. (2008). Estimating the redshift distribution of photometric galaxy samples. *Monthly Notices of the Royal Astronomical Society*, 390(1):118–130. [5, 7, 23]

Lin, Z. and Han, F. (2023). On boosting the power of Chatterjee's rank correlation. *Biometrika*, 110(2):283–299. [3, 4]

Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260. [3]

Morgan, S. L. and Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods and Research*, 35(1):3–60. [2]

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168. [14]

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861. [2, 26]

Noshad, M., Moon, K. R., Sekeh, S. Y., and Hero, A. O. (2017). Direct estimation of information divergence using nearest neighbor ratios. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 903–907. [5, 23, 24, 30]

Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720–1732. [9]

Póczos, B. and Schneider, J. (2011). On the estimation of alpha-divergences. In *2011 International Conference on Artificial Intelligence and Statistics*, pages 609–617. [5]

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer. [2]

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55. [9, 10, 18]

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183. [2]

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701. [7]

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press. [2]

Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585. [2]

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120. [2, 3, 9]

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*, 42(7):1–52. [2]

Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1344. [16, 20]

Shi, H., Drton, M., and Han, F. (2022a). On Azadkia-Chatterjee's conditional dependence coefficient. *Bernoulli*, (in press). [3]

Shi, H., Drton, M., and Han, F. (2022b). On the power of Chatterjee's rank correlation. *Biometrika*, 109(2):317–333. [3]

Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27(1):325–353. [2]

Stein, E. M. (2016). *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press. [26]

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press. [2, 5]

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746. [26]

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer. [36]

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer. [28]

Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1908(134):198–287. [6]

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, 11(2):147–162. [3]

Wang, Y. and Zubizarreta, J. R. (2023). Large sample properties of matching for balance. *Statistica Sinica*, 33:1789–1808. [18]

Yang, S. and Zhang, Y. (2023). Multiply robust matching estimators of average and quantile treatment effects. *Scandinavian Journal of Statistics*, 50:235–265. [18]

Zhao, P. and Lai, L. (2020). Minimax optimal estimation of KL divergence for continuous distributions. *IEEE Transactions on Information Theory*, 66(12):7787–7811. [5, 29]

Zhao, P. and Lai, L. (2022). Analysis of KNN density estimation. *IEEE Transactions on Information Theory*, 68(12):7971–7995. [29]