Statistical Science
2025, Vol. 40, No. 3, 495–514
https://doi.org/10.1214/24-STS972
© Institute of Mathematical Statistics, 2025

Efficient Generalization and Transportation

Zhenghao Zeng, Edward H. Kennedy, Lisa M. Bodnar and Ashley I. Naimi

Abstract. When estimating causal effects, it is important to assess external validity, that is, determine how useful a given study is to inform a practical question for a specific target population. One challenge is that the covariate distribution in the population underlying a study may be different from that in the target population. If some covariates are effect modifiers, the average treatment effect (ATE) may not generalize to the target population. To tackle this problem, we propose new methods to generalize or transport the ATE from a source population to a target population, in the case where the source and target populations have different sets of covariates. When the ATE in the target population is identified, we propose new doubly robust estimators and establish their rates of convergence and limiting distributions. Under regularity conditions, the doubly robust estimators provably achieve the efficiency bound and are locally asymptotic minimax optimal. A sensitivity analysis is provided when the identification assumptions fail. Simulation studies show the advantages of the proposed doubly robust estimator over simple plug-in estimators. Importantly, we also provide minimax lower bounds and higher-order estimators of the target functionals. The proposed methods are applied in transporting causal effects of dietary intake on adverse pregnancy outcomes from an observational study to *the whole U.S. pregnant female population*.

Key words and phrases: Generalization, transportation, influence function, doubly robust estimation, minimax lower bounds, high-order estimation, dietary intake, adverse pregnancy outcomes.

1. INTRODUCTION

In causal inference, researchers typically aim at estimating causal effects in a particular "source" population based on randomized trials or observational studies in that population. When the source population is the sole population of interest, random samples from the source are representative and standard techniques to estimate average treatment effects (ATEs) can be applied to obtain reliable estimates of effects in the whole population. However, in many studies (e.g., randomized trials in clinical medicine, policy analysis) samples are drawn from a different population, unrepresentative of the target of interest [1, 3, 30]. In other words, the source and target populations can be different, and the ATE we obtain from the sample only applies to the source population—and may not generalize to the target population directly. Failing to consider this lack of representation may yield unreliable conclusions that can even be harmful, especially in medicine and policy evaluations [14].

One example of unrepresentative samples is when the distributions of some covariates (e.g., age and BMI) in the target population differ from those in the source population. When some of these covariates are effect modifiers (e.g., age and BMI may modify the effects of some medicine), the ATE in the target population can be quite different from that in the source population. There exists a rich literature in bridging findings from the source population to the target population [11, 13, 16–18, 42, 50]. Most of these papers adopt the idea that we first estimate the probability for a subject to be in the source population and use it as a reweighting term in the proposed estimator. For instance, [18] and [17] proposed three types of estimators based on outcome modeling, inverse prob-

Zhenghao Zeng is a Ph.D. student, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: zhenghaz@andrew.cmu.edu). Edward H. Kennedy is an Associate Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: edward@stat.cmu.edu). Lisa M. Bodnar is a Professor, Department of Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA (e-mail: bodnar@edc.pitt.edu). Ashley I. Naimi is an Associate Professor, Department of Epidemiology, Emory University, Atlanta, Georgia 30329 USA (e-mail: ashley.naimi@emory.edu).

ability weighting and doubly robust-style augmented inverse probability weighting. [42] derived efficient influence functions and robust estimators for three transported estimands: ATE, intent-to-treat ATE and complier ATE. [13] developed a one-step weighting estimator where the weights are learned from a convex optimization problem to simultaneously model the inverse propensity score and outcome regression functions. However, the theoretical properties of the doubly robust estimators have not been well understood. The form of their second-order bias and the conditions required to ensure their asymptotic normality are unclear. Another limitation is that the previous work assumes the sets of covariates in the source population and the target population are the same. In practice, there often exist some covariates that are available in the source, but not in the target. Simply ignoring them is not an efficient way to exploit the samples. For instance, including more covariates in the source may enable us to be more confident in the conditional exchangeability assumption in the source population.

In this work, we fill these two gaps by developing efficiency theory and establishing theoretical guarantees for the doubly robust estimators. In particular, we generalize previous results to accommodate covariate mismatch in the source and target populations. After formulating the generalization and transportation functional of interest, we examine the assumptions required to identify these functionals. When these functionals are identified, we derive the first-order influence functions and establish the asymptotic normality of a doubly robust estimator (under additional conditions). Simulations show how the proposed doubly robust estimator has a smaller estimation error compared to a simple plug-in estimator. To provide a complete story for generalization and transportation of causal effects, we then consider the minimax lower bound and higher-order estimation when the source population and the target population share the same set of covariates. We derive the minimax lower bound and propose a new higher-order estimator based on second-order influence functions, which attains the minimax lower bound in a broader regime than the doubly robust estimator. To the best of our knowledge, the minimax lower bound and high-order estimation have not been studied under generalization and transportation settings. Finally, we apply our proposed methods to transporting the causal effects of fruit and vegetable intake on adverse pregnancy outcomes from an observational study to *the whole U.S. pregnant female population* as an illustration.

The paper is organized as follows. In Section 2, we introduce the setup and notation. In Section 3, we discuss the identification assumptions to identify the ATE in the target population. Efficiency theory and doubly robust methods are provided in Section 4. We then summarize the minimax lower bounds of the target functionals in Section 5 and derive higher-order estimators in Section 6. Simulation studies are presented in Section 7 to explore finite-sample properties of our methods. In Section 8, we provide a data analysis to illustrate the proposed methods, transporting effects of dietary intake on pregnancy outcomes. Finally, we conclude with a discussion in Section 9. All the proofs and additional details on real data analysis are presented in the supplementary materials [59].

2. PRELIMINARIES

In this section, we first introduce the setup in generalization and transportation setting. We then formalize the treatment effects of interests as statistical functionals with the potential outcome framework [41, 44]. Finally, we introduce some notation that will be useful in presenting our results.

2.1 Data Structure

In the generalization and transportation setting, there are typically two populations, that is, a source population and a target population. The source population is usually the underlying population of a randomized trial or observational study and is defined by enrollment processes and inclusion or exclusion criteria of the study. We assume we observe n_1 source samples

$$\mathcal{D}_1 = \{\mathbf{Z}_i = (\mathbf{X}_i, A_i, Y_i, S_i = 1), 1 \le i \le n_1\}$$

from the source population, where $\mathbf{X}$ is a d-dimensional vector containing all covariates, A is the treament assignment and Y is the outcome. We denote S as an indicator such that $S = 1$ if a subject is in the source population and $S = 0$ otherwise. We are interested in estimating the ATE in a target population without direct access to treatment and outcome information on target samples. The observational unit in the target population is $\mathbf{Z} = (\mathbf{V}, S = 0)$ and the target data set $\mathcal{D}_2$ consists of n_2 such realizations, that is,

$$\mathcal{D}_2 = \{\mathbf{Z}_i = (\mathbf{V}_i, S_i = 0), n_1 + 1 \le i \le n_1 + n_2 = n\},$$

where, importantly, $\mathbf{V} \subseteq \mathbf{X}$ represents partial covariates in $\mathbf{X}$, which may be a strict subset. Here, we slightly abuse the notation and denote both the covariates vector and set of covariates as $\mathbf{V}$ (in the target population) and $\mathbf{X}$ (in the source population). In practice, there may be variables present in the target population but not in the source population. Since the information about the treatment effect in the source population is captured by $\mathbf{X}$, covariates that are not included in $\mathbf{X}$ may not help us identify and estimate the ATE in the target population. Therefore, we do not use these variables to generalize or transport the causal effects and assume $\mathbf{V} \subseteq \mathbf{X}$.

2.2 Effects of Interest

Before defining our estimands of interest, we first discuss differences in the target population definition in the generalization and transportation setting. In the generalization setting, the source population is a subset of the target population. After we collect treatment and outcome information and estimate the ATE in the source population, we hope to *generalize* the causal effects to the whole target population. The most natural design is a trial nested within a cohort of eligible individuals [18]. In such designs, researchers collect covariate information for all individuals, but only collect treatment and outcome information from a subset of them. This setup arises in comprehensive cohort studies [35], where only a few patients consent to randomization in a clinical trial, as well as clinical trials embedded in healthcare systems where all individuals' information is collected routinely but only some of them are included in a trial [20].

In contrast, in the transportation setting, the source population is (at least partly) external to the target population [16]. The source population is different from the target population and source samples may not be representative of the target population. The goal is to *transport* the causal effects from the source population to the target population. This setup arises widely in public policy research, where randomized trial or observational study is conducted in selected samples while the target samples are from administrative databases or surveys and can be very different from the samples enrolled in the study (our data example in Section 8 falls into this category, as will be discussed in more detail shortly).

Now we define the generalization and transportation effects of interest. We use the random variable Y^a to denote the potential (counterfactual) outcome we would have observed had a subject received treatment $A = a$, which may be contrary to the fact Y. For simplicity, we consider binary treatment in this work, where $A = 1$ means treatment and $A = 0$ means control. Then the ATE in the target population is

$$\psi := \mathbb{E}\big[Y^1 - Y^0\big] \tag{1}$$

in the generalization case and

$$\theta := \mathbb{E}\big[Y^1 - Y^0 | S = 0\big] \tag{2}$$

in the transportation case. Note that in the generalization case the source population is part of the target population, and hence we take an expectation over the whole population. However, in the transportation case the source population may not be representative of the target population, and we only take an expectation in the target population (i.e., conditioning on $S = 0$).

Under standard identification assumptions (consistency, positivity [40] and exchangeability [25]), the ATE in the source population $\mathbb{E}[Y^1 - Y^0 | S = 1]$ can be identified and efficiently estimated. Mathematically, we do not have $\mathbb{E}[Y^1 - Y^0 | S = 0] = \mathbb{E}[Y^1 - Y^0 | S = 1] = \mathbb{E}[Y^1 - Y^0]$ in general, especially when the effect modifiers have different distributions in the source and the target population. Hence, we need additional assumptions and novel methodology to estimate the treatment effect in the target population. We define the mean potential outcomes in the target population as

$$\psi_a = \mathbb{E}\big[Y^a\big], \quad \theta_a = \mathbb{E}\big[Y^a | S = 0\big] \tag{3}$$

so ψ_a and θ_a correspond to the generalization and transportation cases, respectively. In this paper, we will focus on the identification and estimation of ψ_a and θ_a. After estimating ψ_a or θ_a, the ATE in the target population can be estimated by taking the difference. Other estimands, such as the risk ratio or odds ratio, can be estimated in a similar manner.

2.3 Nuisance Functions and Other Notation

To present our results concisely, we need the following notation on commonly used nuisance functions:

- The propensity score in the source population is $\pi_a(\mathbf{x}) = \mathbb{P}(A = a | \mathbf{X} = \mathbf{x}, S = 1)$. If the source data set $\mathcal{D}_1$ is from a randomized trial, then $\pi_a(\mathbf{x})$ can be known. Otherwise, we would need to estimate it from $\mathcal{D}_1$.
- The conditional probability of being selected into the source population (commonly referred to as the participation probability) is $\rho(\mathbf{v}) = \mathbb{P}(S = 1 | \mathbf{V} = \mathbf{v})$.
- The conditional mean and variance of the outcomes among subjects receiving treatment $A = a$ in the source population are $\mu_a(\mathbf{x}) = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}, A = a, S = 1]$ and $\sigma_a^2(\mathbf{x}) = \mathrm{Var}(Y | \mathbf{X} = \mathbf{x}, A = a, S = 1)$.
- The function obtained by further regressing μ_a on $\mathbf{V}$ in the source population, $\tau_a(\mathbf{v}) = \mathbb{E}[\mu_a(\mathbf{X}) | \mathbf{V} = \mathbf{v}, S = 1]$.

For a univariate function f on variables $\mathbf{Z}$, we use $\mathbb{P}_n[f(\mathbf{Z})]$ or $\mathbb{P}_n[f]$ to denote the sample average over n samples $\frac{1}{n}\sum_{i=1}^n f(\mathbf{Z}_i)$. For a bivariate function g, we use $\mathbb{U}_n[g(\mathbf{Z}_1, \mathbf{Z}_2)]$ or $\mathbb{U}_n(g)$ to denote the U-statistic measure $\frac{1}{n(n-1)}\sum_{i \neq j} g(\mathbf{Z}_i, \mathbf{Z}_j)$. The Hellinger distance $H^2(\mathbb{P}, \mathbb{Q})$ between two distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$H^2(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\int \left[\sqrt{p(\mathbf{x})} - \sqrt{q(\mathbf{x})}\right]^2 \nu(d\mathbf{x})$$

for a dominating measure ν. We say a function f is s-smooth if it is $\lfloor s \rfloor$ times continuously differentiable with derivatives up to order $\lfloor s \rfloor$ bounded by some constant $C > 0$ and $\lfloor s \rfloor$-order derivatives Hölder continuous, that is,

$$\left|D^\beta f(\mathbf{x}) - D^\beta f(\mathbf{x}')\right| \leq C\left\|\mathbf{x} - \mathbf{x}'\right\|_2^{s - \lfloor s \rfloor}$$

for all $\beta = (\beta_1, \ldots, \beta_d)$ with $\sum_i \beta_i = \lfloor s \rfloor$, where $D^\beta = \frac{\partial^\beta}{\partial x_1^{\beta_1} \ldots \partial x_d^{\beta_d}}$ is the differential operator. Hölder class, denoted by $\mathcal{H}(s)$, is the function class containing all s-smooth functions.

We denote the weighted L_2 norm with weight function w as $\|f\|_w = \sqrt{\int f(\mathbf{z})^2 w(\mathbf{z})\, d\mathbb{P}(\mathbf{z})}$ and when the weight function $w = 1$ we abbreviate the notation as $\|f\|$. In this paper, we mainly use $w = \rho\pi_a$ as a weight function. For a matrix $\mathbf{\Omega}$, we let $\|\mathbf{\Omega}\|$ and $\|\mathbf{\Omega}\|_F$ denote its spectral norm and Frobenius norm, respectively. We write $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for a positive constant C and sufficiently large n.

3. IDENTIFICATION

In this section, we discuss sufficient conditions to identify functionals in (3) from the observable data. These assumptions are generalizations of the identification conditions used in [18] and [17] to the case $\mathbf{V} \subseteq \mathbf{X}$. (Note in Section 3.1 we consider sensitivity analysis and allow several of the following assumptions to be violated.)

ASSUMPTION 3.1. Consistency: $Y = Y^a$ if $A = a$.

ASSUMPTION 3.2. No unmeasured confounding in source: $(Y^0, Y^1) \perp\!\!\!\perp A|\mathbf{X}, S = 1$.

ASSUMPTION 3.3. Treatment positivity in source: $\pi_a(\mathbf{X}) > 0 \quad \mathbb{P}(\cdot|S = 1)$ a.s. for all a.

Assumption 3.1 is also known as stable unit treatment value assumption (SUTVA) and requires no interference between different subjects, that is, the outcome for an individual is not affected by other individuals' treatments. Assumption 3.2 is a standard assumption used to identify average treatment effects. It holds if the source data set comes from a randomized trial or if we collect enough covariates in $\mathbf{X}$ so that the treatment process is completely explained by $\mathbf{X}$. Assumption 3.3, also known as the overlap assumption, has been used in causal inference since [40]. It guarantees that every subject in the source population has a positive probability of receiving each treatment a. With these three assumptions, we are able to identify the average treatment effect in the source population. But we need additional assumptions for generalization and transportation.

ASSUMPTION 3.4. Exchangeability between populations: $S \perp\!\!\!\perp Y^a|\mathbf{V}$.

ASSUMPTION 3.5. Positivity of selection: $\rho(\mathbf{V}) > 0$ a.s.

Assumption 3.4 is critical to generalizing/transporting the effects from the source population to the target population [17, 18, 31]. Under Assumption 3.4, we have

$$\mathbb{E}\big[Y^a|\mathbf{V}, S = 1\big] = \mathbb{E}\big[Y^a|\mathbf{V}\big] = \mathbb{E}\big[Y^a|\mathbf{V}, S = 0\big],$$

which further implies

$$\mathbb{E}\big[Y^1 - Y^0|\mathbf{V}, S = 1\big] = \mathbb{E}\big[Y^1 - Y^0|\mathbf{V}, S = 0\big]. \tag{4}$$

Hence, the source population and the target population have the same conditional average treatment effect. We essentially just need (4) to identify the ATE in the target population. To state our results concisely, we formalize the assumption in terms of potential outcome Y^a instead of the contrast $Y^1 - Y^0$. For equality (4) to hold, all effect modifiers that are distributed differently between the source and the target populations must be measured in $\mathbf{V}$. In practice, researchers can first identify effect modifiers from variables shared by both the source and target populations and include them in $\mathbf{V}$. They can then identify additional covariates in the source population that explain the treatment assignment and include them in $\mathbf{X}$, ensuring that Assumption 3.2 is satisfied. Assumption 3.5 requires that in each stratum of effect modifiers $\mathbf{V}$, there is a positive probability of being in the source population for every individual. Thus all members in the target population are represented by some individuals in the source population.

THEOREM 3.6. *Under identification Assumptions 3.1–3.5, the estimands ψ_a and θ_a are identified as*

$$\begin{aligned}
\psi_a &= \mathbb{E}\big\{\mathbb{E}\big[\mathbb{E}(Y|\mathbf{X}, A = a, S = 1)|\mathbf{V}, S = 1\big]\big\} \\
&= \mathbb{E}\big\{\mathbb{E}\big[\mu_a(\mathbf{X})|\mathbf{V}, S = 1\big]\big\} = \mathbb{E}\big[\tau_a(\mathbf{V})\big], \\
\theta_a &= \mathbb{E}\big\{\mathbb{E}\big[\mathbb{E}(Y|\mathbf{X}, A = a, S = 1)|\mathbf{V}, S = 1\big]|S = 0\big\} \\
&= \mathbb{E}\big\{\mathbb{E}\big[\mu_a(\mathbf{X})|\mathbf{V}, S = 1\big]|S = 0\big\} \\
&= \mathbb{E}\big[\tau_a(\mathbf{V})|S = 0\big].
\end{aligned} \tag{5}$$

We can understand the above functionals by evaluating the three iterative expectations. First, we regress the outcome Y on $\mathbf{X}$ among subjects receiving treatment $A = a$ in the source population and obtain $\mu_a(\mathbf{x})$, which contains information on the conditional treatment effect. Then we further regress μ_a on effect modifiers $\mathbf{V}$ in the source population to obtain $\tau_a(\mathbf{V})$, which summarizes the conditional treatment effects within the subset of covariates $\mathbf{V}$. Finally we take the mean of τ_a in the target population and obtain the target functional. The validity of the last step is guaranteed by Assumption 3.4, which implies the information on treatment effects contained in τ_a generalizes to the whole population. The proof of identifiability is provided in the Appendix.

Note that the mean potential outcome in the source population can be written as

$$\mathbb{E}\big[Y^a|S = 1\big] = \mathbb{E}\big[\tau_a(\mathbf{V})|S = 1\big].$$

It differs from the target functional only in the last step, where we average over τ_a in the target population instead of in the source population. When the distributions of effect modifiers $\mathbf{V}$ in the source population and the target populations are different, we have $\mathbb{E}[\tau_a(\mathbf{V})|S = 1] \neq \mathbb{E}[\tau_a(\mathbf{V})|S = 0]$, and thus the treatment effect in the source population may not generalize to the target population directly.

In applications, one needs to carefully assess the five assumptions above. In general, these assumptions are

untestable and their plausibility needs to be evaluated from substantive knowledge on the mechanism of treatment assignment and study participation. If some assumptions are likely to be violated, researchers should perform sensitivity analysis to assess the robustness of their results. We discuss one way of performing sensitivity analysis in the following section.

3.1 Sensitivity Analysis

When the identification assumptions do not hold simultaneously, it is not guaranteed that identification results (5) hold and we cannot identify the target functionals from the observed data. But we can still derive bounds on them under other assumptions (e.g., we may relax conditions required to identify the target functionals [33] or construct bounds via available instrumental variable [2, 32]). In this section, we relax exchangeability and transportability assumptions to derive bounds that provide us with a range of plausible values for the ATE and can be useful in some applications.

ASSUMPTION 3.7. Relaxation of Assumption 3.2: There exists a positive constant δ_1 such that

$$\big|\mathbb{E}\big[Y^a|\mathbf{X}, A=1, S=1\big]-\mathbb{E}\big[Y^a|\mathbf{X}, A=0, S=1\big]\big|\le\delta_1$$

a.s. for all a.

ASSUMPTION 3.8. Relaxation of Assumption 3.4: There exists a positive constant δ_2 such that

$$\big|\mathbb{E}\big[Y^a|\mathbf{V}, S=0\big]-\mathbb{E}\big[Y^a|\mathbf{V}, S=1\big]\big|\le\delta_2 \quad \text{a.s. for all } a.$$

We note that when Assumption 3.2 and Assumption 3.4 hold, we have $\delta_1=\delta_2=0$. Hence, Assumption 3.7 and Assumption 3.8 are indeed relaxations of Assumption 3.2 and Assumption 3.4. In practice, the value δ_1 and δ_2 may come from the domain knowledge of the problem of interests. The following theorem characterizes the bounds on the ATE in the target population when the treatment is binary.

THEOREM 3.9. *Under Assumption 3.1, 3.3, 3.5, 3.7 and 3.8, we have*

$$\begin{aligned}\psi_a\in\big[&\mathbb{E}\big\{\tau_a(\mathbf{V})\big\}-\delta_2\mathbb{P}(S=0)\\&-\delta_1\mathbb{E}\big\{\mathbb{P}(A=1-a|\mathbf{V}, S=1)\big\}, \mathbb{E}\big\{\tau_a(\mathbf{V})\big\}\\&+\delta_2\mathbb{P}(S=0)+\delta_1\mathbb{E}\big\{\mathbb{P}(A=1-a|\mathbf{V}, S=1)\big\}\big],\\ \theta_a\in\big[&\mathbb{E}\big\{\tau_a(\mathbf{V})|S=0\big\}-\delta_2\\&-\delta_1\mathbb{E}\big\{\mathbb{P}(A=1-a|\mathbf{V}, S=1)|S=0\big\},\\&\mathbb{E}\big\{\tau_a(\mathbf{V})|S=0\big\}\\&+\delta_2+\delta_1\mathbb{E}\big\{\mathbb{P}(A=1-a|\mathbf{V}, S=1)|S=0\big\}\big].\end{aligned}$$

Hence, the ATE in the generalization functional $\psi_1-\psi_0$ is in the interval

$$\begin{aligned}&\big[\mathbb{E}\big\{\tau_1(\mathbf{V})-\tau_0(\mathbf{V})\big\}-\delta_1-2\delta_2\mathbb{P}(S=0),\\&\quad\mathbb{E}\big\{\tau_1(\mathbf{V})-\tau_0(\mathbf{V})\big\}+\delta_1+2\delta_2\mathbb{P}(S=0)\big].\end{aligned}$$

The ATE in the transportation functional $\theta_1-\theta_0$ is in the interval

$$\begin{aligned}&\big[\mathbb{E}\big\{\tau_1(\mathbf{V})-\tau_0(\mathbf{V})|S=0\big\}-\delta_1-2\delta_2,\\&\quad\mathbb{E}\big\{\tau_1(\mathbf{V})-\tau_0(\mathbf{V})|S=0\big\}+\delta_1+2\delta_2\big].\end{aligned}$$

Since the efficiency theory in Section 4.2 holds for $\mathbb{E}[\tau_a(\mathbf{V})]$ and $[\mathbb{E}[\tau_a(\mathbf{V})|S=0]$, one can use the doubly robust estimator to estimate them efficiently. When the specific values of δ_1 and δ_2 are available, one can directly construct the bounds in Theorem 3.9. If exact domain knowledge on the precise value of δ_1 and δ_2 is unavailable, we can estimate the bounds as a function of (δ_1, δ_2) and, for example, obtain which values of δ_1 and δ_2 substantially change results (e.g., flip the sign of the treatment effects). For instance, when we are confident about Assumption 3.2 (e.g., the source data set comes from a randomized experiment), we can set $\delta_1=0$. Then the value for δ_2 changes the sign of the effect is $|\mathbb{E}[\tau_1(\mathbf{V})-\tau_0(\mathbf{V})|S=0]|/2$. This value can reflect the robustness of our results when the identification assumptions do not necessarily hold.

4. EFFICIENCY THEORY AND DOUBLY ROBUST ESTIMATION

In this section, we develop the nonparametric theory for estimation of the ATE in the target population. Namely, we first derive the efficient influence function, together with the nonparametric efficiency bound. The nonparametric efficiency bound provides a benchmark for efficient estimation in a nonparametric model, indicating the best possible performance in a local asymptotic minimax sense [55]. Next, we propose a doubly robust estimator of ψ_a and θ_a based on the influence function, which is shown to be asymptotically normal and attain the efficiency bound under additional mild conditions.

4.1 Efficient Influence Function and Efficiency Bound

We first introduce the problem faced by the plug-in estimator and motivate the study of efficient influence functions. Denote the plug-in estimators of nuisance functions $(\mu_a, \tau_a, \pi_a, \rho)$ as $(\hat{\mu}_a, \hat{\tau}_a, \hat{\pi}, \hat{\rho})$. Based on the identification result for ψ_a after equation (5), a plug-in estimator of ψ_a is then given by

$$\hat{\psi}_a=\mathbb{P}_n(\hat{\tau}_a)=\frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_a(\mathbf{V}_i).$$

This plug-in estimator would be $\sqrt{n}$-consistent if we used correct parametric models to estimate the nuisance function τ_a. However, there is generally not sufficient background knowledge to ensure correct specification of such parametric models; thus analysts often use flexible nonparametric methods to avoid model misspecification. However, under such circumstances, the conditional bias

of the plug-in estimator is of order $\|\hat{\tau}_a - \tau_a\|$, which is typically slower than the $\sqrt{n}$-rate, perhaps much slower when the nuisance functions are complex and the number of covariates is large. Hence, the plug-in estimator generally suffers from slow convergence rates and a lack of tractable limiting distributions. These drawbacks make it difficult to estimate ψ_a accurately and perform statistical inference with plug-in estimators.

To address these difficulties, one can derive the efficient influence functions of the target functionals. The efficient influence function is critical in nonparametric efficiency theory [5, 26, 51, 54, 55]. Mathematically, the influence function is the derivative in a von Mises expansion (i.e., distributional Taylor expansion) of the target statistical functional. In the discrete case, it coincides with the Gateaux derivative of the functional when the contamination distribution is a point-mass. Influence functions are important in the following respects. First, the variance of the influence function is equal to the efficiency bound of the target statistical functional, which characterizes the inherent estimation difficulty of the target functional and provides a benchmark to compare against when we construct estimators. Moreover, it allows us to correct for first-order bias in the plug-in estimator and obtain doubly robust-style estimators, which enjoy appealing statistical properties even if nonparametric methods with relatively slow rates are used in nuisance estimation.

In the following discussions, we will first present the efficient influence function of ψ_a and θ_a. We then derive the doubly robust estimator and establish its $\sqrt{n}$-consistency and asymptotic normality under appropriate conditions. The efficient influence functions and efficiency bounds are summarized in the following results.

LEMMA 4.1. *Under an unrestricted nonparametric model, the efficient influence function of ψ_a is given by*

$$\phi_a^{\mathrm{ge}}(\mathbf{Z}) = \frac{I(A=a,S=1)(Y-\mu_a(\mathbf{X}))}{\rho(\mathbf{V})\pi_a(\mathbf{X})} + \frac{I(S=1)(\mu_a(\mathbf{X})-\tau_a(\mathbf{V}))}{\rho(\mathbf{V})} + \tau_a(\mathbf{V}) - \psi_a$$

and the efficient influence function of θ_a is given by

$$\phi_a^{\mathrm{tr}}(Z) = \frac{1}{\mathbb{P}(S=0)}\Big\{ I(S=0)\big[\tau_a(\mathbf{V})-\theta_a\big] + \frac{I(S=1)(1-\rho(\mathbf{V}))(\mu_a(\mathbf{X})-\tau_a(\mathbf{V}))}{\rho(\mathbf{V})} + \frac{I(A=a,S=1)(1-\rho(\mathbf{V}))(Y-\mu_a(\mathbf{X}))}{\rho(\mathbf{V})\pi_a(\mathbf{X})}\Big\}.$$

THEOREM 4.2. *The nonparametric efficiency bound of ψ_a is*

$$\sigma^2_{a,\mathrm{ge}} = \mathbb{E}\Big[\frac{\mathbb{P}(S=1|\mathbf{X})\operatorname{Var}(Y|\mathbf{X},A=a,S=1)}{\rho^2(\mathbf{V})\pi_a(\mathbf{X})}\Big] + \mathbb{E}\Big[\frac{\operatorname{Var}(\mu_a(\mathbf{X})|\mathbf{V},S=1)}{\rho(\mathbf{V})}\Big] + \operatorname{Var}\big(\tau_a(\mathbf{V})\big)$$

and the nonparametric efficiency bound of θ_a is

$$\sigma^2_{a,\mathrm{tr}} = \frac{1}{\mathbb{P}^2(S=0)}\Big\{\mathbb{P}(S=0)\operatorname{Var}\big(\tau_a(\mathbf{V})|S=0\big) + \mathbb{E}\Big[\frac{\mathbb{P}(S=1|\mathbf{X})(1-\rho(\mathbf{V}))^2\operatorname{Var}(Y|\mathbf{X},A=a,S=1)}{\rho^2(\mathbf{V})\pi_a(\mathbf{X})}\Big] + \mathbb{E}\Big[\frac{(1-\rho(\mathbf{V}))^2\operatorname{Var}(\mu_a(\mathbf{X})|\mathbf{V},S=1)}{\rho(\mathbf{V})}\Big]\Big\}.$$

The efficiency bounds in Theorem 4.2 show how particular nuisance quantities determine the estimation difficulty of our target functionals. Specifically, the efficiency bounds of ψ_a and θ_a both depend on:

- The inverse propensity score $1/\pi_a(\mathbf{X})$ measuring how likely an individual will receive treatment $A=a$.
- The conditional variance $\sigma_a^2(\mathbf{x}) = \operatorname{Var}(Y|\mathbf{X}=\mathbf{x},A=a,S=1)$, which measures how much variation of Y can be explained by $\mathbf{X}$ for subjects receiving treatment a in the source population.
- The conditional variance $\operatorname{Var}(\mu_a(\mathbf{X})|\mathbf{V}=\mathbf{v},S=1)$, which measures how much variation of $\mu_a(\mathbf{X})$ can be explained by $\mathbf{V}$ for subjects in the source population.
- The variance of $\tau_a(\mathbf{V})$ in the target population, that is, $\operatorname{Var}(\tau_a(\mathbf{V}))$ in the generalization case and $\operatorname{Var}(\tau_a(\mathbf{V})|S=0)$ in the transportation case.

There are also some differences in the efficiency bounds of the two functionals. First, the efficiency bound in the transportation case depends on the probability of being in the target population $\mathbb{P}(S=0)$ explicitly. Moreover, the first term in the efficiency bound $\sigma^2_{a,\mathrm{ge}}$ depends on $\rho(\mathbf{V})$ via its reciprocal while the first term in $\sigma^2_{a,\mathrm{tr}}$ depends on the inverse odds of being in the source population $(1-\rho(\mathbf{V}))/\rho(\mathbf{V})$. This implies the inverse odds ratio $(1-\rho(\mathbf{V}))/\rho(\mathbf{V})$ may be a more fundamental quantity than $1/\rho(\mathbf{V})$ in transportation problems, and we will see this phenomenon in Section 6 as well. In addition to parameterizing the inverse odds of being in the source population through ρ, one can also express it as

$$\frac{1-\rho(\mathbf{V})}{\rho(\mathbf{V})} = \frac{\mathbb{P}(S=0)\mathbb{P}(\mathbf{V}|S=0)}{\mathbb{P}(S=1)\mathbb{P}(\mathbf{V}|S=1)}$$

and parameterize the density ratio $\mathbb{P}(\mathbf{V}|S=0)/\mathbb{P}(\mathbf{V}|S=1)$. Then in the estimation step, the inverse odds ratio can be obtained by directly estimating the density ratio [36, 47].

The efficient influence functions in Theorem 4.2 generalize those of [17, 18] to the setting where there can be a mismatch between the covariates $\mathbf{V}$ in the target population, and the covariates $\mathbf{X}$ in the source population. To be concrete, in the special case $\mathbf{V}=\mathbf{X}$ the efficient influence functions of ψ_a and θ_a are

$$\phi_a^{\mathrm{ge}}(\mathbf{Z}) = \frac{I(A=a,S=1)(Y-\mu_a(\mathbf{X}))}{\rho(\mathbf{X})\pi_a(\mathbf{X})} + \mu_a(\mathbf{X}) - \psi_a,$$

$$\phi_a^{\mathrm{tr}}(\mathbf{Z}) = \frac{1}{\mathbb{P}(S=0)}\Big\{I(S=0)\big(\mu_a(\mathbf{X})-\theta_a\big) + \frac{I(A=a,S=1)(1-\rho(\mathbf{X}))(Y-\mu_a(\mathbf{X}))}{\rho(\mathbf{X})\pi_a(\mathbf{X})}\Big\}$$

and the corresponding efficiency bounds are

$$\sigma_{a,\mathrm{ge}}^2 = \mathbb{E}\Big[\frac{\mathrm{Var}(Y|\mathbf{X},A=a,S=1)}{\rho(\mathbf{X})\pi_a(\mathbf{X})}\Big] + \mathrm{Var}\big(\mu_a(\mathbf{X})\big),$$

$$\sigma_{a,\mathrm{tr}}^2 = \frac{1}{\mathbb{P}^2(S=0)}\Big\{\mathbb{P}(S=0)\,\mathrm{Var}\big(\mu_a(\mathbf{X})|S=0\big) + \mathbb{E}\Big[\frac{(1-\rho(\mathbf{X}))^2\,\mathrm{Var}(Y|\mathbf{X},A=a,S=1)}{\rho(\mathbf{X})\pi_a(\mathbf{X})}\Big]\Big\}.$$

These results can be derived separately starting from the functional $\psi_a = \mathbb{E}[\mathbb{E}(Y|\mathbf{X},A=a,S=1)]$ and $\theta_a = \mathbb{E}[\mathbb{E}(Y|\mathbf{X},A=a,S=1)|S=0]$. Alternatively, one can set $\mathbf{V}=\mathbf{X}$ in Lemma 4.1 and Theorem 4.2 to obtain the results by noting the second terms vanish in each formula due to $\mu_a(\mathbf{X}) = \tau_a(\mathbf{V})$ and $\mathrm{Var}(\mu_a(\mathbf{X})|\mathbf{X},S=1)=0$. From this perspective, the second terms in the general case $\mathbf{V}\subset\mathbf{X}$ come from an extra step where we regress the conditional mean $\mu_a(\mathbf{X})$ on possible effect modifiers $\mathbf{V}$.

As we mentioned above, the efficiency bound characterizes the fundamental statistical difficulty of estimating the target functionals, and acts as a nonparametric analog of the Cramer–Rao bound. Specifically, no estimator can have smaller mean square error than the efficiency bound in a local asymptotic minimax sense, as summarized in the following Corollary 4.3.

COROLLARY 4.3. *For any estimators $\hat{\psi}_a$ and $\hat{\theta}_a$, we have*

$$\inf_{\delta>0}\liminf_{n\to\infty}\sup_{\mathbb{Q}:\mathrm{TV}(\mathbb{P},\mathbb{Q})<\delta} n\mathbb{E}_{\mathbb{Q}}\big[\{\hat{\psi}_a-\psi_a(\mathbb{Q})\}^2\big] \geq \sigma_{a,\mathrm{ge}}^2(\mathbb{P}),$$

$$\inf_{\delta>0}\liminf_{n\to\infty}\sup_{\mathbb{Q}:\mathrm{TV}(\mathbb{P},\mathbb{Q})<\delta} n\mathbb{E}_{\mathbb{Q}}\big[\{\hat{\theta}_a-\theta_a(\mathbb{Q})\}^2\big] \geq \sigma_{a,\mathrm{tr}}^2(\mathbb{P}),$$

where $\mathrm{TV}(\mathbb{P},\mathbb{Q})$ *is the total variation distance between* $\mathbb{P}$ *and* $\mathbb{Q}$ *and* $\sigma_{a,\mathrm{ge}}^2(\mathbb{P})$ *and* $\sigma_{a,\mathrm{tr}}^2(\mathbb{P})$ *are the nonparametric efficiency bounds in Theorem 4.2 evaluated at* $\mathbb{P}$.

We have characterized the efficiency bounds with efficient influence functions, which implies that (without further assumptions) the asymptotic local minimax mean squared error of any estimator scaled by a factor of n cannot be smaller than these bounds [55]. The next step is to correct for the first-order bias of plug-in-style estimators by instead deriving doubly robust estimators, as detailed in the following subsection.

4.2 Doubly Robust Estimation

After deriving the influence functions, we can correct for the first-order bias of the plug-in estimator via the following doubly robust estimators:

$$\hat{\psi}_a^{\mathrm{dr}} = \mathbb{P}_n\Big\{\frac{I(A=a,S=1)(Y-\hat{\mu}_a(\mathbf{X}))}{\hat{\rho}(\mathbf{V})\hat{\pi}_a(\mathbf{X})} + \frac{I(S=1)(\hat{\mu}_a(\mathbf{X})-\hat{\tau}_a(\mathbf{V}))}{\hat{\rho}(\mathbf{V})} + \hat{\tau}_a(\mathbf{V})\Big\}, \tag{6}$$

$$\hat{\theta}_a^{\mathrm{dr}} = \frac{1}{\widehat{\mathbb{P}}(S=0)}\mathbb{P}_n\Big\{I(S=0)\hat{\tau}_a(\mathbf{V}) + \frac{I(A=a,S=1)(1-\hat{\rho}(\mathbf{V}))(Y-\hat{\mu}_a(\mathbf{X}))}{\hat{\rho}(\mathbf{V})\hat{\pi}_a(\mathbf{X})} + \frac{I(S=1)(1-\hat{\rho}(\mathbf{V}))(\hat{\mu}_a(\mathbf{X})-\hat{\tau}_a(\mathbf{V}))}{\hat{\rho}(\mathbf{V})}\Big\}. \tag{7}$$

The doubly robust estimators combine simple plug-in-style estimators with inverse-probability-weighted estimators to correct for the first-order bias. For instance, in the doubly robust estimator $\hat{\psi}_a^{\mathrm{dr}}$ the last term $\mathbb{P}_n[\hat{\tau}_a(\mathbf{V})]$ is the outcome regression-based plug-in estimator and the first two terms are centered inverse-probability-weighted terms motivating from the influence function in Lemma 4.1. Compared with the well-known doubly robust estimator for the ATE

$$\mathbb{P}_n\Big\{\frac{I(A=a)(Y-\hat{\mu}_a(\mathbf{X}))}{\hat{\pi}_a(\mathbf{X})}+\hat{\mu}_a(\mathbf{X})\Big\},$$

in the generalization and transportation setting the participation probability also needs to be modeled and incorporated into the reweighting terms. Moreover, an extra term appears in (6) and (7) due to further regressing $\mu_a(\mathbf{X})$ on $\mathbf{V}$, as similarly discussed in Section 4.1.

For simplicity, we define the uncentered influence function terms in the brackets above as

$$\varphi_a^{\mathrm{ge}}(\mathbf{Z}) = \frac{I(A=a,S=1)(Y-\mu_a(\mathbf{X}))}{\rho(\mathbf{V})\pi_a(\mathbf{X})} + \frac{I(S=1)(\mu_a(\mathbf{X})-\tau_a(\mathbf{V}))}{\rho(\mathbf{V})}+\tau_a(\mathbf{V}),$$

$$\varphi_a^{\mathrm{tr}}(\mathbf{Z}) = \frac{I(A=a,S=1)(1-\rho(\mathbf{V}))(Y-\mu_a(\mathbf{X}))}{\rho(\mathbf{V})\pi_a(\mathbf{X})} + \frac{I(S=1)(1-\rho(\mathbf{V}))(\mu_a(\mathbf{X})-\tau_a(\mathbf{V}))}{\rho(\mathbf{V})} + I(S=0)\tau_a(\mathbf{V}).$$

The following theorems characterize the properties of these new doubly robust estimators.

THEOREM 4.4 (Doubly robust estimation of generalization functional). *Suppose the nuisance functions* $(\hat{\mu}_a,\hat{\tau}_a,\hat{\pi}_a,\hat{\rho})$ *are estimated from a separate independent sample. Further assume our estimates satisfy* $\|\hat{\varphi}_a^{\mathrm{ge}} - \varphi_a^{\mathrm{ge}}\|_2 = o_{\mathbb{P}}(1)$, *and* $\hat{\rho}(\mathbf{V}),\hat{\pi}_a(\mathbf{X}) \geq \epsilon > 0$ *for some positive constant* ϵ. *Then we have*

$$\hat{\psi}_a^{\mathrm{dr}} - \psi_a = \mathbb{P}_n\big(\phi_a^{\mathrm{ge}}\big) + O_{\mathbb{P}}\big(\|\hat{\mu}_a-\mu_a\|\|\hat{\pi}_a-\pi_a\| + \|\hat{\rho}-\rho\|\|\hat{\tau}_a-\tau_a\|\big) + o_{\mathbb{P}}(1/\sqrt{n}).$$

If the nuisance estimators further satisfy the following convergence rate:

$$\|\hat{\mu}_a - \mu_a\|\|\hat{\pi}_a - \pi_a\| = o_{\mathbb{P}}(1/\sqrt{n}),$$
$$\|\hat{\rho} - \rho\|\|\hat{\tau}_a - \tau_a\| = o_{\mathbb{P}}(1/\sqrt{n}),$$

then $\hat{\psi}_a^{\mathrm{dr}}$ *is* $\sqrt{n}$*-consistent and asymptotically normal, with asymptotic variance equal to the nonparametric efficiency bound* $\sigma^2_{a,\mathrm{ge}}$ *of Theorem* 4.2*, and so also locally asymptotic minimax optimal in the sense of Corollary* 4.3.

THEOREM 4.5 (Doubly robust estimation of transportation functional). *Suppose the nuisance functions* $\hat{\mu}_a$, $\hat{\tau}_a$, $\hat{\pi}_a$, $\hat{\rho}$ *are estimated from a separate independent sample. Further assume our estimates satisfy* $\|\hat{\varphi}_a^{\mathrm{tr}} - \varphi_a^{\mathrm{tr}}\|_2 = o_{\mathbb{P}}(1)$, $\mathbb{P}(S=0) > 0$, $\hat{\rho}(\mathbf{V})$, $\hat{\pi}_a(\mathbf{X}) \geq \epsilon > 0$ *for some positive constant* ϵ*. Then we have*

$$\hat{\theta}_a^{\mathrm{dr}} - \theta_a = \mathbb{P}_n(\phi_a^{\mathrm{tr}}) + O_{\mathbb{P}}\big(\|\hat{\mu}_a - \mu_a\|\|\hat{\pi}_a - \pi_a\| + \|\hat{\rho} - \rho\|\|\hat{\tau}_a - \tau_a\|\big) + o_{\mathbb{P}}(1/\sqrt{n}).$$

If the nuisance estimators further satisfy the following convergence rate:

$$\|\hat{\mu}_a - \mu_a\|\|\hat{\pi}_a - \pi_a\| = o_{\mathbb{P}}(1/\sqrt{n}),$$
$$\|\hat{\rho} - \rho\|\|\hat{\tau}_a - \tau_a\| = o_{\mathbb{P}}(1/\sqrt{n}),$$

then $\hat{\theta}_a^{\mathrm{dr}}$ *is* $\sqrt{n}$*-consistent and asymptotically normal, with asymptotic variance equal to the nonparametric efficiency bound* $\sigma^2_{a,\mathrm{tr}}$ *of Theorem* 4.2*, and so also locally asymptotic minimax optimal in the sense of Corollary* 4.3.

REMARK 1. For simplicity, we assume all the nuisance estimators are constructed from a separate independent sample with the same size n as the estimation sample over which $\mathbb{P}_n$ takes an average. Using the same sample to both estimate the nuisance functions and average the (uncentered) influence functions may also yield similar results by further assuming empirical process assumptions to avoid overfitting. For instance, one can assume the nuisance functions and their estimates belong to a Donsker class and arrive at similar estimation guarantees. However, such assumptions are hard to verify in practice and sample splitting enables us to get rid of them: one can randomly split the data in folds and use different folds to estimate the nuisance functions and average the influence functions. To recover full sample size efficiency, one can swap the folds, repeat the same procedures and finally average the results, known as cross-fitting and commonly used in the literature [6, 15, 28, 37, 61]. In this paper, all the results are based on a single split procedure, with the understanding that extending to procedures based on cross-fitting is straightforward.

We note that in Theorems 4.4 and 4.5 we do not require that each individual nuisance function converges at $\sqrt{n}$-rate as might be required in the plug-in estimator case. The condition is instead on the product of convergence rates, that is, $\|\hat{\mu}_a - \mu_a\|\|\hat{\pi}_a - \pi_a\| = o_{\mathbb{P}}(1/\sqrt{n})$ and $\|\hat{\tau}_a - \tau_a\|\|\hat{\rho} - \rho\| = o_{\mathbb{P}}(1/\sqrt{n})$. This shows the key property of doubly robust estimators: after we correct for the first-order bias, the error only involves second-order products, and hence is "doubly small." In applications, such conditions on the convergence rate are much easier to satisfy. Estimators like ours that have errors involving multiple nuisance functions are sometimes referred to as multiply robust estimators [49], since there are multiple ways in which the error term is $o_{\mathbb{P}}(1/\sqrt{n})$. For instance, (1) quarter rate $\|\hat{\mu}_a - \mu_a\| = o_{\mathbb{P}}(n^{-1/4})$ and $\|\hat{\pi}_a - \pi_a\| = O_{\mathbb{P}}(n^{-1/4})$ or (2) $\|\hat{\mu}_a - \mu_a\| = o_{\mathbb{P}}(1)$ and $\|\hat{\pi}_a - \pi_a\| = O_{\mathbb{P}}(n^{-1/2})$ (e.g., we know exactly the parametric model for propensity score) both satisfy the condition; further, $n^{-1/4}$-style rates can be attained under appropriate smoothness, sparsity or other structural assumptions. So, we may apply flexible nonparametric methods (e.g., random forests) or high-dimensional models (e.g., Lasso regression) to estimate the nuisance functions and still maintain the $\sqrt{n}$-consistency and asymptotic normality of our effect estimator; this is the main advantage of doubly robust estimators over plug-ins. In the special case $\mathbf{V} = \mathbf{X}$ (i.e., the source and the target population share the same sets of covariates), conditions in Theorems 4.4 and Theorem 4.5 are reduced to $\|\hat{\mu}_a - \mu_a\|\|\hat{\pi}_a - \pi_a\| = o_{\mathbb{P}}(1/\sqrt{n})$ and $\|\hat{\mu}_a - \mu_a\|\|\hat{\rho} - \rho\| = o_{\mathbb{P}}(1/\sqrt{n})$. Compared with the ATE case where we only require $\|\hat{\mu}_a - \mu_a\|\|\hat{\pi}_a - \pi_a\| = o_{\mathbb{P}}(1/\sqrt{n})$, extra conditions on the convergence rates of modeling participation probability is needed in the generalization and transportation problem.

We conclude this section with additional comments on estimating $\tau_a(\mathbf{V})$. In practice, a simple plug-in approach is to first estimate $\mu_a(\mathbf{X})$ for each data point and then further regress $\hat{\mu}_a$ on partial covariates $\mathbf{V}$ in the source data set, known as regression with estimated or imputed outcomes [21, 27]. However, the estimation error of τ_a is closely tied to that of μ_a. When $\hat{\mu}_a$ suffers from a slow convergence rate, $\hat{\tau}_a$ may inherit this slow rate as well. Faster convergence rates of estimating $\tau_a(\mathbf{V})$ can be achieved by adopting the stability framework in [27] or orthogonal statistical learning framework in [21]. For example, instead of directly regressing $\hat{\mu}_a(\mathbf{X})$ on $\mathbf{V}$, one can estimate μ_a and π_a, and construct a pseudo-outcome

$$\hat{g}(\mathbf{Z}) = \frac{I(A=a)(Y - \hat{\mu}_a(\mathbf{X}))}{\hat{\pi}_a(\mathbf{X})} + \hat{\mu}_a(\mathbf{X})$$

for each data point in the source data set. Similar to how doubly robust estimators reduce bias in estimating the ATE, regressing $\hat{g}(\mathbf{Z})$ on $\mathbf{V}$ yields an estimator of τ_a with an error rate depending on the product of the estimation rates for $\hat{\mu}_a$ and $\hat{\pi}_a$ (under appropriate conditions from [27]). This estimator offers a faster convergence rate than naive plug-in approaches and can achieve the oracle estimation rate for τ_a in a broader regime. For further details, we refer readers to [27].

5. MINIMAX LOWER BOUNDS

In Section 4, we established local asymptotic minimax optimality of doubly robust estimators (under certain conditions). In this section, we will examine the minimax lower bounds from a global perspective, that is, the minimax rate over suitable model classes, more generally when parametric $\sqrt{n}$-rates are not attainable. The minimax rate (in terms of mean-squared-error) of a statistical functional $\theta = \theta(\mathbb{P})$ over model class $\mathcal{P}$ is defined as

$$\inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \big(\mathbb{E}_{\mathbb{P}}\big(\hat{\theta} - \theta(\mathbb{P})\big)^2\big),$$

where the infimum ranges over all possible estimators. It provides an important benchmark to compare against in constructing estimators. If one estimator has estimation error guarantees matching the minimax rate, then one may stop searching for estimators that can achieve a smaller estimation error and conclude the estimator is optimal in terms of worst-case rates. If the minimax rate is not achieved, one may study alternative estimators with smaller statistical error or study sharper bounds for the problem. In this section, we derive the fundamental minimax lower bounds in estimating ATE in the target population under the special case $\mathbf{V} = \mathbf{X}$ (so the source population and target population share the same set of covariates). We introduce the ideas and techniques that are useful in deriving minimax rates in Appendix A. The minimax lower bounds for estimating the ATE in the target population, derived using these techniques, are presented in the remainder of this section.

5.1 Minimax Lower Bounds in Generalization and Transportation

In the generalization and transportation setting, we consider the target functionals in the case $\mathbf{V} = \mathbf{X}$. When the identification assumptions hold, we can identify the effects as

$$\begin{aligned}\psi_a &= \mathbb{E}\big[Y^a\big] = \mathbb{E}\big\{\mathbb{E}[Y|\mathbf{X}, S=1, A=a]\big\},\\ \theta_a &= \mathbb{E}\big[Y^a|S=0\big]\\ &= \mathbb{E}\big\{\mathbb{E}[Y|\mathbf{X}, S=1, A=a]|S=0\big\}.\end{aligned} \tag{8}$$

We restrict the range of covariates $\mathbf{X}$ as $[0,1]^d$ in this section. Consider the following model class for the generalization and transportation functionals, respectively:

$$\begin{aligned}\mathcal{P}_{\mathrm{ge}} = \Big\{&(f, \rho, \pi_a, \mu_a) : \frac{1}{\rho\pi_a} \text{ is } \alpha\text{-smooth},\\ &\mu_a \text{ is } \beta\text{-smooth},\\ &f\rho\pi_a = 1/2, \rho\pi_a \text{ and } \mu_a \text{ are bounded away from}\\ &0 \text{ and } 1\Big\},\end{aligned}$$

$$\begin{aligned}\mathcal{P}_{\mathrm{tr}} = \Big\{&(f, \rho, \pi_a, \mu_a) : \frac{1-\rho}{\rho\pi_a} \text{ is } \alpha\text{-smooth},\\ &\mu_a \text{ is } \beta\text{-smooth},\\ &f\rho\pi_a = 1/2, \rho\pi_a \text{ and } \mu_a \text{ are bounded away}\\ &\text{from 0 and 1}\Big\}.\end{aligned}$$

Here, f is the density of covariates $\mathbf{X}$. We note that the selection/treatment probability $\rho\pi_a$ is jointly parameterized in the model class $\mathcal{P}_{\mathrm{ge}}$. In other words, one can intuitively view $I(S=1, A=a)$ (i.e., both selected in the source population and receive treatment a) as a new treatment. The probability of getting treated under this "compound" treatment is exactly $\rho\pi_a$. One minor difference between $\mathcal{P}_{\mathrm{ge}}$ and $\mathcal{P}_{\mathrm{tr}}$ is that in $\mathcal{P}_{\mathrm{ge}}$ we impose smoothness conditions on $1/\rho\pi_a$ while in $\mathcal{P}_{\mathrm{tr}}$ the smoothness condition is imposed on $(1-\rho)/\rho\pi_a$. As pointed out in the discussion on efficiency bounds in Section 4.1 and quadratic estimation in Section 6, the inverse odds ratio $(1-\rho)/\rho$ turns out to be a more fundamental quantity in transportation problems. To be consistent with these results, we also impose a smoothness condition on the inverse odds ratio in deriving the minimax rate.

The following theorem characterizes the minimax lower bounds in estimating generalization and transportation functionals. The strategy to construct two distribution classes $\mathbb{P}_{\boldsymbol{\lambda}}$ and $\mathbb{Q}_{\boldsymbol{\lambda}}$ is similar to the techniques in proving the minimax rate for ATE, as in [39].

THEOREM 5.1. *Let $s = (\alpha+\beta)/2$ denote the average smoothness of the nuisance functions. The minimax rate of generalization functional ψ_a over $\mathcal{P}_{\mathrm{ge}}$ is lower bounded by*

$$\inf_{\hat{\psi}_a} \sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{ge}}} \big(\mathbb{E}_{\mathbb{P}}(\hat{\psi}_a - \psi_a)^2\big)^{1/2} \gtrsim \begin{cases} n^{-1/(1+d/4s)} & \text{if } s < d/4,\\ n^{-1/2} & \text{otherwise.}\end{cases}$$

The minimax rate of transportation functional θ_a over $\mathcal{P}_{\mathrm{tr}}$ is lower bounded by

$$\inf_{\hat{\theta}_a} \sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{tr}}} \big(\mathbb{E}_{\mathbb{P}}(\hat{\theta}_a - \theta_a)^2\big)^{1/2} \gtrsim \begin{cases} n^{-1/(1+d/4s)} & \text{if } s < d/4,\\ n^{-1/2} & \text{otherwise.}\end{cases}$$

Although the minimax lower bounds above are the same as those for the ATE, we believe these results are not trivial, and presenting them helps us gain a more complete and comprehensive insight into generalization and transportation functionals. Following the efficiency theory for the ATE, the doubly robust estimator of ψ_a and θ_a may achieve the minimax rate under some regimes. For instance, under the conditions in Theorem 4.4 and assuming $\|\hat{\pi}_a\hat{\rho} - \pi\rho\|\|\hat{\mu}_a - \mu_a\| = o_{\mathbb{P}}(1/\sqrt{n})$, the doubly robust estimator is $\sqrt{n}$-consistent and attains the minimax rate (this corresponds to the smooth regime $s \geq d/4$). However, in the less smooth regime $s < d/4$, even if we

can estimate $\rho\pi_a$ at a minimax rate (i.e., $\|\hat{\pi}_a\hat{\rho} - \pi\rho\| = O_{\mathbb{P}}(n^{-\frac{\alpha}{2\alpha+d}})$ and $\|\hat{\mu}_a - \mu_a\| = O_{\mathbb{P}}(n^{-\frac{\beta}{2\beta+d}})$), the necessary condition for the doubly robust estimator to achieve the rate $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$ is

$$\frac{\alpha}{2\alpha+d} + \frac{\beta}{2\beta+d} \geq \frac{2\alpha+2\beta}{2\alpha+2\beta+d},$$

which can be a restrictive condition and motivates us to improve the doubly robust estimator. The main potential drawback of the doubly robust estimator is that we only correct for the first-order bias of the plug-in estimator. In the following section, we propose a higher-order/quadratic estimator based on second-order von Mises expansion [38], which also takes second-order bias into consideration and achieves the minimax lower bound in a broader regime.

6. HIGHER-ORDER ESTIMATION

In this section we study a higher-order (quadratic) estimator of the functionals of interest. An introduction to quadratic Von Mises calculus is provided in Appendix B. We apply the second-order Von Mises expansion to generalization and transportation functionals in the case $\mathbf{V} = \mathbf{X}$ and propose quadratic estimators of them in Section 6.1. The error guarantees of our estimators are then established under mild assumptions.

6.1 Quadratic Estimators in Generalization and Transportation

We now present our higher-order estimator. We first define some objects that will be useful in presenting the theory of this section. Namely, let

$$\boldsymbol{\Omega} = \int \mathbf{b}(\mathbf{x})\mathbf{b}(\mathbf{x})^\top \rho(\mathbf{x})\pi_a(\mathbf{x})\, dF(\mathbf{x}),$$

$$\widehat{\boldsymbol{\Omega}} = \int \mathbf{b}(\mathbf{x})\mathbf{b}(\mathbf{x})^\top \hat{\rho}(\mathbf{x})\hat{\pi}_a(\mathbf{x})\, d\hat{F}(\mathbf{x}),$$

$$\Pi_{\mathbf{b}}(g)(\mathbf{x}) = \mathbf{b}(\mathbf{x})^\top \boldsymbol{\Omega}^{-1} \int \mathbf{b}(\mathbf{u})g(\mathbf{u})\rho(\mathbf{u})\pi_a(\mathbf{u})\, dF(\mathbf{u}),$$

$$\widehat{\Pi}_{\mathbf{b}}(g)(\mathbf{x}) = \mathbf{b}(\mathbf{x})^\top \widehat{\boldsymbol{\Omega}}^{-1} \int \mathbf{b}(\mathbf{u})g(\mathbf{u})\rho(\mathbf{u})\pi_a(\mathbf{u})\, dF(\mathbf{u}).$$

Here, $\mathbf{b} : \mathbb{R}^d \mapsto \mathbb{R}^k$ is a k-dimensional basis (k corresponds to the dimension of the subspace L in Section B). $\boldsymbol{\Omega}$ is the Gram matrix of basis $\mathbf{b}$ if we define the inner product between two functions f, g as $\langle f, g\rangle = \int f(\mathbf{x})g(\mathbf{x})\rho(\mathbf{x})\pi_a(\mathbf{x})\, dF(\mathbf{x})$. $\widehat{\boldsymbol{\Omega}}$ is the estimated Gram matrix where all nuisance functions are replaced with their estimators. $\Pi_{\mathbf{b}}(g)$ is the projection of g onto the linear span of $\mathbf{b}$ and $\widehat{\Pi}_{\mathbf{b}}(g)$ is the estimated projection of g where the gram matrix $\boldsymbol{\Omega}$ is replaced with its estimator $\widehat{\boldsymbol{\Omega}}$. With this notation, the noncentered first-order and (approximate) second-order influence functions of the generalization functional are

$$\phi^{\mathrm{ge}}_{a,1}(\mathbf{Z}) = \frac{I(S=1, A=a)}{\rho(\mathbf{X})\pi_a(\mathbf{X})}\big(Y - \mu_a(\mathbf{X})\big) + \mu_a(\mathbf{X}),$$

$$\begin{aligned}\phi^{\mathrm{ge}}_{a,2}(\mathbf{Z}_1, \mathbf{Z}_2) = &-I(S_1=1, A_1=a)\big(Y_1 - \mu_a(\mathbf{X}_1)\big) \\ &\times \mathbf{b}(\mathbf{X}_1)^\top \boldsymbol{\Omega}^{-1}\mathbf{b}(\mathbf{X}_2) \\ &\times \frac{I(S_2=1, A_2=a) - \rho(\mathbf{X}_2)\pi_a(\mathbf{X}_2)}{\rho(\mathbf{X}_2)\pi_a(\mathbf{X}_2)}.\end{aligned}$$

Following the discussions in Section B, a quadratic estimator is

$$\hat{\psi}^{\mathrm{qr}}_a = \mathbb{P}_n\big[\hat{\phi}^{\mathrm{ge}}_{a,1}(\mathbf{Z})\big] + \mathbb{U}_n\big[\hat{\phi}^{\mathrm{ge}}_{a,2}(\mathbf{Z}_1, \mathbf{Z}_2)\big],$$

where all the nuisance functions and $\boldsymbol{\Omega}$ are replaced with their estimators. Note that the plug-in estimator $\psi(\hat{\mathbb{P}})$ in the general quadratic estimator (10) cancels the centralization term in the centered first-order influence function $\phi_1(\mathbf{Z}, \hat{\mathbb{P}})$. Hence, in $\hat{\psi}^{\mathrm{qr}}_a$ (where $\phi^{\mathrm{ge}}_{a,1}(\mathbf{Z})$ is the noncentered influence function) the plug-in term disappears. In the following discussions, we first present a general theorem summarizing the conditional bias and variance of the quadratic estimator $\hat{\psi}^{\mathrm{qr}}_a$, without assuming any conditions on the convergence rates of nuisance estimation. Then we examine the assumptions needed for the quadratic estimator $\hat{\psi}^{\mathrm{qr}}_a$ to achieve the minimax optimal rate.

THEOREM 6.1 (Quadratic estimation of generalization functional). *Assume all the nuisance functions $\hat{\mu}_a$, $\hat{\rho}$, $\hat{\pi}_a$, $\hat{F}$ are estimated from a separate training sample D^n. Further assume that $\rho, \pi, \hat{\rho}, \hat{\pi}$ are all bounded away from zero, the eigenvalues of $\boldsymbol{\Omega}$, $\widehat{\boldsymbol{\Omega}}$ are bounded away from zero and infinity. Then the conditional bias and variance of $\hat{\psi}^{\mathrm{qr}}_a$ (giving the training data to estimate the nuisance functions) are bounded as*

$$\begin{aligned}&\big|\mathbb{E}\big[\hat{\psi}^{\mathrm{qr}}_a \big| D^n\big] - \psi_a\big| \\ &\quad\lesssim \left\|(\mathbf{I} - \Pi_{\mathbf{b}})\left(\frac{1}{\hat{\rho}\hat{\pi}_a} - \frac{1}{\rho\pi_a}\right)\right\|_w \big\|(\mathbf{I} - \Pi_{\mathbf{b}})(\hat{\mu}_a - \mu_a)\big\|_w \\ &\qquad + \left\|\frac{1}{\hat{\rho}\hat{\pi}_a} - \frac{1}{\rho\pi_a}\right\|_w \|\hat{\mu}_a - \mu_a\|_w \big\|\widehat{\boldsymbol{\Omega}}^{-1} - \boldsymbol{\Omega}^{-1}\big\|,\end{aligned}$$

$$\mathrm{Var}\big(\hat{\psi}^{\mathrm{qr}}_a \big| D^n\big) \lesssim \frac{1}{n} + \frac{k}{n^2},$$

where the weighted L_2 norm of a function g is $\|g\|^2_w = \int g^2(\mathbf{x})\pi_a(\mathbf{x})\rho(\mathbf{x})\, dF(\mathbf{x})$.

The boundedness assumption on the eigenvalues of $\boldsymbol{\Omega}$ can be implied by boundedness of ρ, π_a and $\frac{dF}{d\nu}$ (the density of F with respect to an underlying measure ν) together with the assumption that $\int \mathbf{b}(\mathbf{x})\mathbf{b}(\mathbf{x})^\top d\nu(\mathbf{x})$ is positive definite. See Proposition 2.1 in [4] and Proposition 8 in [29] for more detailed discussions. From Theorem 6.1,

we see the conditional bias is mainly composed of two parts. The first term

$$\left\|(\mathbf{I}-\Pi_{\mathbf{b}})\left(\frac{1}{\hat{\rho}\hat{\pi}_a}-\frac{1}{\rho\pi_a}\right)\right\|_w\|(\mathbf{I}-\Pi_{\mathbf{b}})(\hat{\mu}_a-\mu_a)\|_w$$

is the approximation error of the projection $\Pi_{\mathbf{b}}$ and corresponds to the representational error discussed in Section B. The second term

$$\left\|\frac{1}{\hat{\rho}\hat{\pi}_a}-\frac{1}{\rho\pi_a}\right\|_w\|\hat{\mu}_a-\mu_a\|_w\|\widehat{\mathbf{\Omega}}^{-1}-\mathbf{\Omega}^{-1}\|$$

is a third-order error term and corresponds to the remainder term $R_3(\hat{\mathbb{P}},\mathbb{P})$ discussed in Section B. There is also an extra term k/n^2 in the conditional variance of the proposed higher-order estimator $\hat{\psi}_a^{\mathrm{qr}}$. Hence, we need to carefully select k to balance the representational error, third-order error term and the extra term in the variance.

Note that we need some approximation guarantees of the basis $\mathbf{b}$ to ensure that the representational error is small. Concretely, we impose the following uniform approximation assumption on $\mathbf{b}$.

ASSUMPTION 6.2. For any $s>0$ and $g\in\mathcal{H}(s)$, the basis $\mathbf{b}$ satisfies

$$\|(\mathbf{I}-\Pi_{\mathbf{b}})g\|_w\lesssim k^{-s/d}.$$

Assumption 6.2 holds for a wide class of basis functions. For instance, when the function class containing g is supported on a convex and compact subset of $\mathbb{R}^d$, the approximation is valid even with the uniform norm $\|\cdot\|_\infty$ when the basis uses spline, CDV wavelet or local polynomial partition series.

Under the conditions in Theorem 6.1, if we assume $1/\rho\pi_a, 1/\hat{\rho}\hat{\pi}_a\in\mathcal{H}(\alpha)$ and $\hat{\mu}_a,\mu_a\in\mathcal{H}(\beta)$, by the approximation property of the basis $\mathbf{b}$, we have

$$\left\|(\mathbf{I}-\Pi_{\mathbf{b}})\left(\frac{1}{\hat{\rho}\hat{\pi}_a}-\frac{1}{\rho\pi_a}\right)\right\|_w\lesssim k^{-\alpha/d},$$
$$\|(\mathbf{I}-\Pi_{\mathbf{b}})(\hat{\mu}_a-\mu_a)\|_w\lesssim k^{-\beta/d}.$$

In the less smooth regime $s=(\alpha+\beta)/2<d/4$, set $k\sim n^{2d/(d+2\alpha+2\beta)}$ to balance the representational error and variance. Further assume

$$\left\|\frac{1}{\hat{\rho}\hat{\pi}_a}-\frac{1}{\rho\pi_a}\right\|_w\|\hat{\mu}_a-\mu_a\|_w\|\widehat{\mathbf{\Omega}}^{-1}-\mathbf{\Omega}^{-1}\|\lesssim n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}.$$

Then we see that this estimator can achieve the minimax lower bound. In the smooth regime $s=(\alpha+\beta)/2\geq d/4$, we can set k similarly and assume

$$\left\|\frac{1}{\hat{\rho}\hat{\pi}_a}-\frac{1}{\rho\pi_a}\right\|_w\|\hat{\mu}_a-\mu_a\|_w\|\widehat{\mathbf{\Omega}}^{-1}-\mathbf{\Omega}^{-1}\|\lesssim n^{-1/2},$$

and again this estimator achieves the minimax lower bound. Compared with the condition required for the doubly robust estimator to achieve the minimax lower bound

$$\|\hat{\rho}\hat{\pi}_a-\rho\pi_a\|\|\hat{\mu}_a-\mu_a\|=o_{\mathbb{P}}(n^{-1/2}),$$

since we have a third-order error term for the quadratic estimator, our higher-order/quadratic estimator achieves the minimax lower bound in a broader regime.

For the transportation functional θ_a, we consider a related functional

$$\begin{aligned}\eta_a&=\mathbb{E}\big[I(S=0)\mathbb{E}(Y|\mathbf{X},A=a,S=1)\big]\\&=\mathbb{E}\big[I(S=0)\mu_a(\mathbf{X})\big]\\&=\mathbb{P}(S=0)\theta_a.\end{aligned}$$

Since $\mathbb{P}(S=0)$ can be estimated at $\sqrt{n}$-rate, to estimate θ_a at a minimax optimal rate we only need to estimate η_a at an optimal rate. The first-order and approximate second-order influence functions for η_a are

$$\begin{aligned}\phi_{a,1}^{\mathrm{tr}}(\mathbf{Z})&=\frac{I(S=1,A=a)(1-\rho(\mathbf{X}))}{\rho(\mathbf{X})\pi_a(\mathbf{X})}(Y-\mu_a(\mathbf{X}))\\&\quad+I(S=0)\mu_a(\mathbf{X}),\\\phi_{a,2}^{\mathrm{tr}}(\mathbf{Z}_1,\mathbf{Z}_2)&\\&=I(S_1=1,A_1=a)(Y_1-\mu_a(\mathbf{X}_1))\mathbf{b}(\mathbf{X}_1)^\top\mathbf{\Omega}^{-1}\mathbf{b}(\mathbf{X}_2)\\&\quad\times\frac{I(S_2=0)\rho(\mathbf{X}_2)\pi_a(\mathbf{X}_2)-(1-\rho(\mathbf{X}_2))I(A_2=a,S_2=1)}{\rho(\mathbf{X}_2)\pi_a(\mathbf{X}_2)}.\end{aligned}$$

The quadratic estimator is

$$\hat{\eta}_a^{\mathrm{qr}}=\mathbb{P}_n\big[\hat{\phi}_{a,1}^{\mathrm{tr}}(\mathbf{Z})\big]+\mathbb{U}_n\big[\hat{\phi}_{a,2}^{\mathrm{tr}}(\mathbf{Z}_1,\mathbf{Z}_2)\big].$$

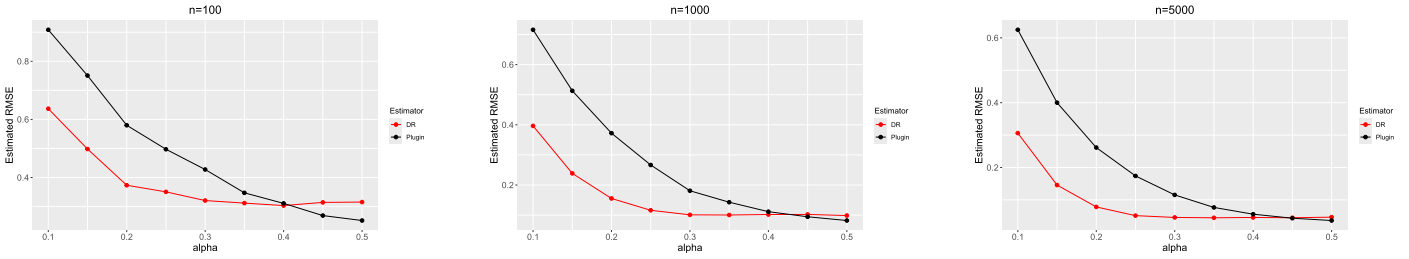
The following theorem, which is the analogous version of Theorem 6.1 in the transportation setting, summarizes the estimation guarantee of the quadratic estimator $\hat{\eta}_a^{\mathrm{qr}}$.

THEOREM 6.3 (Quadratic estimation of transportation functional). *Assume all the nuisance functions $\hat{\mu}_a$, $\hat{\rho}$, $\hat{\pi}_a$, $\hat{F}$ are estimated from a separate training sample D^n. Further assume that $\rho,\pi,\hat{\rho},\hat{\pi}$ are all bounded away from zero, and the eigenvalues of $\mathbf{\Omega}$, $\widehat{\mathbf{\Omega}}$ are bounded away from zero and infinity. The conditional bias and variance of $\hat{\eta}_a$ (giving the training data to estimate the nuisance functions) are bounded as*

$$\begin{aligned}&|\mathbb{E}\big[\hat{\eta}_a^{\mathrm{qr}}|D^n\big]-\eta_a|\\&\lesssim\left\|(\mathbf{I}-\Pi_{\mathbf{b}})\left(\frac{1-\hat{\rho}}{\hat{\rho}\hat{\pi}_a}-\frac{1-\rho}{\rho\pi_a}\right)\right.\\&\quad\times\left\|_w\|(\mathbf{I}-\Pi_{\mathbf{b}})(\hat{\mu}_a-\mu_a)\|_w\right.\\&\quad+\left\|\frac{1-\hat{\rho}}{\hat{\rho}\hat{\pi}_a}-\frac{1-\rho}{\rho\pi_a}\right\|_w\|\hat{\mu}_a-\mu_a\|_w\|\widehat{\mathbf{\Omega}}^{-1}-\mathbf{\Omega}^{-1}\|,\\&\mathrm{Var}\big(\hat{\eta}_a^{\mathrm{qr}}|D^n\big)\lesssim\frac{1}{n}+\frac{k}{n^2},\end{aligned}$$

where the weighted L_2 norm of a function g is $\|g\|_w^2=\int g^2(\mathbf{x})\pi_a(\mathbf{x})\rho(\mathbf{x})\,dF(\mathbf{x})$.

Now the story is the same as the generalization case except we need to replace $1/\rho\pi_a$ with $(1-\rho)/\rho\pi_a$ (i.e., we impose smoothness assumption on $(1-\rho)/\rho\pi_a$ and



(a) n=100 (b) n=1000 (c) n=5000

FIG. 1. *RMSE vs.* α.

$(1-\hat{\rho})/\hat{\rho}\hat{\pi}_a$). With the approximation property of basis **b** (Assumption 6.2) and further assume a bound on the convergence rate of the nuisance function estimators

$$\left\| \frac{1-\hat{\rho}}{\hat{\rho}\hat{\pi}_a} - \frac{1-\rho}{\rho\pi_a} \right\|_w \|\hat{\mu}_a - \mu_a\|_w \|\widehat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}\|,$$

we can prove that the quadratic estimator $\hat{\eta}_a^{\text{qr}}$ achieves the minimax lower bound in a broader regime than the doubly robust estimator.

7. SIMULATION STUDY

In this section, we examine the performance of doubly robust estimators empirically. Specifically, we examine the performance of the plug-in-style estimator and the doubly robust estimator under varying nuisance convergence rates in Section 7.1 and in the presence of potential model misspecification in Section 7.2.

7.1 Estimation Error and Nuisance Estimation Error

Consider the following data generating process: $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5) \sim N(0, \mathbf{I}_5)$, $\mathbf{V} = (X_1, X_2, X_3)$. Given n samples, generate S according to $\rho(\mathbf{V}) = \mathbb{P}(S = 1|\mathbf{V}) = 0.5$ (participation probability). In the source population, set $\pi_1(\mathbf{x}) = \text{expit}(0.3x_1 - 0.3x_3)$ and simulate the treatment $A \sim \text{Bernoulli}(\pi_1(\mathbf{X}))$. Consider the following linear potential outcome model:

$$\mu_1(\mathbf{x}) = 1.5x_1 + x_4 + 1, \mu_0(\mathbf{x}) = x_1,$$

and $Y = A\mu_1(\mathbf{X}) + (1-A)\mu_0(\mathbf{X}) + N(0,1)$. So, we have

$$\tau_1(\mathbf{v}) = \mathbb{E}\big[\mu_1(\mathbf{X})|\mathbf{V} = \mathbf{v}, S = 1\big] = 1.5x_1 + 1, \tau_0(\mathbf{v}) = x_1.$$

The transportation functional is $\theta_1 = \mathbb{E}[\tau_1(\mathbf{V})|S = 0] = 1$. The nuisance estimators are $\hat{\mu}_a(\mathbf{x}) = \mu_a(\mathbf{x}) + \epsilon_{1,n}$, $\hat{\tau}_a(\mathbf{v}) = \tau_a(\mathbf{v}) + \epsilon_{2,n}$, $\hat{\rho}(\mathbf{v}) = \text{expit}(\text{logit}(\rho(\mathbf{v})) + \epsilon_{3,n})$, $\hat{\pi}(\mathbf{x}) = \text{expit}(\text{logit}(\pi_1(\mathbf{x})) + \epsilon_{4,n})$, where $\epsilon_{i,n} \sim N(n^{-\alpha}, n^{-2\alpha})$. This construction guarantees that the root mean square errors (RMSE) of $\hat{\pi}_1$, $\hat{\rho}$, $\hat{\mu}_a$, $\hat{\tau}_a$ are of order $O(n^{-\alpha})$. Then we can use different values of α to evaluate the performance of the doubly robust estimator and plug-in estimator when the nuisance functions are estimated with different convergence rates $O(n^{-\alpha})$. Specifically, we set the possible values of α to be a sequence ranging from 0.1 to 0.5 by a step of 0.05. In each replication, we generate the data and use the doubly robust estimator and the plug-in estimator to estimate the functional $\theta_1 = 1$. This process is replicated 1000 times for sample size $n = 100, 1000, 5000$ and the RMSE is computed. We report the simulation results in Figure 1.

From Figure 1, we see when the sample size is large and the convergence rate of the nuisance functions is slower than the parametric rate $O(n^{-1/2})$; the RMSE of doubly robust estimators can be much lower than the plug-in estimators. This coincides with our theoretical results. Note that although the performance of the plug-in estimator can be as good as the doubly robust estimator when the convergence rate is approximately the parametric rate, in practice we may be unable to correctly specify a parametric model for nuisance functions, and hence we are unlikely to achieve the parametric rate. So, the doubly robust estimator is recommended in generalizing and transporting causal effects from the source population to the target population in real data analysis. In future work, we will compare against the performance of our higher-order estimator.

7.2 Estimation Error and Model Misspecification

Consider the following setting: $\mathbf{X} = (X_1, X_2, X_3, X_4) \sim N(0, \mathbf{I}_4)$, $\mathbf{V} = (X_1, X_2, X_3)$. Given n samples, generate S according to $\rho(\mathbf{v}) = \mathbb{P}(S = 1|\mathbf{v}) = 0.2x_1 + 0.4x_2 - 0.5x_3^2$. In the source population, set $\pi_1(\mathbf{x}) = \text{expit}(0.3x_1 - 0.2x_2 + 0.5x_3^2)$ and simulate the treatment $A \sim \text{Bernoulli}(\pi_1(\mathbf{X}))$. Consider the following linear potential outcome model:

$$\mu_1(\mathbf{x}) = 1.5x_1 + 0.5x_2 + x_3^2 + x_4,$$

and $Y = A\mu_1(\mathbf{X}) + N(0,1)$. So, we have

$$\tau_1(\mathbf{v}) = \mathbb{E}\big[\mu_1(\mathbf{X})|\mathbf{V} = \mathbf{v}, S = 1\big] = 1.5x_1 + 0.5x_2 + x_3^2.$$

The generalization functional is $\psi_1 = \mathbb{E}[\tau_1(\mathbf{V})] = 1$. We fit (generalized) linear regression models for the nuisance functions μ_1, τ_1, π_1, ρ. To evaluate estimator performance under model misspecification, we consider scenarios where the outcome models μ_1, τ_1 or the propensity score models π_1, ρ are correctly specified or misspecified, with misspecification treating the quadratic term

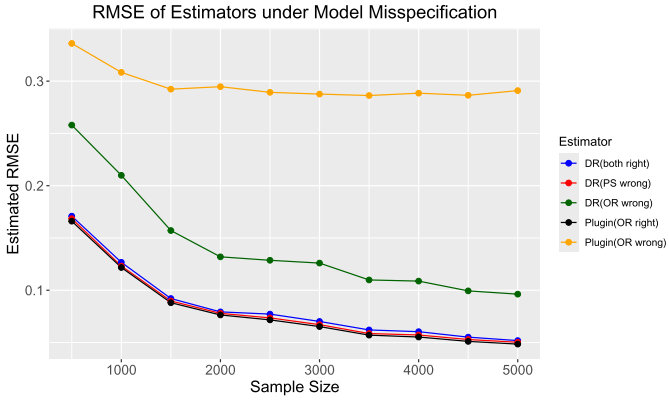


FIG. 2. *Performance of estimators under model misspecification.*

in X_3 as linear. We generate samples with size $n \in \{500, 1000, \ldots, 5000\}$, apply two estimators under different model misspecifications, and compute the RMSE. The results are summarized in Figure 2.

When the outcome model is misspecified, the plug-in-style estimator, which relies solely on outcome modeling, becomes inconsistent, as illustrated in Figure 2. In contrast, the doubly robust estimator models both the outcome and treatment assignment, making it more robust to model misspecification. As long as either the outcome models τ_1, μ_1 or the propensity score models π_1, ρ are correctly specified, the doubly robust estimator remains consistent, aligning with the theoretical results in Theorem 4.4.

8. DATA ANALYSIS

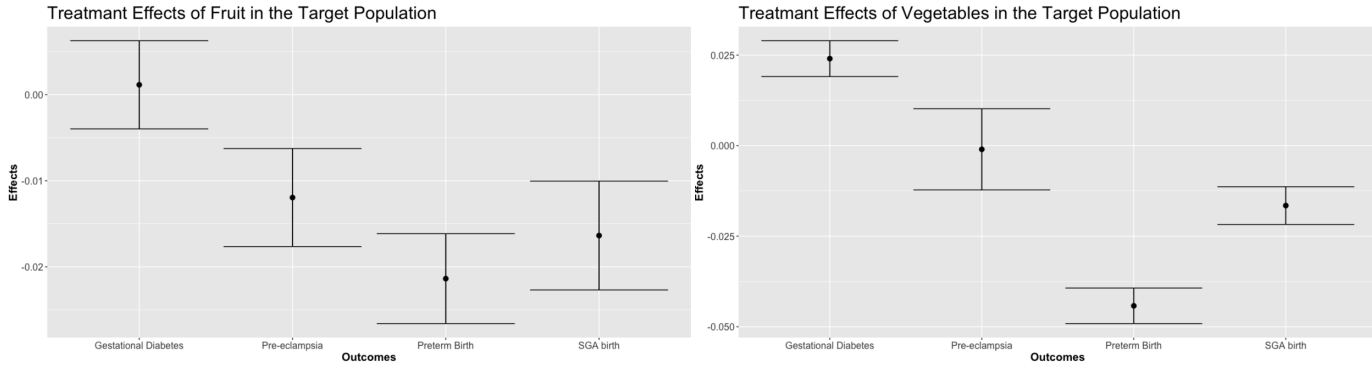
In this section, we illustrate the proposed method with a real data example. Specifically, we aim to transport the causal effects of fruit and vegetable intake on adverse pregnancy outcomes from an observational study to all pregnancies in the U.S. We first introduce the background and motivation of the problem in Section 8.1. The necessity of transportation is assessed in supplementary materials [59], that is, we show the distributions of covariates in two populations are different and some covariates may modify effects. In Section 8.2, we apply the proposed method to two data sets and estimate the ATE of different dietary components in the target population (*the whole U.S. pregnant female population*). Finally, sensitivity analysis is performed when we are not confident in the exchangeability and transportability assumption in Section 8.3.

8.1 Background and Motivation

Preterm birth, small-for-gestational-age birth, preeclampsia and gestational diabetes are adverse pregnancy and birth outcomes that contribute to one-quarter of infant deaths in the U.S. and pose a tremendous economic and emotional burden for societies and families [12, 19, 46]. Maternal nutrition is one of the few known modifiable risk factors for adverse pregnancy outcomes [45]. Preventing poor outcomes by optimizing preconception dietary patterns is therefore a public health priority. We previously showed that diets with a high density of fruits and vegetables were associated with a reduced risk of poor pregnancy outcomes [8]. In the current analysis, we estimated the causal effects of fruit and vegetable intake on adverse pregnancy outcomes in *the whole U.S. population of pregnant women*. Ideally, we would conduct randomized controlled trials on this population, or a random sample of it, and apply standard causal inference techniques to estimate the ATE. However, sampling and doing experiments on the U.S. pregnant population is not feasible. Therefore, we will transport the causal effects obtained from our prior work using data from the Nulliparous Pregnancy Outcomes Study: Monitoring mothers-to-be (nuMoM2b) to the U.S. pregnant population.

As described in detail previously [22], nuMoM2b enrolled 10,083 people in 8 U.S. medical centers from 2010 to 2013. Eligibility criteria included a viable singleton pregnancy, 6–13 completed weeks of gestation at enrollment, and no previous pregnancy that lasted ≥ 20 weeks of gestation. All participants provided informed, written consent and the study was approved by each site's institutional review board. At enrollment (6–13 completed weeks of gestation), women completed a semiquantitative food frequency questionnaire querying usual periconceptional dietary intake. The full dietary assessment approach has been published [8]. The treatments of interest are servings of fruits and vegetables per day. Other food groups served as important confounders: whole grains, dairy products, total protein foods, seafood and plant proteins, fatty acids, refined grains, sodium and "empty" calories. At least 30 days after delivery, a trained certified chart abstractor recorded final birth outcomes, medical history, and delivery diagnoses and complications. This information provides us with data on responses: preterm birth, SGA birth, gestational diabetes and preeclampsia. Data were also ascertained on maternal age, race, education, prepregnancy body mass index, smoking status, marital status, insurance status and working status. In our analysis, we set the threshold to be the 80% quantile of total fruit or vegetable intake. Total fruit/vegetable intake higher than the 80% quantile is considered treated ($A = 1$), otherwise not treated ($A = 0$). For the outcome, when the adverse pregnancy outcome occurs, we let $Y = 1$. Otherwise $Y = 0$. We will focus on the causal effects of fruit and vegetable intake on the aforementioned adverse pregnancy outcomes.

However, participants in the nuMoM2b cohort may not be representative of the U.S. pregnant population. For instance, 23% of participants received an education beyond



(a) Fruit (b) Vegetables

FIG. 3. *Effects of fruit and vegetable intake in the whole U.S. pregnant female population.*

college compared with 10% of the U.S. pregnant population. Hence, the distributions of covariates in study participants and target population are quite different. The nuMoM2b cohort may not be representative of the target population. It is possible that some covariates, such as age and education level will modify the effects of dietary components. Then the estimates in [8] based on Numom study will not immediately generalize to the U.S. population of women. To estimate the ATE in *the whole U.S. pregnant female population*, we find a U.S. representative sample from the National Survey of Family Growth (NSFG), which contains information on 9553 women in the U.S. The documentation of the data is available at https://www.cdc.gov/nchs/nsfg/nsfg_2015_2017_puf.htm.

8.2 Transportation

Having observed that transportation methods are necessary based on results in Supplementary Materials [59], here we assume the five identification assumptions in Section 3 and use our doubly robust estimator to estimate the ATE in the target population. The covariate sets that we use are $\mathbf{X} = \{$education, age, race, marital status, insurance, work, smoking, number of cigarettes, BMI, other dietary components (e.g., protein)$\}$ and $\mathbf{V} = \{$education, age, race, marital status, insurance, work, smoking, number of cigarettes$\}$. In particular, we adapt our estimators to these two data sets in two aspects. First, there exist some units with extremely small propensity score $\hat{\pi}_a(\mathbf{X})$ in the source data set. Since we need to reweight each sample in the nuMoM2b cohort by the inverse of the probability of getting treated, our doubly robust estimator may suffer from high instability if we directly use these small propensity scores. So, we enforce all the propensity scores $\pi_a(\mathbf{X}_i)$'s to be in the range [0.01, 0.99] (i.e., the positivity constant ϵ is set to be 0.01 in our analysis). Similarly, we also enforce all the participation probabilities $\rho(\mathbf{V}_i)$'s to be greater than 0.01.

The other point is that the sampling process of NSFG data set undergoes multiple stages and is more complicated than simple random sampling (SRS). According to examples given by CDC, we can approximate that sampling process with a stratified cluster sampling procedure and estimate the mean of a variable with appropriate weights. The variance of the estimator can also be obtained from the theory of stratified cluster sampling. The details on adjusting our estimator based on stratified cluster sampling can be found in the Supplementary Materials [59].

Our methods with above adjustments are applied to each combination of treatments (fruit intake, vegetable intake) and outcomes (preterm birth, SGA birth, gestational diabetes, preeclampsia). All the nuisance functions are fitted with "SL.ranger" (random forests) and "SL.glmnet" (penalized GLM) and "SL.mean" in the R package "SuperLearner." Five-fold cross-fitting is used to guarantee the sample splitting condition in Theorem 4.5. The results are summarized in Figure 3.

Detailed analysis on model selection (i.e., choice of models in SuperLearner), choice of positivity constant ϵ and potential outliers is presented in the Supplementary Materials [59], which justifies the choice in our analysis. The effects of high fruit intake on preterm birth, preeclampsia, gestational diabetes and SGA birth are -0.0214 (95% CI $[-0.0266, -0.0162]$), -0.012 (95% CI $[-0.0176, -0.00626]$), 0.00114 (95% CI $[-0.00398, 0.00626]$), -0.0164 (95% CI $[-0.0227, -0.0101]$), respectively. The effects of high vegetable intake on preterm birth, preeclampsia, gestational diabetes and SGA birth are -0.0442 (95% CI $[-0.0491, -0.0393]$), -0.00102 (95% CI $[-0.0122, 0.0102]$), 0.024 (95% CI [0.0191, 0.029]), -0.0166 (95% CI $[-0.0218, -0.0113]$).

From the results above, we see the effects of high fruit intake on preterm birth, preeclampsia and SGA birth are significantly negative at level 0.05 in the target population, which implies eating more fruit potentially causes a lower risk of these adverse pregnancy outcomes. For the results on vegetables, the effect of high vegetable intake on preterm birth and SGA birth is significantly negative. The strict interpretation for the effects of high vegetable

intake on preterm birth is: Compared with participants whose vegetable intake is less than 80% quantile of vegetable intake in the nuMoM2b cohort, women with higher vegetable intake have 4 fewer preterm births for every 100 women in *the whole U.S. pregnant female population*. Other combinations of treatments and outcomes can be similarly interpreted. We also see (significant) positive effects of high vegetable intake on gestational diabetes, which shows eating more vegetables may potentially increase the risk of getting gestational diabetes. This result is counterintuitive. One potential problem is the identification assumptions may not hold, so we cannot interpret our results from a causal perspective. As a result, we will perform sensitivity analysis in the following section, to deal with possible violations of the identification assumptions.

8.3 Sensitivity Analysis and Discussion

Here, we focus on the effects of high vegetable intake on gestational diabetes as an example. According to the discussions in Section 3.1, under Assumption 3.7 and Assumption 3.8 the bound for the ATE in the target population is given by

$$[0.024 - \delta_1 - 2\delta_2, 0.024 + \delta_1 + 2\delta_2].$$

We visualize the range of the bounds when only exchangeability is violated ($\delta_2 = 0$) or only transportability is violated ($\delta_1 = 0$) in Figure 4, respectively.

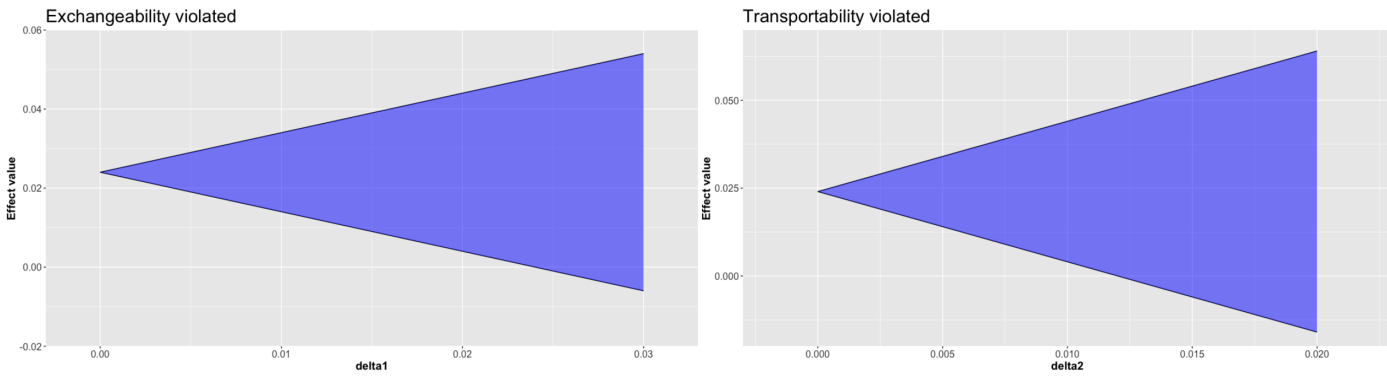
The plots should be understood as follows: (a) For a specific value δ_1, the interval that covers the ATE is given by the intersection of the blue region and the line $x = \delta_1$. (b) Can be interpreted similarly. From Figure 4, we see if only exchangeability is violated (i.e., $\delta_2 = 0$), then the critical value for δ_1 to turn around the result is 0.024. If only transportability is violated (i.e., $\delta_1 = 0$), then the critical value for δ_2 to turn around the result is 0.012.

Considering the covariate sets $\mathbf{X} = \{$education, age, race, marital status, insurance, work, smoking, number of cigarettes, BMI, other dietary components (e.g., protein)$\}$ and $\mathbf{V} = \{$education, age, race, marital status, insurance, work, smoking, number of cigarettes$\}$, we find most of the covariates are categorical. The information contained in categorical variables is usually less than that in continuous ones. Furthermore, there may be some confounders or effect modifiers that are not measured in the data set. For instance, body mass index (BMI) may be an effect modifier but we do not have information on it in the NSFG data set, and hence cannot take it into account. Therefore, the categorical variables we include in the analysis may not provide sufficient information for the identification Assumptions 3.2 and 3.4 to hold. Hence, sensitivity analysis is quite necessary in our problem.

From the perspective of the high vegetable intake treatment itself, it may be that the increased risk is due to the majority of vegetables being fried potatoes, which are less healthy than other vegetables.

9. DISCUSSION

In this paper, we extend the generalization and transportation techniques in [17, 18] to the case where a different sets of covariates can be used in the source population and the target population. To be concrete, we summarize the sufficient assumptions to identify the ATE in the target population. We also provide methods to perform sensitivity analysis when the identification assumptions fail. The first-order influence functions and efficiency bounds are derived for the target statistical functionals when they are identified. We further propose a doubly robust estimator based on the first-order influence functions and establish its asymptotic normality under proper conditions. Simulation studies show the advantage of the doubly robust estimator over the plug-in estimator. We also study the minimax lower bounds and higher-order/quadratic estimation based on second-order von Mises expansion in the case where the source population and the target population share the same set of covariates (i.e., $\mathbf{V} = \mathbf{X}$). Although we rely on similar techniques used in the ATE case [38, 39], these results are nontrivial and important in



(a) Only exchangeability is violated ($\delta_2 = 0$)

(b) Only transportability is violated($\delta_1 = 0$)

FIG. 4. *Sensitivity analysis when treatment is high vegetable intake and outcome is gestational diabetes.*

understanding the properties of generalization and transportation functional. Finally, we illustrate the proposed methods with an interesting example, where we transport the causal effects of fruit and vegetable intake on adverse pregnancy outcomes from an observational study to *the whole U.S. pregnant female population*.

In this paper, we only consider the minimax rate and quadratic estimation in the special case $\mathbf{V} = \mathbf{X}$. In the general case $\mathbf{X} = (\mathbf{W}, \mathbf{V})$, the target functionals in (5) involve a triple integral and can be viewed as "cubic functionals" [34, 48]. Such functionals are not well understood in the literature and it is more challenging to construct adversarial settings to establish the minimax lower bounds or derive higher-order influence functions to correct for the second-order bias. One special case is to assume the covariates $\mathbf{V}$ are discrete and we can write

$$\psi_a = \sum_{\mathbf{v}} p(\mathbf{v}) \mathbb{E}\big[\mathbb{E}(Y|\mathbf{X}, A = a, S = 1)|\mathbf{V} = \mathbf{v}, S = 1\big].$$

Now we only need to estimate $\mathbb{E}[\mathbb{E}(Y|\mathbf{X}, A = a, S = 1)|\mathbf{V} = \mathbf{v}, S = 1] = \mathbb{E}[\mathbb{E}(Y|\mathbf{W}, A = a, \mathbf{V} = \mathbf{v}, S = 1)|\mathbf{V} = \mathbf{v}, S = 1]$. Note that this functional is equivalent to estimating ATE among individuals with $\mathbf{V} = \mathbf{v}$, $S = 1$ using covariate set $\mathbf{W}$, and hence the standard ATE theory applies. We leave the general case as future work. Moreover, while [9] and [58] have demonstrated that higher-order estimators perform well in settings with continuous treatments or discrete covariates, the implementation and selection of tuning parameters remain open problems for future research.

Another relevant literature focuses on combining multiple data sources to estimate treatment effects [23, 24, 57], where some sources may have unmeasured confounders or may not represent the target population. Extending our methods to this setting, where generalization and transportation are considered with multiple data sets from the source population, is a potential avenue for future research. Additionally, other interesting questions in generalization and transportation include whether it is possible to generalize or transport time-varying treatment effects, where the treatment or exposure changes over time. In the classical ATE framework, several approaches exist for sensitivity analysis, where different sensitivity models are used to address potential violations of identification assumptions [10, 53, 56, 60]. Formalizing similar sensitivity models for use in generalization and transportation settings remains an open question for future investigation. A further practical challenge is how to appropriately select confounders $\mathbf{X}$ and effect modifiers $\mathbf{V}$. [43] demonstrated that researchers can focus on variables that are both effect modifiers and differentially distributed across groups in $\mathbf{V}$. However, identifying this set of variables is typically not straightforward in applications. Therefore, it is crucial to develop principled methods for selecting variables to adjust for in the generalization and transportation of causal effects.

APPENDIX A: GENERAL STRATEGIES TO PROVE MINIMAX LOWER BOUNDS

We will use the following lemma (adapted from Le Cam's convex hull method), which summarizes the main ingredients in deriving minimax lower bound for a statistical functional, to derive the minimax lower bounds.

LEMMA A.1. *Let $\mathbb{P}_{\boldsymbol{\lambda}}$ and $\mathbb{Q}_{\boldsymbol{\lambda}}$ denote distributions in $\mathcal{P}$ indexed by a vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ with corresponding n-dimensional product measure $\mathbb{P}^n_{\boldsymbol{\lambda}}$ and $\mathbb{Q}^n_{\boldsymbol{\lambda}}$. Let π be a probability distribution of $\boldsymbol{\lambda}$. The functional of interest is $\psi : \mathcal{P} \mapsto \mathbb{R}$. If*

$$H^2\left(\int \mathbb{P}^n_{\boldsymbol{\lambda}}\, d\pi(\boldsymbol{\lambda}), \int \mathbb{Q}^n_{\boldsymbol{\lambda}}\, d\pi(\boldsymbol{\lambda})\right) \le \alpha < 2$$

and

$$\left|\psi(\mathbb{P}_{\boldsymbol{\lambda}}) - \psi(\mathbb{Q}_{\boldsymbol{\lambda}'})\right| \ge s > 0$$

for all $\boldsymbol{\lambda}, \boldsymbol{\lambda}'$, then

$$\inf_{\widehat{\psi}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}\{|\widehat{\psi} - \psi(\mathbb{P})|\} \ge \frac{s}{2}\left(\frac{1 - \sqrt{\alpha(1 - \alpha/4)}}{2}\right).$$

As pointed out in Lemma A.1, the general method to derive minimax lower bound is to construct two distributions that are statistically indistinguishable (i.e., the distance between two distributions is small) but the target parameters of two distributions are sufficiently separated. In nonparametric regression problems, to arrive at the minimax rate it suffices to apply Le Cam's two-point method and construct a pair of distributions [52]. In nonlinear functional estimation, typically at least one distribution needs to be a mixture to obtain a tight bound [39]. Intuitively, by considering mixture distributions, the signal from the distribution is further neutralized, which yields distributions with smaller statistical distance while the target parameter is still separated, and thus gives tighter lower bounds. New techniques for bounding the distance between mixture distributions have been developed in the literature. For instance, [7] considered the case where only one distribution is a mixture and [39] further extended the bound to the case where both distributions are mixtures. In our analysis, we will use the following bound on Hellinger distance from [39].

LEMMA A.2 ([39]). *Let $\mathbb{P}_{\boldsymbol{\lambda}}$ and $\mathbb{Q}_{\boldsymbol{\lambda}}$ denote distributions in $\mathcal{P}$ indexed by a vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ with corresponding n-dimensional product measure $\mathbb{P}^n_{\boldsymbol{\lambda}}$ and $\mathbb{Q}^n_{\boldsymbol{\lambda}}$ and let $\mathcal{Z} = \cup_{j=1}^k \mathcal{Z}_j$ be a partition of the sample space satisfying*:

1. *$\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{Z}_j) = \mathbb{Q}_{\boldsymbol{\lambda}}(\mathcal{Z}_j) = p_j$ for all $\boldsymbol{\lambda}$;*
2. *the conditional distributions $I_{\mathcal{Z}_j} d\mathbb{P}_{\boldsymbol{\lambda}}/p_j$, $I_{\mathcal{Z}_j} d\mathbb{Q}_{\boldsymbol{\lambda}}/p_j$ only depend on λ_j (do not depend on λ_k for $k \ne j$).*

Let p_λ and q_λ denote the density of $\mathbb{P}_\lambda$ and $\mathbb{Q}_\lambda$, respectively. Given a product distribution π on λ, define $p = \int p_\lambda \, d\pi(\lambda)$, $q = \int q_\lambda \, d\pi(\lambda)$ and

$$\delta_1 = \max_j \sup_\lambda \int_{\mathcal{Z}_j} \frac{(p_\lambda - p)^2}{p_\lambda p_j} \, d\nu,$$

$$\delta_2 = \max_j \sup_\lambda \int_{\mathcal{Z}_j} \frac{(q_\lambda - p_\lambda)^2}{p_\lambda p_j} \, d\nu,$$

$$\delta_3 = \max_j \sup_\lambda \int_{\mathcal{Z}_j} \frac{(q - p)^2}{p_\lambda p_j} \, d\nu$$

for a dominating measure ν. If $np_j(1 \vee \delta_1 \vee \delta_2) \le b$ for all j and $b^{-1} \le p_\lambda \le b$ for a positive constant b, then there exists a constant C that only depends on b such that

$$H^2\Big(\int \mathbb{P}_\lambda^n \, d\pi(\lambda), \int \mathbb{Q}_\lambda^n \, d\pi(\lambda)\Big) \le Cn^2\Big(\max_j p_j\Big)(\delta_1\delta_2 + \delta_2^2) + Cn\delta_3.$$

The idea of deriving minimax lower bound now becomes clear. We first carefully construct two classes of distributions indexed by λ such that the target parameters of two distribution classes are separated. Then we can apply Lemma A.2 to bound the Hellinger distance between the mixture distributions of two distribution classes (with the same mixture distribution on λ). Finally, we apply Lemma A.1 to obtain the minimax lower bound. In Section 5.1, we apply this idea to the generalization and transportation setting.

APPENDIX B: QUADRATIC VON MISES CALCULUS

As discussed in Section 4.1, the first-order influence function is the derivative in the first-order von Mises expansion of the statistical functional $\psi = \psi(\mathbb{P})$. Mathematically, we have

$$\psi(\hat{\mathbb{P}}) - \psi(\mathbb{P}) = -\int \phi_1(\mathbf{z}, \hat{\mathbb{P}}) \, d\mathbb{P}(\mathbf{z}) + R_2(\hat{\mathbb{P}}, \mathbb{P}),$$

where $\phi_1(\mathbf{z}, \mathbb{P})$ is the influence function of $\psi(\mathbb{P})$, $\psi(\hat{\mathbb{P}})$ is the plug-in estimator and $R_2(\hat{\mathbb{P}}, \mathbb{P})$ is the second-order reminder. This suggests that the plug-in estimator has first-order bias $-\int \phi_1(\mathbf{z}, \hat{\mathbb{P}}) \, d\mathbb{P}(\mathbf{z})$. Equivalently, we can write

$$\psi(\mathbb{P}) = \psi(\hat{\mathbb{P}}) + \int \phi_1(\mathbf{z}, \hat{\mathbb{P}}) \, d\mathbb{P}(\mathbf{z}) + R_2(\hat{\mathbb{P}}, \mathbb{P}),$$

which motivates us to correct for the first-order bias and arrive at the doubly robust estimator

$$\hat{\psi}^{\mathrm{dr}} = \psi(\hat{\mathbb{P}}) + \mathbb{P}_n\{\phi_1(\mathbf{Z}, \hat{\mathbb{P}})\}.$$

Under further empirical process assumptions or sample splitting assumptions, we can show the dominating term in the conditional bias of $\hat{\psi}^{\mathrm{dr}}$ is $R_2(\hat{\mathbb{P}}, \mathbb{P})$, which is usually a second-order error term (e.g., $R_2(\hat{\mathbb{P}}, \mathbb{P})$ is of order $\|\hat{\mu}_a - \mu_a\|\|\hat{\pi}_a - \pi_a\| + \|\hat{\rho} - \rho\|\|\hat{\tau}_a - \tau_a\|$ in generalization and transportation functionals). At this point, one may consider further expressing $\psi(\mathbb{P})$ as a second-order Von Mises expansion and hopefully the remainder error term can be of "third-order small". This motivates the study of second-order influence functions. Mathematically, we can write

$$\begin{aligned}
&\psi(\mathbb{P}) - \psi(\hat{\mathbb{P}}) \\
&\quad = \int \phi_1(\mathbf{z}, \hat{\mathbb{P}}) \, d(\mathbb{P} - \hat{\mathbb{P}})(\mathbf{z}) \\
&\qquad + \frac{1}{2}\int \phi_2(\mathbf{z}_1, \mathbf{z}_2, \hat{\mathbb{P}}) \, d(\mathbb{P} - \hat{\mathbb{P}}) \\
&\qquad \times (\mathbb{P} - \hat{\mathbb{P}})(\mathbf{z}_1, \mathbf{z}_2) \\
&\qquad + R_3(\hat{\mathbb{P}}, \mathbb{P}) \\
&\quad = \int \phi_1(\mathbf{z}, \hat{\mathbb{P}}) \, d\mathbb{P}(\mathbf{z}) \\
&\qquad + \frac{1}{2}\int \phi_2(\mathbf{z}_1, \mathbf{z}_2, \hat{\mathbb{P}}) \, d(\mathbb{P} \times \mathbb{P})(\mathbf{z}_1, \mathbf{z}_2) \\
&\qquad + R_3(\hat{\mathbb{P}}, \mathbb{P}),
\end{aligned} \tag{9}$$

where $\phi_1(\mathbf{z}, \mathbb{P})$ is the first-order influence function and $\phi_2(\mathbf{z}_1, \mathbf{z}_2, \mathbb{P})$ is the second-order influence function. Here, we assume both ϕ_1 and ϕ_2 are centered: $\int \phi_1(\mathbf{z}, \mathbb{P}) \, d\mathbb{P}(\mathbf{z}) = 0$ and $\int \phi_2(\mathbf{z}_1, \mathbf{z}_2, \mathbb{P}) \, d\mathbb{P}(\mathbf{z}_i) = 0$ for $i = 1, 2$. A strict mathematical definition of second-order influence function requires the notion of second-order score and we refer to [38] for more detailed discussions. When the second-order influence function exists, it can be derived by differentiating the first-order influence function (i.e., deriving the influence function of the first-order influence function), as summarized in the following lemma.

LEMMA B.1 ([38]). *Suppose $\{\mathbb{P}_t, t \ge 0\}$ is a sufficiently smooth submodel and $\phi_1 : \mathcal{Z} \mapsto \mathbb{R}$, $\phi_2 : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$ are measurable functions that satisfy*

$$\begin{aligned}
\frac{\partial \psi(\mathbb{P}_t)}{\partial t} &= \int \phi_1(\mathbf{z}, \mathbb{P}_t) \frac{\partial p_t(\mathbf{z})}{\partial t} \, d\mathbf{z}, \quad t \ge 0, \\
\frac{\partial \phi_1(\mathbf{z}_1, \mathbb{P}_t)}{\partial t}\Big|_{t=0} &= \int \phi_2(\mathbf{z}_1, \mathbf{z}_2, \mathbb{P}) \\
&\qquad \times \frac{\partial \log p_t(\mathbf{z}_2)}{\partial t}\Big|_{t=0} d\mathbb{P}(\mathbf{z}_2)
\end{aligned}$$

$\forall \mathbf{z}_1 \in \mathcal{Z}$. Then the function ϕ_2 is a second-order influence function of ψ.

After deriving the second-order influence function, we replace the integral with respect to $\mathbb{P}$ and $\mathbb{P} \times \mathbb{P}$ with a sample average and U-statistics measure in (9), and arrive at the general form of the quadratic estimator for a statistical functional ψ as follows:

$$\begin{aligned}
\hat{\psi}^{\mathrm{qr}} &= \psi(\hat{\mathbb{P}}) + \mathbb{P}_n\{\phi_1(\mathbf{Z}, \hat{\mathbb{P}})\} \\
&\quad + \frac{1}{2}\mathbb{U}_n\{\phi_2(\mathbf{Z}_1, \mathbf{Z}_2, \hat{\mathbb{P}})\}.
\end{aligned} \tag{10}$$

Since we have corrected for both first-order and second-order bias in the plug-in estimator, ideally we may hope $R_3(\hat{\mathbb{P}}, \mathbb{P})$ is a third-order error term, and thus the bias of $\hat{\psi}^{\mathrm{qr}}$ is smaller than $\hat{\psi}^{\mathrm{dr}}$. Unfortunately, the nonexistence of a second-order influence function is typical for many functionals, implying that a third-order error term $R_3(\hat{\mathbb{P}}, \mathbb{P})$ may not be attainable. However, we may still be able to correct for "partial second-order bias." The key is to find a finite-dimensional subspace L and define an "approximate second-order influence function" based on L, which induces an additional representational error to $R_3(\hat{\mathbb{P}}, \mathbb{P})$. The dimension of L is carefully selected to balance the representation error, variance of the quadratic estimator and the remainder term in von Mises expansion. In Section 6.1, we illustrate how to construct quadratic estimators based on second-order influence functions in generalization and transportation settings.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referee and the Editor for their constructive comments that improved the quality of this paper.

FUNDING

This work is supported by NIH R01 Grant HD102313.

SUPPLEMENTARY MATERIAL

Supplementary Derivations, Proofs and Results (DOI: 10.1214/24-STS972SUPPA; .pdf). Proof of all results, additional details in real data analysis.

Supplementary R Code (DOI: 10.1214/24-STS972SUPPB; .zip). The file transportation_codes.zip provides codes for simulations and data analysis.

REFERENCES

[1] ALLCOTT, H. (2015). Site selection bias in program evaluation. *Q. J. Econ.* **130** 1117–1165.

[2] BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1176.

[3] BELL, S. H., OLSEN, R. B., ORR, L. L. and STUART, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educ. Eval. Policy Anal.* **38** 318–335.

[4] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186** 345–366. MR3343791 https://doi.org/10.1016/j.jeconom.2015.02.014

[5] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

[6] BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393. MR1065550

[7] BIRGÉ, L. and MASSART, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29. MR1331653 https://doi.org/10.1214/aos/1176324452

[8] BODNAR, L. M., CARTUS, A. R., KIRKPATRICK, S. I., HIMES, K. P., KENNEDY, E. H., SIMHAN, H. N., GROBMAN, W. A., DUFFY, J. Y., SILVER, R. M. et al. (2020). Machine learning as a strategy to account for dietary synergy: An illustration based on dietary intake and adverse pregnancy outcomes. *Amer. J. Clin. Nutr.* **111** 1235–1243.

[9] BONVINI, M. and KENNEDY, E. H. (2022). Fast convergence rates for dose-response estimation. arXiv preprint. Available at arXiv:2207.11825.

[10] BONVINI, M. and KENNEDY, E. H. (2022). Sensitivity analysis via the proportion of unmeasured confounding. *J. Amer. Statist. Assoc.* **117** 1540–1550. MR4480730 https://doi.org/10.1080/01621459.2020.1864382

[11] BUCHANAN, A. L., HUDGENS, M. G., COLE, S. R., MOLLAN, K. R., SAX, P. E., DAAR, E. S., ADIMORA, A. A., ERON, J. J. and MUGAVERO, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *J. Roy. Statist. Soc. Ser. A* **181** 1193–1209. MR3876388 https://doi.org/10.1111/rssa.12357

[12] BUTLER, A. S., BEHRMAN, R. E. et al. (2007). *Preterm Birth: Causes, Consequences, and Prevention*. National Academies Press, Washington, DC.

[13] CHATTOPADHYAY, A., COHN, E. R. and ZUBIZARRETA, J. R. (2024). One-step weighting to generalize and transport treatment effect estimates to a target population. *Amer. Statist.* **78** 280–289. MR4777181 https://doi.org/10.1080/00031305.2023.2267598

[14] CHEN, I. Y., PIERSON, E., ROSE, S., JOSHI, S., FERRYMAN, K. and GHASSEMI, M. (2021). Ethical machine learning in healthcare. *Annu. Rev. Biomed. Data Sci.* **4** 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

[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. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[16] COLE, S. R. and STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Amer. J. Epidemiol.* **172** 107–115. https://doi.org/10.1093/aje/kwq084

[17] DAHABREH, I. J., ROBERTSON, S. E., STEINGRIMSSON, J. A., STUART, E. A. and HERNÁN, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Stat. Med.* **39** 1999–2014. MR4105273 https://doi.org/10.1002/sim.8426

[18] DAHABREH, I. J., ROBERTSON, S. E., TCHETGEN, E. J., STUART, E. A. and HERNÁN, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* **75** 685–694. MR3999190 https://doi.org/10.1111/biom.13009

[19] DALL, T. M., YANG, W., HALDER, P., PANG, B., MASSOUDI, M., WINTFELD, N., SEMILLA, A. P., FRANZ, J. and HOGAN, P. F. (2014). The economic burden of elevated blood glucose levels in 2012: Diagnosed and undiagnosed diabetes, gestational diabetes mellitus, and prediabetes. *Diabetes Care* **37** 3172–3179. https://doi.org/10.2337/dc14-1036

[20] FIORE, L. D. and LAVORI, P. W. (2016). Integrating randomized comparative effectiveness research with patient care. *N. Engl. J. Med.* **374** 2152–2158. https://doi.org/10.1056/NEJMra1510057

[21] FOSTER, D. J. and SYRGKANIS, V. (2023). Orthogonal statistical learning. *Ann. Statist.* **51** 879–908. MR4630373 https://doi.org/10.1214/23-AOS2258

[22] HAAS, D. M., PARKER, C. B., WING, D. A., PARRY, S., GROBMAN, W. A., MERCER, B. M., SIMHAN, H. N., HOFFMAN, M. K., SILVER, R. M. et al. (2015). A description of the methods of the nulliparous pregnancy outcomes study: Monitoring mothers-to-be (nuMoM2b). *Amer. J. Obstet. Gynecol.* **212** 539–e1.

[23] HAN, L., HOU, J., CHO, K., DUAN, R. and CAI, T. (2021). Federated adaptive causal estimation (face) of target treatment effects. arXiv preprint. Available at arXiv:2112.09313.

[24] HAN, L., SHEN, Z. and ZUBIZARRETA, J. (2023). Multiply robust federated estimation of targeted average treatment effects. In *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, eds.) **36** 70453–70482. Curran Associates, Red Hook.

[25] HERNAN, M. A. and ROBINS, J. M. (2023). *Causal Inference. Chapman & Hall/CRC Monographs on Statistics & Applied Probab.* CRC Press, Boca Raton.

[26] KENNEDY, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. arXiv preprint. Available at arXiv:2203.06469.

[27] KENNEDY, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electron. J. Stat.* **17** 3008–3049. MR4667730 https://doi.org/10.1214/23-ejs2157

[28] KENNEDY, E. H., BALAKRISHNAN, S. and G'SELL, M. (2020). Sharp instruments for classifying compliers and generalizing causal effects. *Ann. Statist.* **48** 2008–2030. MR4134784 https://doi.org/10.1214/19-AOS1874

[29] KENNEDY, E. H., BALAKRISHNAN, S., ROBINS, J. M. and WASSERMAN, L. (2024). Minimax rates for heterogeneous causal effect estimation. *Ann. Statist.* **52** 793–816. MR4744196 https://doi.org/10.1214/24-aos2369

[30] KENNEDY-MARTIN, T., CURTIS, S., FARIES, D., ROBINSON, S. and JOHNSTON, J. (2015). A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* **16** 1–14.

[31] KERN, H. L., STUART, E. A., HILL, J. and GREEN, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *J. Res. Educ. Eff.* **9** 103–127. https://doi.org/10.1080/19345747.2015.1060282

[32] LEVIS, A. W., BONVINI, M., ZENG, Z., KEELE, L. and KENNEDY, E. H. (2023). Covariate-assisted bounds on causal effects with instrumental variables. arXiv preprint. Available at arXiv:2301.12106.

[33] LUEDTKE, A. R., DIAZ, I. and VAN DER LAAN, M. J. (2015). The statistics of sensitivity analyses.

[34] MUKHERJEE, R., TCHETGEN, E. T. and ROBINS, J. (2015). Lepski's method and adaptive estimation of nonlinear integral functionals of density. arXiv preprint. Available at arXiv:1508.00249.

[35] OLSCHEWSKI, M. and SCHEURLEN, H. (1985). Comprehensive cohort study: An alternative to randomized consent design in a breast preservation trial. *Methods Inf. Med.* **24** 131–134.

[36] QIN, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **85** 619–630. MR1665814 https://doi.org/10.1093/biomet/85.3.619

[37] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman. Inst. Math. Stat. (IMS) Collect.* **2** 335–421. IMS, Beachwood, OH. MR2459958 https://doi.org/10.1214/193940307000000527

[38] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. W. (2009). Quadratic semiparametric von Mises calculus. *Metrika* **69** 227–247. MR2481922 https://doi.org/10.1007/s00184-008-0214-3

[39] ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

[40] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

[41] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

[42] RUDOLPH, K. E. and VAN DER LAAN, M. J. (2017). Robust estimation of encouragement design intervention effects transported across sites. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1509–1525. MR3731673 https://doi.org/10.1111/rssb.12213

[43] RUDOLPH, K. E. WILLIAMS, N., STUART, E. A. and DÍAZ, I. (2024). Improving efficiency in transporting average treatment effects.

[44] SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

[45] STEPHENSON, J., HESLEHURST, N., HALL, J., SCHOENAKER, D. A., HUTCHINSON, J., CADE, J. E., POSTON, L., BARRETT, G., CROZIER, S. R. et al. (2018). Before the beginning: Nutrition and lifestyle in the preconception period and its importance for future health. *Lancet* **391** 1830–1841.

[46] STEVENS, W., SHIH, T., INCERTI, D., TON, T. G., LEE, H. C., PENEVA, D., MACONES, G. A., SIBAI, B. M. and JENA, A. B. (2017). Short-term costs of preeclampsia to the United States health care system. *Amer. J. Obstet. Gynecol.* **217** 237–248.

[47] SUGIYAMA, M., SUZUKI, T. and KANAMORI, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge Univ. Press, Cambridge. With a foreword by Thomas G. Dietterich. MR2895762 https://doi.org/10.1017/CBO9781139035613

[48] TCHETGEN, E., LI, L., ROBINS, J. and VAN DER VAART, A. (2008). Minimax estimation of the integral of a power of a density. *Statist. Probab. Lett.* **78** 3307–3311. MR2479495 https://doi.org/10.1016/j.spl.2008.07.001

[49] TCHETGEN TCHETGEN, E. J. and SHPITSER, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Ann. Statist.* **40** 1816–1845. MR3015045 https://doi.org/10.1214/12-AOS990

[50] TIPTON, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *J. Educ. Behav. Stat.* **38** 239–266.

[51] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data. Springer Series in Statistics*. Springer, New York. MR2233926

[52] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

[53] VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Ann. Intern. Med.* **167** 268–274. https://doi.org/10.7326/M16-2607

[54] VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality. Springer Series in Statistics*. Springer, New York. MR1958123 https://doi.org/10.1007/978-0-387-21700-0

[55] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[56] YADLOWSKY, S., NAMKOONG, H., BASU, S., DUCHI, J. and TIAN, L. (2022). Bounds on the conditional and average treatment effect with unobserved confounding factors. *Ann. Statist.* **50** 2587–2615. MR4505372 https://doi.org/10.1214/22-aos2195

[57] YANG, S. and DING, P. (2020). Combining multiple observational data sources to estimate causal effects. *J. Amer. Statist. Assoc.* **115** 1540–1554. MR4143484 https://doi.org/10.1080/01621459.2019.1609973

[58] ZENG, Z., BALAKRISHNAN, S., HAN, Y. and KENNEDY, E. H. (2024). Causal inference with high-dimensional discrete covariates. arXiv preprint. Available at arXiv:2405.00118.

[59] ZENG, Z., KENNEDY, E. H., BODNAR, L. M. and NAIMI, A. I. (2025). Supplement to "Efficient generalization and transportation." https://doi.org/10.1214/24-STS972SUPPA, https://doi.org/10.1214/24-STS972SUPPB

[60] ZHANG, B. and TCHETGEN TCHETGEN, E. J. (2022). A semiparametric approach to model-based sensitivity analysis in observational studies. *J. Roy. Statist. Soc. Ser. A* **185** S668–S691. MR4547527 https://doi.org/10.1111/rssa.12946

[61] ZHENG, W. and VAN DER LAAN, M. J. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation. Technical Report, U.C. Berkeley.