Average Adjusted Association: Efficient Estimation with High Dimensional Confounders

Sung Jae Jun
Department of Economics,
Pennsylvania State University,
University Park, PA 16802, USA

Sokbae Lee
Department of Economics,
Columbia University,
New York, NY, 10027, USA

Abstract

The log odds ratio is a well-established metric for evaluating the association between binary outcome and exposure variables. Despite its widespread use, there has been limited discussion on how to summarize the log odds ratio as a function of confounders through averaging. To address this issue, we propose the Average Adjusted Association (AAA), which is a summary measure of association in a heterogeneous population, adjusted for observed confounders. To facilitate the use of it, we also develop efficient double/debiased machine learning (DML) estimators of the AAA. Our DML estimators use two equivalent forms of the efficient influence function, and are applicable in various sampling scenarios, including random sampling, outcome-based sampling, and exposure-based sampling. Through real data and simulations, we demonstrate the practicality and effectiveness of our proposed estimators in measuring the AAA.

1 INTRODUCTION

There are several statistical measures of association, among which the (log) odds ratio is one of the most popular ones. The odds ratio has been frequently used in medicine, biostatistics and epidemiology because it is simple, it has a natural interpretation in the standard logistic regression model, and it is invariant under various sampling designs that include case-control studies. See Breslow (1976); Breslow and Powers (1978); Breslow and Day (1980) for early work and Bland and Altman (2000); Norton et al. (2018) for how to use the odds ratio in practice.

To describe the object of interest and the background more concretely, let Y denote a binary outcome, T a binary exposure and X a vector of measured confounders/covariates. The conditional odds ratio between Y and T given $X = x$, which we denote by $\mathrm{OR}(x)$, provides a complete picture of the *adjusted association* between Y and T for different subpopulations defined by different values of x. Conditioning on $X = x$ matters for two reasons: first, it is more plausible to make a causal interpretation out of $\mathrm{OR}(x)$ than the unconditional odds ratio between Y and T (e.g., Holland and Rubin, 1988; Greenland et al., 1999); second, the association can be heterogeneous over different subpopulations so that $\mathrm{OR}(x)$ is a complicated function of x in general.

There are several approaches to modeling $\mathrm{OR}(\cdot)$. It is most common to parametrize the function $\mathrm{OR}(\cdot)$ typically via a logistic regression model, but it may suffer from misspecification. Therefore, effort has been made to mitigate the problem from the perspective of doubly robust estimation (e.g., Chen, 2007; Tchetgen Tchetgen et al., 2010; Tchetgen Tchetgen, 2013). On the contrary, some authors have emphasized heterogeneity, advocating nonparametric estimation of the function $\mathrm{OR}(\cdot)$ (e.g., Chen et al., 2011; Hui and Geenens, 2013). However, fully nonparametric approaches generally suffer from the curse of dimensionality and they are often not a practical option in finite samples. In fact, all the numerical experiments in Chen et al. (2011) and Hui and Geenens (2013) are limited with the scalar X.

In this paper, we consider the case where the dimension of X is high. We introduce a new summary measure of association, which we call the *average adjusted association*, by taking the average of conditional log odds ratios over covariates. That is, we think of $\theta_0 := \mathbb{E}\{\log \mathrm{OR}(X)\}$ as a summary measure of adjusted association in a heterogeneous population. The issues of what to condition on for X and how to summarize $\mathrm{OR}(\cdot)$ have been recognized in the literature (e.g., Muller and MacLehose, 2014; Greenland et al., 1999). However, to the best of our knowledge, the average adjusted association has not been formally studied in the literature, let alone how to deal with high dimensional X. Indeed, many recently published papers in both

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

natural and social sciences have used the odds ratio to report their findings but, to the best of our knowledge, they are all based on strong parametric assumptions that restrict heterogeneous association: a logistic regression model is the most popular, where heterogeneity in association needs to be pre-specified. For example, recent studies on determinants of callbacks to job applications (Farber et al., 2016), association of Alzheimer's dementia with genotypes (Reiman et al., 2020), and police use of force with respect to race (Hoekstra and Sloan, 2022) are all based on logistic regression and pre-specified odds ratios. It is worth noting that odds ratios are natural in these examples because callbacks, dementia, and use of force are all rare events that occur with small probabilities.

As we depart from parametric models, the odds ratio can be a complicated function of observed characteristics. It is a natural and common practice to summarize a heterogeneous quantity by taking an average. For instance, when the treatment effect is heterogeneous, we frequently focus on the average treatment effect (e.g., Hirano et al., 2003; Ray and Szabo, 2019; Shi et al., 2019) as a summary parameter. Despite the popularity and usefulness of the average treatment effect and related quantities, the literature on such a statistic based on the odds ratios is surprisingly sparse. The only exception we are aware of is the Mantel-Haenszel approach that averages the (log) odds ratio across finite strata. This paper fills this important gap.

Treating the function $\log\{\mathrm{OR}(\cdot)\}$ nonparametrically, we derive two equivalent forms of the efficient influence function for θ_0, which we then explore to construct efficient double/debiased machine learning (DML) estimators of θ_0. We take this approach because the class of DML estimators are generally more advocated than conventional plug-in efficient estimators particularly when X is high-dimensional (see, e.g., Chernozhukov et al., 2018; Lewis and Syrgkanis, 2021, among many others). The efficient influence function can be expressed in terms of either prospective probabilities (i.e., the probabilities of the outcome conditional on the exposure and confounders) or retrospective ones (i.e., the probabilities of the exposure given the outcome and confounders). Therefore, we have two types of DML estimators: they are asymptotically equivalent and efficient, but they are not the same in finite samples. Our work is the first to derive the forms of the efficient influence function and to propose suitable DML estimators of θ_0. We also provide easy-to-follow computational and inferential algorithms for implementation. Given the basic nature of this research, we do not see potential negative societal impacts of our work, although we should mention that it generally requires much caution to draw causal inference from statistical association.

Related Literature There are a couple of recent papers on automatic DML estimators (Chernozhukov et al., 2022a,b). However, applying automatic orthogonalization to our setting is not necessarily better because (i) automation may induce additional approximation errors and (ii) the explicit formula obtained in this paper can provide useful insight into the estimation problem.

General theory for semiparametric efficiency is well-developed in the literature (e.g., Ai and Chen, 2012; Ackerberg et al., 2014). However, if we used the general framework, we would need to verify all the regularity conditions and our proof would not be self-contained. This type of verification is not needed for our setting because we can directly calculate the efficient influence function for θ_0 with a more preliminary proof technique.

Replication Files The replication files for all the numerical results are available at `https://github.com/sokbae/replication-JunLee-2023-AISTATS`.

2 AVERAGE ADJUSTED ASSOCIATION

The odds ratio can be expressed by using either prospective or retrospective probabilities: i.e., for all x in the support $\mathcal{X}$ of X,

$$\begin{aligned}\mathrm{OR}(x)&\\ &:= \frac{\mathbb{P}(Y=1 \mid T=1, X=x)}{\mathbb{P}(Y=1 \mid T=0, X=x)}\frac{\mathbb{P}(Y=0 \mid T=0, X=x)}{\mathbb{P}(Y=0 \mid T=1, X=x)},\\ &= \frac{\mathbb{P}(T=1 \mid Y=1, X=x)}{\mathbb{P}(T=1 \mid Y=0, X=x)}\frac{\mathbb{P}(T=0 \mid Y=0, X=x)}{\mathbb{P}(T=0 \mid Y=1, X=x)},\end{aligned}$$

where the second equality is known as the invariance property of the odds ratio, which can be verified by the Bayes rule (see, e.g., Cornfield, 1951). The two expressions of OR can be used to develop two different machine learning estimators.

Since the function OR is an infinite-dimensional object, it is generally difficult to estimate with high precision or even to communicate estimation results in a fully nonparametric manner, unless the dimension of x is limited to 1 or 2. In this context, we propose, as a scalar summary measure of association, to take the expectation of $\log \mathrm{OR}(\cdot)$ using the probability distribution of X. Specifically, we define

$$\theta_0 := \mathbb{E}\{\log \mathrm{OR}(X)\}, \tag{2.1}$$

which can be understood as the *Average Adjusted Association* (AAA) for the entire population. We have taken the logarithm before taking expectation because $\log \mathrm{OR}(x)$ corresponds to a coefficient in the traditional logistic model; for instance, if $\mathbb{P}(Y=1 \mid T=t, X=x) = G(\alpha_0 + \alpha_1 t + \alpha_2^\intercal x)$, where $G(s) = \exp(s)/\{1+\exp(s)\}$, then $\log \mathrm{OR}(x) = \alpha_1$ for all $x \in \mathcal{X}$. But the essence of our results does not rely on this structure; all of our results can be straightforwardly modified to the case of aggregating without the logarithm. Since $\mathrm{OR}(\cdot)$ is a nonlinear function in general, θ_0 differs from the log odds evaluated at the average value of X.

3 EFFICIENT INFLUENCE FUNCTION

We now characterize the efficient influence function for estimating θ_0 and present two equivalent expressions. We first state an assumption that will be used throughout the paper. Recall that Y and T are binary variables, and let $\mathcal{X}$ be the support of X.

Assumption 3.1. *There exists a constant $\epsilon > 0$ such that for all $t, y \in \{0, 1\}$ and for all $x \in \mathcal{X}$, we have $\epsilon \leq \mathbb{P}(Y = y, T = t \mid X = x) \leq 1 - \epsilon$.*

Assumption 3.1 is to ensure that all the four joint outcomes of (Y, T) occur with positive probability conditional on any value of $X = x \in \mathcal{X}$. Therefore, conditioning on any outcome of (Y, T) does not exclude any value of X from the support: i.e., assumption 3.1 implies that the joint support of (Y, T, X) is given by $\{0, 1\} \times \{0, 1\} \times \mathcal{X}$. This requirement may not be trivial in some applications, unless we restrict out attention to a certain subpopulation. For example, if Y represents prostate cancer, then it is reasonable to focus on the subpopulation of men. We are implicit about this type of (extra) conditioning throughout the analysis.

Also, under assumption 3.1, both of the prospective and retrospective representations of OR are well-defined for any $x \in \mathcal{X}$. Further, the existence of such an $\epsilon > 0$ in Assumption 3.1 guarantees that $x \mapsto \mathrm{OR}(x)$ is uniformly bounded from below by zero and from above by infinity.

Below we discuss the efficient influence function for θ_0 when assumption 3.1 is the only restriction imposed on the distribution of (Y, T, X).

For $y, t \in \{0, 1\}$, define

$$\begin{aligned} &\Delta_{pt}(Y, T, X) \\ &:= \frac{T^t(1-T)^{1-t}\{Y - \mathbb{P}(Y = 1 \mid T = t, X)\}}{\mathbb{P}(Y = 1 \mid T = t, X)\mathbb{P}(Y = 0 \mid T = t, X)}, \\ &\Delta_{ry}(Y, T, X) \\ &:= \frac{Y^y(1-Y)^{1-y}\{T - \mathbb{P}(T = 1 \mid Y = y, X)\}}{\mathbb{P}(T = 1 \mid Y = y, X)\mathbb{P}(T = 0 \mid Y = y, X)}. \end{aligned}$$

Further, define

$$\begin{aligned} &F_p(Y, T, X) \\ &:= \log \mathrm{OR}(X) - \theta_0 + \frac{\Delta_{p1}(Y, T, X)}{\mathbb{P}(T = 1 \mid X)} - \frac{\Delta_{p0}(Y, T, X)}{\mathbb{P}(T = 0 \mid X)}, \end{aligned} \tag{3.1}$$

$$\begin{aligned} &F_r(Y, T, X) \\ &:= \log \mathrm{OR}(X) - \theta_0 + \frac{\Delta_{r1}(Y, T, X)}{\mathbb{P}(Y = 1 \mid X)} - \frac{\Delta_{r0}(Y, T, X)}{\mathbb{P}(Y = 0 \mid X)}. \end{aligned} \tag{3.2}$$

The efficient influence function for θ_0 is given in the following theorem.

Theorem 3.1. *Suppose that Assumption 3.1 holds. Then, the semiparametrically efficient influence function for θ_0 is given by $F_p(Y, T, X) = F_r(Y, T, X)$.*

Theorem 3.1 includes equality between $F_p(Y, T, X)$ and $F_r(Y, T, X)$; F_p is based on the prospective expression of OR, whereas F_r uses the retrospective one. Theorem 3.1 is proved in two steps: (i) by direct calculation, as in Hahn (1998), θ_0 is pathwise differentiable along regular parametric submodels in the sense of Newey (1990, 1994); (ii) the pathwise derivative is an element of the tangent space, from which we can obtain the semiparametric efficiency bound V_{eff} for θ_0:

$$V_{\text{eff}} := \mathbb{E}\{F_p^2(Y, T, X)\} = \mathbb{E}\{F_r^2(Y, Y, X)\}.$$

The bound V_{eff} can be achieved by double/debiased machine learning (DML) estimators based on the efficient influence function, i.e., the representation in either (3.1) or (3.2). The DML approach has an advantage that it is robust to local perturbation on the unknown functions that need to be estimated in the first step, which is known as the Neyman orthogonality property (see, e.g., Chernozhukov et al., 2018).

In the remaining part of this section we formally show that the moment condition based on either (3.1) or (3.2) is indeed robust to local perturbation on the nonparametric components. For this purpose, note first that each of F_p and F_r depends on three nonparametric elements, i.e.,

$$\eta_{p0}(x) := \begin{pmatrix} \mathbb{E}(Y \mid T = 0, X = x) \\ \mathbb{E}(Y \mid T = 1, X = x) \\ \mathbb{E}(T \mid X = x) \end{pmatrix}, \tag{3.3}$$

$$\eta_{r0}(x) := \begin{pmatrix} \mathbb{E}(T \mid Y = 0, X = x) \\ \mathbb{E}(T \mid Y = 1, X = x) \\ \mathbb{E}(Y \mid X = x)] \end{pmatrix}, \tag{3.4}$$

respectively. Let $\mathcal{G}$ be a space of (measurable) functions on $\mathcal{X}$ such that $g \in \mathcal{G}$ satisfies $0 < \inf g(x) \leq \sup g(x) < 1$: see Assumption 3.1. Let $\tilde{F}_p(\cdot)[Y, T, X]$ and $\tilde{F}_r(\cdot)[Y, T, X]$ denote the functionals defined on $\mathcal{G}^3$ such that $\tilde{F}_p(\eta_{p0})[Y, T, X] = F_p(Y, T, X)$ and $\tilde{F}_r(\eta_{r0})[Y, T, X] = F_r(Y, T, X)$. Then, Neyman orthogonality is concerned about the Gateaux derivatives of $\tilde{F}_p$ and $\tilde{F}_r$ at η_{p0} and η_{r0}, respectively.

Theorem 3.2. *Suppose that Assumption 3.1 holds. Then, the Gateaux derivative of $\tilde{F}_p(\cdot)[Y, T, X]$ at η_{p0} has conditional mean zero given X almost surely: i.e.,*

$$\mathbb{E}\Big[\partial \tilde{F}_p\{\eta_{p0} + \gamma(\eta - \eta_{p0})\}[Y, T, X]/\partial\gamma\Big|_{\gamma=0} \Bigm| X\Big] = 0 \ \ a.s.$$

for any direction η. The same is true for $\tilde{F}_r(\cdot)[Y, T, X]$ at η_{r0} as well.

Theorem 3.2 says that both $F_p(Y,T,X)$ and $F_r(Y,T,X)$ provide Neyman orthogonal moments conditional on X. One implication of the local robustness property is that the first step nonparametric estimation of η_{p0} (or η_{r0}) in estimating θ_0 will not have any first-order consequence, i.e., the limiting distribution would be the same as if η_{p0} (respectively, η_{r0}) were known. In other words, all the adjustment terms that are needed to address the effect of the first step estimation are already reflected in F_p (respectively, in F_r).

4 DML ESTIMATORS

We now describe a couple of double/debiased machine learning (DML) estimators of θ_0, for which we use the functions F_p and F_r as estimating equations. We assume that a random sample $\{(Y_i, T_i, X_i^\intercal)^\intercal : i = 1, 2, \dots, n\}$ is available, where X_i is allowed to be high dimensional.

4.1 Computational Algorithms

Let $K \geq 2$ be some fixed integer (say, 5, 10 or 20). For simplicity, assume that n is divisible by K. Let $\{I_k : k = 1, \dots, K\}$ denote a K-fold partition of $\{1, \dots, n\}$ such that $|I_k| = n/K$ for each k. Suppose that one estimates all the conditional probabilities appearing in (3.1) or (3.2), depending on which one to use for estimation, via machine-learning estimators by using observations that belong to $I_k^c = \{1, \dots, n\} \setminus I_k$ for each k. Using data in I_k^c to estimate conditional probabilities evaluated at the points in I_k is reminiscent of the traditional leave-one-out method.

In using the prospective formula in (3.1), we start with

$$
\begin{aligned}
&\hat{p}_{p,t,k}(x) \\
&:= \widehat{\mathbb{P}}_{\mathrm{ML},k}(Y = 1 \mid T = t, X = x) \quad \text{for } t = 0, 1, \\
&\widehat{\mathrm{OR}}_{p,k}(x) \\
&:= \frac{\widehat{\mathbb{P}}_{\mathrm{ML},k}(Y = 1 \mid T = 1, X = x)}{\widehat{\mathbb{P}}_{\mathrm{ML},k}(Y = 0 \mid T = 1, X = x)} \\
&\times \frac{\widehat{\mathbb{P}}_{\mathrm{ML},k}(Y = 0 \mid T = 0, X = x)}{\widehat{\mathbb{P}}_{\mathrm{ML},k}(Y = 1 \mid T = 0, X = x)}, \\
&\widehat{w}_{p,k}(x) \\
&:= \widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 1 \mid X = x),
\end{aligned}
$$

where $\widehat{\mathbb{P}}_{\mathrm{ML},k}$ denotes a machine-learning estimator of a probability model using observations that belong to I_k^c. We then define the prospective DML estimator $\widehat{\theta}_p$ of θ_0 by

$$
\widehat{\theta}_p := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \widehat{\psi}_{i,p,k}, \tag{4.1}
$$

where

$$
\begin{aligned}
\widehat{\psi}_{i,p,k} &:= \log \widehat{\mathrm{OR}}_{p,k}(X_i) \\
&\quad + \frac{T_i}{\widehat{w}_{p,k}(X_i)} \frac{\{Y_i - \hat{p}_{p,1,k}(X_i)\}}{\hat{p}_{p,1,k}(X_i)\{1 - \hat{p}_{p,1,k}(X_i)\}} \\
&\quad - \frac{(1 - T_i)}{\{1 - \widehat{w}_{p,k}(X_i)\}} \frac{\{Y_i - \hat{p}_{p,0,k}(X_i)\}}{\hat{p}_{p,0,k}(X_i)\{1 - \hat{p}_{p,0,k}(X_i)\}}.
\end{aligned} \tag{4.2}
$$

The estimator $\hat{\theta}_p$ is asymptotically normal and efficient as we will show in Section 4.2. We have summarized the estimation procedure in Algorithm 1.

Algorithm 1: Prospective DML estimator of θ_0

```
Input: $\{(Y_i, T_i, X_i) : i = 1, \dots, n\}$, integer $K \geq 2$,
       machine learning methods for estimating
       probability models
Output: estimate of $\theta_0$ and its standard error
Construct a $K$-fold partition $\{I_k : k = 1, \dots, K\}$ of
  $\{1, \dots, n\}$ of approximately equal size;
For each $k$, use observations belonging to $I_k^c$ to obtain
  machine learning estimates of
  $\mathbb{P}(Y = 1 \mid T = 1, X = x)$,
  $\mathbb{P}(Y = 1 \mid T = 0, X = x)$ and $\mathbb{P}(T = 1 \mid X = x)$,
  respectively;
For each $k$, use observations belonging to $I_k$ to
  construct $\widehat{\psi}_{i,p,k}$ in equation (4.2);
Obtain the estimate of $\theta_0$ by equation (4.1) and its
  standard error $\widehat{\sigma}_p / \sqrt{n}$ by

    $\widehat{\sigma}_p^2 := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \left\{ \widehat{\psi}_{i,p,k} - \widehat{\theta}_p \right\}^2.$   (4.3)
```

The retrospective DML estimator $\widehat{\theta}_r$ of θ_0 is defined analogously. That is, we start with

$$
\begin{aligned}
&\hat{p}_{r,y,k}(x) \\
&:= \widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 1 \mid Y = y, X = x) \quad \text{for } y = 0, 1, \\
&\widehat{\mathrm{OR}}_{r,k}(x) \\
&:= \frac{\widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 1 \mid Y = 1, X = x)}{\widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 0 \mid Y = 1, X = x)} \\
&\times \frac{\widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 0 \mid Y = 0, X = x)}{\widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 1 \mid Y = 0, X = x)}, \\
&\widehat{w}_{r,k}(x) \\
&:= \widehat{\mathbb{P}}_{\mathrm{ML},k}(Y = 1 \mid X = x),
\end{aligned}
$$

and we define

$$
\widehat{\theta}_r := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \widehat{\psi}_{i,r,k}, \tag{4.4}
$$

where

$$\begin{aligned}\widehat{\psi}_{i,r,k} &:= \log \widehat{\mathrm{OR}}_{r,k}(X_i) \\ &+ \frac{Y_i}{\widehat{w}_{r,k}(X_i)} \frac{\{T_i - \hat{p}_{r,1,k}(X_i)\}}{\hat{p}_{r,1,k}(X_i)\{1-\hat{p}_{r,1,k}(X_i)\}} \\ &- \frac{(1-Y_i)}{\{1-\widehat{w}_{r,k}(X_i)\}} \frac{\{T_i - \hat{p}_{r,0,k}(X_i)\}}{\hat{p}_{r,0,k}(X_i)\{1-\hat{p}_{r,0,k}(X_i)\}}. \end{aligned} \tag{4.5}$$

The algorithm for the retrospective DML estimator can be stated easily by making simple modifications in Algorithm 1. We omit details for brevity.

Before we finish this subsection, we remark that the consistency of $\widehat{\theta}_p$ and $\widehat{\theta}_r$ does not require that $\widehat{w}_{p,k}$ and $\widehat{w}_{r,k}$ consistently estimate w_p and w_r; indeed, $F_p(Y,T,X)$ and $F_r(Y,T,X)$ have mean zero even if w_p and w_r deviate from the truth.

4.2 Asymptotic Distributions

Let $\|\cdot\|_{P,2}$ denote the $L_2(P)$-norm, where P is the probability distribution of (Y,T,X): i.e., $\|a\|_{P,2} := \max_{1\le\ell\le d}\left\{\mathbb{E}[a_\ell^2(Y,T,X)]\right\}^{1/2}$ for a d-dimensional vector-valued function $a := (a_1,\ldots,a_d)$. For each k, let

$$\widehat{\eta}_{n,p,k}(X) := \begin{pmatrix} \widehat{\mathbb{P}}_{\mathrm{ML},k}(Y=1 \mid T=0,X) \\ \widehat{\mathbb{P}}_{\mathrm{ML},k}(Y=1 \mid T=1,X) \\ \widehat{\mathbb{P}}_{\mathrm{ML},k}(T=1 \mid X) \end{pmatrix}$$

denote the vector of machine learning estimators of η_{p0} defined in (3.3), using observations belonging to I_k^c. The first step estimators $\widehat{\eta}_{n,p,k}$ are inputs to the second step in (4.1). Likewise, let $\widehat{\eta}_{n,r,k}$ be the vector of machine learning estimators of θ_{r0} defined in (3.4), which will be used for the retrospective DML estimator of θ_0.

Assumption 4.1 (First-Stage Estimation)**.** *There exist sequences $\delta_n \ge n^{-1/2}$ and τ_n of positive constants both approaching zero such that for each $k=1,\ldots,K$,*

$$\begin{aligned}\|\widehat{\eta}_{n,p,k} - \eta_{p0}\|_{P,2} &\le \delta_n n^{-1/4}, \\ \|\widehat{\eta}_{n,r,k} - \eta_{r0}\|_{P,2} &\le \delta_n n^{-1/4},\end{aligned}$$

with probability no less than $1-\tau_n$.

Assumption 4.1 is a high-level assumption that may not be trivial if X is high dimensional. For instance, it may fail even when all the conditional probabilities are logistically specified and they are estimated by the method of maximum likelihood if X is high dimensional (e.g., Sur and Candès, 2019; Zhao et al., 2022). However, Assumption 4.1 is known to be attainable for a variety of machine learning methods. The primitive conditions for ℓ_1-penalized logit estimators are worked out by van de Geer (2008) and Belloni et al. (2016) among others. A maximum likelihood approach with some adjustment for the dimensionality and signal strength of X as in e.g., Yadlowsky et al. (2021) is another possibility although we do not pursue the latter in this paper.

An application of Theorems 3.1 and 3.2 of Chernozhukov et al. (2018) gives the following result that formally justifies the estimation and inference methods proposed in Section 4.1.

Theorem 4.1. *Let $\{\mathcal{P}_n : n \ge 1\}$ be a sequence of sets of probability distributions of (Y,T,X). Suppose that for all $n \ge 3$ and $P \in \mathcal{P}_n$, Assumptions 3.1 and 4.1 hold and that we have a random sample $\{(Y_i,T_i,X_i^\intercal)^\intercal : i=1,\ldots,n\}$. Then, uniformly over $P \in \mathcal{P}_n$,*

$$\sqrt{n}\frac{(\widehat{\theta}_p - \theta_0)}{\widehat{\sigma}_p} \to_d \mathbb{N}(0,1) \text{ and } \sqrt{n}\frac{(\widehat{\theta}_r - \theta_0)}{\widehat{\sigma}_r} \to_d \mathbb{N}(0,1).$$

Furthermore, both $\widehat{\sigma}_p^2 \to_p V_{\mathit{eff}}$ and $\widehat{\sigma}_r^2 \to_p V_{\mathit{eff}}$ uniformly over $P \in \mathcal{P}_n$.

Theorem 4.1 establishes that both the prospective and retrospective DML estimators are asymptotically normal and efficient. However, asymptotic equivalence does not imply that it is irrelevant which one to use between $\widehat{\theta}_p$ and $\widehat{\theta}_r$ in finite samples. In fact, the first steps of the two estimators involve different nonparametric regression functions, and therefore they generally lead to different estimates. Comparing the estimates and their standard errors can be a useful diagnostic check in practice; we recommend reporting inference results based on both estimators.

5 FURTHER DISCUSSIONS

5.1 Nonparametric Estimation and Double Robustness

Since we do not use any parametric specification to identify and estimate θ_0, potential misspecification is not a concern, at least asymptotically: it is a concern only to the extent that our choices of nonparametric estimators must satisfy Assumption 4.1. However, nonparametric approaches do rely on several input parameters, which are important for the performance of the estimators in finite samples. In this regard we show that there is a nonparametric version of double robustness.

For every $x \in \mathcal{X}$, we implicitly *define* $\varphi_{p0}(x)$, $\varphi_{r0}(x)$, and $\vartheta_0(x)$ by the following equations: for $t,y \in \{0,1\}$,

$$\mathbb{P}(Y=1 \mid T=t, X=x) = \frac{\exp\{\varphi_{p0}(x) + t\vartheta_0(x)\}}{1+\exp\{\varphi_{p0}(x)+t\vartheta_0(x)\}}, \tag{5.1}$$

$$\mathbb{P}(T=1 \mid Y=y, X=x) = \frac{\exp\{\varphi_{r0}(x) + y\vartheta_0(x)\}}{1+\exp\{\varphi_{r0}(x)+y\vartheta_0(x)\}}. \tag{5.2}$$

For instance, equation (5.1) with $t=0$ defines $\varphi_{p0}(x)$. Here, we have four conditional probabilities to define three objects. This is because the four conditional probabilities are restricted by the invariance property of the odds ratio. Indeed, simple algebra shows that $\vartheta_0(x) = \log \mathrm{OR}(x)$, on which no restrictions have been imposed.

Let $\mathcal{H}$ be the class of functions on $\mathcal{X}$, which contains $\varphi_{p0}, \varphi_{r0}$, and ϑ_0. We then consider the function $m : \mathcal{H}^3 \times \{0,1\}^2 \times \mathcal{X} \to \mathbb{R}$ defined by

$$\begin{aligned} &m(\varphi_p, \varphi_r, \vartheta, Y, T, X) \\ &= \{Y - \Lambda_0(\varphi_p, X)\}\{T - \Lambda_0(\varphi_r, X)\}\exp\{-\vartheta(X)TY\}, \end{aligned}$$

where $\Lambda_0(\varphi, x) = \exp\{\varphi(x)\}/\big[1 + \exp\{\varphi(x)\}\big]$.

Theorem 5.1. *Suppose that Assumption 3.1 holds. Then, for any $\varphi_p, \varphi_r \in \mathcal{H}$, we have*

$$\begin{aligned} &\mathbb{E}\{m(\varphi_{p0}, \varphi_r, \vartheta_0, Y, T, X) \mid X\} \\ &\qquad = \mathbb{E}\{m(\varphi_p, \varphi_{r0}, \vartheta_0, Y, T, X) \mid X\} = 0 \quad a.s. \end{aligned}$$

Further, for any $\vartheta \in \mathcal{H}$ and $x \in \mathcal{X}$ such that $\vartheta(x) \neq \vartheta_0(x)$,

$$\mathbb{E}\{m(\varphi_{p0}, \varphi_{r0}, \vartheta, Y, T, X) \mid X = x\} \neq 0 \quad a.s.$$

Theorem 5.1 is a nonparametric extension of the double robustness idea of Tchetgen Tchetgen et al. (2010) and Tchetgen Tchetgen (2013): i.e., either $\mathbb{P}(Y = 1 \mid T = 0, X = x) = \varphi_{p0}(x)$ or $\mathbb{P}(T = 1 \mid Y = 0, X = x) = \varphi_{r0}(x)$ (but not both) needs to be correctly specified to estimate $\vartheta_0(x) = \log \mathrm{OR}(x)$ consistently. Further, the second assertion shows that the function m can be used to identify ϑ_0 once φ_{p0} and φ_{r0} are given by their definition. It can be shown that the efficiency bound for estimating $\theta_0 = \mathbb{E}\{\vartheta_0(X)\}$ by using the conditional moment equation in Theorem 5.1 is given by

$$\begin{aligned} &F_{\mathrm{DR}}(Y, T, X) = \vartheta_0(X) - \theta_0 \\ &\quad + \frac{1}{\mathbb{P}(Y = 1, T = 1 \mid X)} \frac{m(\varphi_{p0}, \varphi_{r0}, \vartheta_0, Y, T, X)}{m(\varphi_{p0}, \varphi_{r0}, \vartheta_0, 1, 1, X)}, \end{aligned}$$

which coincides with $F_p(Y, T, X) = F_r(Y, T, X)$. Therefore, we do not lose anything in terms of semiparametric efficiency in using the moment function m to estimate θ_0.

Therefore, it is possible to have an extra DML-based doubly robust algorithm for efficient estimation of θ_0. However, we do not pursue this possibility in the current paper for a couple of reasons. As we described in Section 4, the DML approach requires splitting the sample into multiple subsamples, but it often causes computational issues in practice when multiple tiers of nonparametric estimation are involved as in the current setup. Also, since our approach does not require any parametric specification at all, double robustness seems to have limited merits; there is no misspecification in the limit in the approach described in Section 4.

5.2 Interpretation: Association vs. Causation

One can rely on Theorem 4.1 to conduct statistical inference on θ_0. For instance, using the prospective estimator $\hat{\theta}_p$, a (symmetric two-sided) 95% confidence interval for θ_0 can be obtained in the usual manner, i.e., $\hat{\theta}_p \pm 1.96 \cdot \hat{\sigma}_p/\sqrt{n}$. Here, we emphasize that θ_0 is understood as a simple association parameter to which we do not give any causal interpretation at this stage. However, θ_0 can be used for causal inference under a few extra assumptions.

In order to discuss causal inference, let $Y(t)$ denote the potential outcome when the treatment is exogenously fixed at t; i.e., the observed outcome Y is equal to $Y(T) = Y(1)T + Y(0)(1 - T)$. Then, $\mathrm{OR}(x)$ is related with the following causal parameters:

$$\begin{aligned} \vartheta_{\mathrm{OR}}(x) &= \frac{\mathbb{P}\{Y(1) = 1 \mid X = x\}}{\mathbb{P}\{Y(1) = 0 \mid X = x\}} \frac{\mathbb{P}\{Y(0) = 0 \mid X = x\}}{\mathbb{P}\{Y(0) = 1 \mid X = x\}}, \\ \vartheta_{\mathrm{RR}}(x) &= \frac{\mathbb{P}\{Y(1) = 1 \mid X = x\}}{\mathbb{P}\{Y(0) = 1 \mid X = x\}}. \end{aligned}$$

That is, $\vartheta_{\mathrm{OR}}(x)$ represents the adjusted causal odds ratio, while $\vartheta_{\mathrm{RR}}(x)$ is the adjusted causal relative risk parameter. Below we summarize some facts that are known in the literature; see, e.g., Holland and Rubin (1988) and Jun and Lee (2021) for more detail.

1. (*No confounding*) Suppose that there is no confounding conditional on $X = x$: i.e., for $t = 0, 1$, $Y(t)$ and T are independent given $X = x$. Then, $\mathrm{OR}(x) = \vartheta_{\mathrm{OR}}(x)$.

2. (*No confounding + MTR*) If there is no confounding conditional on $X = x$ and the treatment is potentially beneficial but it never hurts, i.e., $Y(1) \geq Y(0)$ almost surely, which is termed as the Monotone Treatment Response (MTR) assumption (e.g., Manski, 1997), then $1 \leq \theta_{\mathrm{RR}}(x) \leq \vartheta_{\mathrm{OR}}(x) = \mathrm{OR}(x)$, where the inequalities are sharp.

3. (*Confounding + MTR/MTS*) Suppose that there may be confounding even if we condition on $X = x$ but that the treatment is potentially beneficial (MTR). Further, suppose that those who have chosen to take the treatment have no smaller chance of "success" than those who have opted out, i.e., $\mathbb{P}\{Y(t) = 1 \mid T = 1, X = x\} \geq \mathbb{P}\{Y(t) = 1 \mid T = 0, X = x\}$, which is termed as the Monotone Treatment Selection (MTS) assumption (e.g., Manski and Pepper, 2000). Then, $1 \leq \theta_{\mathrm{RR}}(x) \leq \mathrm{OR}(x)$, where the inequalities are sharp.

Therefore, the average parameter θ_0 and Theorem 4.1 can be used for causal inference on the entire population. For example, if the researcher is willing to assume that the treatment was randomly assigned conditional on

X, then the usual symmetric confidence interval such as $\hat{\theta}_r \pm 1.96 \cdot \hat{\sigma}_r$ is a confidence interval for $\mathbb{E}\{\vartheta_{\mathrm{OR}}(X)\}$. There is no general relationship between $\mathbb{E}\{\vartheta_{\mathrm{OR}}(X)\}$ and $\mathbb{E}\{\vartheta_{\mathrm{RR}}(X)\}$, but if $\mathbb{P}(Y = 1)$ is close to zero (known as the rare-disease assumption), then the two causal parameters are known to be close to each other as long as there is no confounding given X. So, the symmetric confidence interval of θ_0 can also be understood as an approximate confidence interval of $\mathbb{E}\{\vartheta_{\mathrm{RR}}(X)\}$ in this scenario.

The no-confounding assumption or the rare-disease assumption can be unrealistic in some applications. For instance, many treatments of interest in social sciences such as education choices are deliberate decisions, and in such cases the no-confounding assumption is unrealistic. Even so, if one is willing to assume that education is potentially beneficial but it never hurts and that those who deliberately chose to take higher education is generally no less likely to "succeed" than those who did not, then θ_0 can be understood as a sharp upper bound on $\mathbb{E}\{\vartheta_{\mathrm{RR}}(X)\}$. Therefore, a one-sided confidence interval such as $[1,\ \hat{\theta}_r + 1.64 \cdot \hat{\sigma}_r]$ can be reported if the causal parameter $\mathbb{E}\{\vartheta_{\mathrm{RR}}(X)\}$ is of interest.

5.3 Averaging over a Subpopulation

In practice there may be a subpopulation of particular interest, in which case averaging over the entire population may not provide the most relevant summary statistic. For example, consider the population of patients with a certain type of cancer. Suppose that Y and T, respectively, indicate five-year survival (say, $Y = 1$ for survival and $Y = 0$ for death) and a certain type of treatment (say, $T = 1$ for treatment and $T = 0$ for no treatment). Here, it may be relevant to summarize the association between Y and T for those who received the treatment, i.e., $\theta_T(1) := \mathbb{E}\{\log \mathrm{OR}(X) \mid T = 1\}$. Alternatively, the association between Y and T for those who survived the cancer, which can be captured by $\theta_Y(1) := \mathbb{E}\{\log \mathrm{OR}(X) \mid Y = 1\}$, can be an interesting quantity to look at. Of course, if the average adjusted association between T and Y is homogeneous across X, then there will be no difference among $\theta_0, \theta_T(1)$, and $\theta_Y(1)$. However, in general, they are all distinct and can be substantially different, depending on the degree of heterogeneity.

Also, $\theta_T(1)$ and $\theta_Y(1)$ can be of interest if our access to a random sample is limited. For example, $\theta_T(1)$ is point identifiable even when we only have a treatment-based sample, where a half of the sample comes from the patients who received the treatment and the other half is from those who did not. Similarly, an outcome-based sample (e.g., case-control studies) is sufficient to identify $\theta_Y(1)$. The statistical analysis of $\theta_T(1)$ and $\theta_Y(1)$ is similar to that of θ_0 and we do not repeat it here.

In addition, there has been increasing interest in estimating individual level treatment effects (e.g., Shalit et al., 2017). Averaging over the entire population may have a risk of over-simplification: for example, if the association is positive for some values of X, and it is negative for other values of X, then the overall average association may be close to zero. With this motivation in mind, suppose that there is a vector of low-dimensional confounders (say, Z) such that we are interested in estimating $z \mapsto \mathbb{E}\{\log \mathrm{OR}(X) \mid Z = z\}$. Then it can be estimated by projecting $\log \mathrm{OR}(X)$ on a low-dimensional space of Z as in Ogburn et al. (2015), Lee et al. (2017), and Semenova and Chernozhukov (2020). However, it is a topic of future research to develop this idea formally.

6 NUMERICAL STUDIES

In this section, we provide numerical results to illustrate the usefulness of our approach.

6.1 Top Income and Higher Education

We start with a real-data example. Table 1 summarizes data from American Community Survey (ACS) 2018, cross-tabulating the likelihood of top income by educational attainment. The sample is restricted to white males residing in California with at least a bachelor's degree. It is extracted from IPUMS USA (Ruggles et al., 2019). The ACS is an ongoing annual survey by the US Census Bureau that provides key information about the US population.

Table 1: Top income and education

	Beyond bachelor's		Total
Top income	$T = 0$	$T = 1$	
$Y = 0$	10,533	6,362	16,895
$Y = 1$	397	524	921
Total	10,930	6,886	17,816

The binary outcome variable 'Top income' (Y) is defined to be one if a respondent's annual total pre-tax wage and salary income is top-coded. In ACS 2018, the threshold income for top-coding is different across states. In our sample extract, the top-coded income bracket has median income \$565,000 and the next highest income that is not top-coded is \$327,000. The binary exposure variable (T) is defined to be one if a respondent has a master's degree, a professional degree, or a doctoral degree.

To adjust for individual differences, we include age and industry code as covariates (X). In particular, cubic B-splines of age with 17 inner knots as well as 254 industry dummies are included in this specification, which can be viewed as a high-dimensional setting.

Specifically, we implement ℓ_1-penalized logistic estimation with `glmnet` package in R (Friedman et al., 2010) to estimate $\mathbb{P}(Y = 1|T = t, X = x), t = 0, 1$ and

$\mathbb{P}(T = 1|X = x)$ for the prospective model (respectively, $\mathbb{P}(T = 1|Y = y, X = x), y = 0, 1$ and $\mathbb{P}(Y = 1|X = x)$ for the retrospective model) with 10-fold cross-fitting. The underlying assumption here is that the B-spline terms plus the industry dummies are rich enough to approximate $\mathbb{P}(Y = 1|T = t, X = x)$ as well as $\mathbb{P}(T = 1|X = x)$ for the prospective model (respectively, $\mathbb{P}(T = 1|Y = y, X = x)$ as well as $\mathbb{P}(Y = 1|X = x)$ for the retrospective model). The penalization tuning parameter is chosen by cross-validation (that is, `lambda.min` in the `glmnet` package). Here, we focus on ℓ_1-penalized estimators among other possible machine learning estimators because the primitive conditions for Assumption 4.1 are well established for ℓ_1-penalized logit estimators (e.g., van de Geer, 2008; Belloni et al., 2016), as we mentioned in Section 4.2.

Table 2: Numerical results

Panel A: θ_0	Prospective	Retrospective
Estimate	0.72	0.71
Standard Error	(0.12)	(0.10)
Panel B: $\exp(\theta_0)$	Prospective	Retrospective
Estimate	2.06	2.04
95% Confidence Interval	[1.61,2.63]	[1.67,2.49]

Table 2 reports estimation results. Looking at Panel A, the prospective estimate of θ_0 is 0.72, which is almost the same as the retrospective estimate of 0.71. In Panel B, we present point estimates of $\exp(\theta_0)$ and its confidence intervals using asymptotic normality obtained in Theorem 4.1.

The estimates of $\exp(\theta_0)$ are comparable to the usual odds ratio in terms of its scale; therefore, they can be interpreted similarly. Furthermore, as can be seen from Table 1, top income is a rare event (that is, the sample proportion of $\mathbb{P}(Y = 1)$ is approximately 0.05). When the outcome of interest is rare, an average of the conditional odds ratio approximates the average of conditional relative risk, namely the average of the ratio between $\mathbb{P}(Y = 1|T = 1, X)/\mathbb{P}(Y = 1|T = 0, X)$. Hence, obtaining a higher-level degree is associated with doubling the chance of earning very high incomes. The 95% confidence interval for the prospective estimate is $[1.61, 2.63]$ (respectively, [1.67,2.49] for the retrospective model). As discussed in Section 5.2, θ_0 can be interpreted as the upper bound on the causal parameter $\mathbb{E}\{\log \vartheta_{\mathrm{RR}}(X)\}$ if one assumes the MTR/MTS assumptions here.

Recall that the covariates consists of cubic B-splines of age as well as industry dummies, resulting in 274 regressors. As $n = 17, 816$, one may simply try to estimate a parametric logistic regression model with the same set of regressors. However, it turns out that this flexible parametric approach suffers from a couple of numerical issues: (i) a very small number of estimated coefficients are `NA` due to multicollinearity; (ii) some of predicted probabilities are numerically zero. As a result, the conditional odds ratios are not defined for all values of the regressors. To resolve these problems, we make some ad hoc adjustments: (i) we ignore the problematic regressors by setting their coefficients to be zero; (ii) we set a lower bound on the fitted probabilities. Specifically, any fitted probability less than 1e-6 is set to be 1e-6. Then, we estimate θ_0 by simply plugging the fitted probabilities into the formula of $\theta_0 = \mathbb{E}\{\log \mathrm{OR}(X)\}$. The parametric plug-in estimates with the ad hoc adjustments turn out to be 0.38 (prospective estimate) and 0.88 (retrospective estimate). The large difference between the prospective and the retrospective estimates indicates that there is an anomaly in the plug-in estimates. In addition, we also consider plug-in estimation of θ_0 using ℓ_1-penalized logistic estimation with the same specifications and tuning parameters as in DML estimation. Hence, in this case, the plug-in estimator is different from the DML estimator in that (i) it uses a different estimating equation and (ii) it does not use cross-fitting. The resulting plug-in estimates are 0.78 (prospective estimate) and 0.68 (retrospective estimate). They look more similar to the DML estimators; however, there is no theoretically proven result regarding how to conduct inference with the ℓ_1-penalized plug-in estimators.

6.2 A Monte Carlo Experiment

We turn to a Monte Carlo experiment to make a more systematic comparison between the plug-in and DML estimators. We generate observations in the following way: (i) the covariates are randomly drawn from the empirical distribution of the ACS sample; (ii) the binary exposure variable is generated from a logit model with $\mathbb{P}(T = 1|X) = G(\alpha_0 + \alpha_1 Age + \alpha_2 Age^2)$, where $G(\cdot)$ is the logit link function and the parameters $(\alpha_0, \alpha_1, \alpha_2)$ are chosen by fitting the logit model with the ACS sample; (iii) the binary outcome variable is drawn from a logit model with $\mathbb{P}(Y = 1|T, X) = G(\beta_0 + \beta_1 T + \beta_2 Age + \beta_3 Age^2)$, where the parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$ are again chosen by fitting a logit model with the ACS sample. In this experimental design, the true model is such that $\theta_0 = \beta_1$ and the industry effects are null. However, we fit exactly the same specifications as in the previous real-data example to examine the differences between the plug-in and DML estimators. The only change made here is that 5-fold cross-validation is adopted for DML estimation to speed up Monte Carlo simulations. The sample size is $n = 5, 000$ and the number of Monte Carlo replications is 500.

Table 3 summarizes the results of the experiments. The prospective DML estimator has a much smaller mean bias than the prospective plug-in estimator without increasing the standard deviation. Further, its size (the probability of excluding the true value of θ_0 in the confidence interval) is close to the 10% nominal level. The retrospective DML estimator does not perform as well as the prospective DML

Table 3: Results of the Monte Carlo Experiment

Estimator	Mean Bias	Standard Deviation	Size (10%)
Prospective plug-in	0.12	0.16	NA
Retrospective plug-in	0.12	0.16	NA
Prospective DML	0.05	0.16	0.89
Retrospective DML	0.09	0.16	0.84

estimator. This is due to the fact that experimental data are generated from a prospective logit model. Overall, the results of the experiment verify that the DML estimators are superior to the plug-in estimators when the underlying machine learning estimators are the ℓ_1-penalized logistic regression estimators.

7 CONCLUSIONS

Our proposed DML estimators offer a novel way of estimating the summary measure of association, namely the AAA functional θ_0. In particular, we provide a method for statistical inference on θ_0 based on asymptotic normality of our efficient DML estimators.

This paper has focused on binary outcome and exposure. However, it is possible to define the AAA functional beyond the current setup. For example, following Tchetgen Tchetgen et al. (2010), we can define the conditional odds ratio function as

$$\mathrm{OR}(x) := \frac{f(y \mid t, x)}{f(y \mid t_0, x)} \frac{f\left(y_0 \mid t_0, x\right)}{f\left(y_0 \mid t, x\right)},$$

where Y and T can take either discrete values, continuous values, or a mixture of both; (y_0, t_0) is a user specified point in the sample space; and $f(y \mid t, x)$ is the conditional density of Y given $T = t$ and $X = x$ with respect to a dominating measure μ. It is a topic of future research to develop this idea formally.

Acknowledgements

We would like to thank a meta-reviewer and four anonymous reviewers for helpful comments.

References

Ackerberg, D., Chen, X., Hahn, J., and Liao, Z. (2014). Asymptotic Efficiency of Semiparametric Two-step GMM. *Review of Economic Studies*, 81(3):919–943.

Ai, C. and Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170(2):442–457.

Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619.

Bland, J. M. and Altman, D. G. (2000). The odds ratio. *BMJ*, 320(7247):1468.

Breslow, N. (1976). Regression analysis of the log odds ratio: A method for retrospective studies. *Biometrics*, pages 409–416.

Breslow, N. and Powers, W. (1978). Are there two logistic regressions for retrospective studies? *Biometrics*, pages 100–105.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research I. The Analysis of Case-Control Studies*, volume 1. International Agency for Research on Cancer, Lyon, France.

Chen, H. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, 63(2):413–421.

Chen, Z., Shi, N.-Z., and Gao, W. (2011). Nonparametric estimation of the log odds ratio for sparse data by kernel smoothing. *Statistics & probability letters*, 81(12):1802–1807.

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

Chernozhukov, V., Newey, W., Quintas-Martínez, V. M., and Syrgkanis, V. (2022a). RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3901–3914. PMLR.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022b). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11(6):1269–1275.

Farber, H. S., Silverman, D., and Von Wachter, T. (2016). Determinants of callbacks to job applications: An audit study. *American Economic Review*, 106(5):314–318.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46.

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

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

Hoekstra, M. and Sloan, C. (2022). Does race matter for police use of force? evidence from 911 calls. *American economic review*, 112(3):827–860.

Holland, P. W. and Rubin, D. B. (1988). Causal inference in retrospective studies. *Evaluation Review*, 12(3):203–231.

Hui, F. K. and Geenens, G. (2013). A nonparametric measure of local association for two-way contingency tables. *Computational Statistics & Data Analysis*, 68:98–110.

Jun, S. J. and Lee, S. (2021). Causal inference under outcome-based sampling with monotonicity assumptions. arXiv:2004.08318.

Lee, S., Okui, R., and Whang, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, 32(7):1207–1225.

Lewis, G. and Syrgkanis, V. (2021). Double/debiased machine learning for dynamic treatment effects. In *Advances in Neural Information Processing Systems*, volume 34.

Manski, C. F. (1997). Monotone treatment response. *Econometrica*, 65(6):1311–1334.

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1010.

Muller, C. J. and MacLehose, R. F. (2014). Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. *International journal of epidemiology*, 43(3):962–970.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382.

Norton, E., Dowd, B., and Maciejewski, M. (2018). Odds ratios–current best practice and use. *JAMA*, 320(1):84–85.

Ogburn, E. L., Rotnitzky, A., and Robins, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396.

Ray, K. and Szabo, B. (2019). Debiased Bayesian inference for average treatment effects. In *Advances in Neural Information Processing Systems*, volume 32.

Reiman, E. M., Arboleda-Velasquez, J. F., Quiroz, Y. T., Huentelman, M. J., Beach, T. G., Caselli, R. J., Chen, Y., Su, Y., Myers, A. J., Hardy, J., et al. (2020). Exceptionally low likelihood of alzheimer's dementia in apoe2 homozygotes from a 5,000-person neuropathological study. *Nature communications*, 11(1):667.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., and Sobek, M. (2019). IPUMS USA: Version 9.0 [dataset]. `https://doi.org/10.18128/D010.V9.0`.

Semenova, V. and Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *Econometrics Journal*, 24(2):264–289.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, volume 32.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Tchetgen Tchetgen, E. J. (2013). On a closed-form doubly robust estimator of the adjusted odds ratio for a binary exposure. *American journal of epidemiology*, 177(11):1314–1316.

Tchetgen Tchetgen, E. J., Robins, J. M., and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97(1):171–180.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645.

Yadlowsky, S., Yun, T., McLean, C. Y., and D'Amour, A. (2021). SLOE: A faster method for statistical inference in high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, volume 34.

Zhao, Q., Sur, P., and Candes, E. J. (2022). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861.

Supplementary Materials for "Average Adjusted Association: Efficient Estimation with High Dimensional Confounders"

A Proofs of Theorems

Proof of Theorem 3.1: The proof of Theorem 3.1 has two parts. We first establish the equivalence result in Theorem 3.1 and then show that $F_r(Y, T, X)$ is the efficient influence function of θ_0.

Proof of the equivalence result in Theorem 3.1:

The claim of $F_p(Y, T, X) = F_r(Y, T, X)$ can be verified by checking the four cases of (Y, T) equal to $(0, 0)$, $(0, 1)$, $(1, 0)$, or $(1, 1)$. Below we will do this and show that

$$\frac{\Delta_{p1}(Y,T,X)}{\mathbb{P}(T=1 \mid X)} - \frac{\Delta_{p0}(Y,T,X)}{\mathbb{P}(T=0 \mid X)} = \frac{\Delta_{r1}(Y,T,X)}{\mathbb{P}(Y=1 \mid X)} - \frac{\Delta_{r0}(Y,T,X)}{\mathbb{P}(Y=0 \mid X)}. \tag{A.1}$$

In what follows we will use abbreviations like $P_{T|X}(t|x), P_{Y|TX}(y|t,x)$, etc., to denote $\mathbb{P}(T = t|X = x), \mathbb{P}(Y = y|T = t, X = x)$, and similar objects. Now, the left-hand side of equation (A.1) can be expressed as

$$TYa_{TY}(X) - Ta_T(X) - Ya_Y(X) + a_o(X),$$

where

$$\begin{aligned}
a_{TY}(x) &= \frac{1}{P_{T|X}(1|X)P_{Y|TX}(1|1,x)P_{Y|TX}(0|1,x)} + \frac{1}{P_{T|X}(0|x)P_{Y|TX}(1|0,x)P_{Y|TX}(0|0,x)},\\
a_T(x) &= \frac{1}{P_{T|X}(1|x)P_{Y|TX}(0|1,x)} + \frac{1}{P_{T|X}(0|x)P_{Y|TX}(0|0,x)},\\
a_Y(x) &= \frac{1}{P_{T|X}(0|x)P_{Y|TX}(1|0,x)P_{Y|TX}(0|0,x)},\\
a_o(x) &= \frac{1}{P_{T|X}(0|x)P_{Y|TX}(0|0,x)}.
\end{aligned}$$

Similarly, the right-hand side of equation (A.1) is

$$TYb_{TY}(X) - Tb_T(X) - Yb_Y(X) + b_o(X),$$

where

$$\begin{aligned}
b_{TY}(x) &= \frac{1}{P_{Y|X}(1|X)P_{T|YX}(1|1,x)P_{T|YX}(0|1,x)} + \frac{1}{P_{Y|X}(0|x)P_{T|YX}(1|0,x)P_{T|YX}(0|0,x)},\\
b_Y(x) &= \frac{1}{P_{Y|X}(1|x)P_{T|YX}(0|1,x)} + \frac{1}{P_{Y|X}(0|x)P_{T|YX}(0|0,x)},\\
b_T(x) &= \frac{1}{P_{Y|X}(0|x)P_{T|YX}(1|0,x)P_{T|YX}(0|0,x)},\\
b_o(x) &= \frac{1}{P_{T|X}(0|x)P_{T|YX}(0|0,x)}.
\end{aligned}$$

Here, $a_o(x) = b_o(x)$ by the Bayes rule. Also,

$$\begin{aligned}
a_Y(x) = \frac{P_{T|X}(0|x)}{P_{YT|X}(1,0|x)P_{YT|X}(0,0|x)} &= \frac{P_{YT|X}(0,0|x) + P_{YT|X}(1,0|x)}{P_{YT|X}(1,0|x)P_{YT|X}(0,0|x)}\\
&= \frac{1}{P_{YT|X}(1,0|x)} + \frac{1}{P_{YT|X}(0,0|x)} = b_Y(X).
\end{aligned}$$

Similarly, $a_T(x) = b_T(x)$ and $a_{TY}(x) = b_{TY}(x)$ follows from simple algebra. □

Therefore, the proof of Theorem 3.1 will be complete if we show that $F_r(Y,T,X)$ is the efficient influence function of θ_0. Before we do this, we prove several lemmas. Let $P_y(X) = \mathbb{P}(T = 1 \mid Y = y, X)$ and let $\tilde{\gamma} = (p, \gamma)^\intercal$ be the parameter that denotes smooth regular parametric submodels, where γ parametrizes the conditional likelihood

$$\mathcal{L}_y(T,X) = f_{X|Y}(X \mid y)P_y(X)^T\{1 - P_y(X)\}^{1-T}$$

and p is the parameter whose true value is $p_0 = \mathbb{P}(Y = 1)$. So, regular parametric submodels will be denoted by using $f_{X|Y}(x \mid y;\gamma)$ and $P_y(x;\gamma)$, along with the parameter γ. The truth will be denoted by $\tilde{\gamma}_0 = (p_0, \gamma_0)^\intercal$. We will use the symbol $\partial_\gamma g(\gamma_0)$ to denote the derivative of the function g with respect to γ evaluated at γ_0.

For $y = \{0, 1\}$, define

$$S_{X|Y}(X \mid y) := \partial_\gamma \log f_{X|Y}(X \mid y;\gamma_0) \quad \text{and} \quad A_y(X) := \frac{\partial_\gamma P_y(X;\gamma_0)}{P_y(X)\{1 - P_y(X)\}}.$$

Note that $S_{X|Y}(X \mid y)$ is restricted only by $\mathbb{E}\{S_{X|Y}(X \mid y) \mid Y = y\} = 0$, while the derivatives $\partial_\gamma P_y(X;\gamma_0)$ are unrestricted.

Lemma A.1. *The tangent space is given by the set of functions of the form*

$$s(Y,T,X) = (1 - Y)\big[\tilde{a}_0(X) + \{T - P_0(X)\}\tilde{b}_0(X)\big] + Y\big[\tilde{a}_1(X) + \{T - P_1(X)\}\tilde{b}_1(X)\big] + \kappa(Y - p_0),$$

where κ is a constant, and the functions $\tilde{a}_y$ and $\tilde{b}_y$ for $y = 0, 1$ are such that $\mathbb{E}\{\tilde{a}_y(X) \mid Y = y\} = 0$ and $\mathbb{E}\{s^2(Y,T,X)\} < \infty$.

Proof. The score along the regular parametric submodels at $\tilde{\gamma}_0$ can be expressed as $S(Y,T,X) = \big(S_p(Y,T,X)\ S_\gamma(Y,T,X)\big)^\intercal$, where

$$S_p(Y,T,X) = \frac{Y - p_0}{p_0(1 - p_0)}, \tag{A.2}$$

$$S_\gamma(Y,T,X) = YS_{\gamma,1}(T,X) + (1 - Y)S_{\gamma,0}(T,X), \tag{A.3}$$

where $S_{\gamma,y}(T,X) = S_{X|Y}(X|y) + \{T - P_y(X)\}A_y(X)$ for $y = 0, 1$. Noting that $S_{X|Y}(X \mid y)$ is only restricted by the conditional mean zero condition and $\partial_\gamma P_y(X;\gamma_0)$ is unrestricted, the lemma follows from linear combinations of $S_p(Y,T,X)$ and $S_\gamma(Y,T,X)$. □

Consider $\theta_0(\tilde{\gamma})$ defined by

$$\theta_0(\tilde{\gamma}) := \theta_Y(1;\gamma)p + \theta_Y(0;\gamma)(1 - p),$$

where

$$\theta_Y(y;\gamma) = \int_{\mathcal{X}} \log \mathrm{OR}_\mathrm{r}(x;\gamma)f_{X|Y}(x|y;\gamma)dx.$$

Lemma A.2. *The derivatives of $\theta_0(\cdot)$ at $\tilde{\gamma}_0$ are given by*

$$\begin{cases} \partial_p\theta_0(\tilde{\gamma}_0) = \theta_Y(1) - \theta_Y(0), \\ \partial_\gamma\theta_0(\tilde{\gamma}_0) = \partial_\gamma\theta_Y(1;\gamma_0)p_0 + \partial_\gamma\theta_Y(0;\gamma_0)(1 - p_0), \end{cases}$$

where

$$\partial_\gamma\theta_Y(y;\gamma_0) = \int_{\mathcal{X}} \{A_1(x) - A_0(x) + \log \mathrm{OR}_\mathrm{r}(x)S_{X|Y}(x \mid y)\}f_{X|Y}(x \mid y)dx \tag{A.4}$$

Proof. It follows by direct calculation by using

$$\partial_\gamma \mathrm{OR}_\mathrm{r}(x;\gamma_0) = \partial_\gamma P_1(x;\gamma_0)\frac{\{1 - P_0(x)\}}{P_0(x)\{1 - P_1(x)\}^2} - \partial_\gamma P_0(x;\gamma_0)\frac{P_1(x)}{P_0^2(x)\{1 - P_1(x)\}} = \{A_1(x) - A_0(x)\}\mathrm{OR}(x). \quad \square$$

Proof of the influence function result in Theorem 3.1:

We are now ready to complete the proof of Theorem 3.1; we will show that $F_r(Y,T,X)$ is the efficient influence function of θ_0. First, note that

$$F_r(Y,T,X) = (Y-p_0)\{\theta_Y(1)-\theta_Y(0)\} + YF_{r1}(T,X) + (1-Y)F_{r0}(T,X),$$

where

$$\begin{aligned} F_{r1}(T,X) &:= \log \mathrm{OR_r}(X) - \theta_Y(1) + \frac{1}{Q(X)}\frac{T-P_1(X)}{P_1(X)\{1-P_1(X)\}}, \\ F_{r0}(T,X) &:= \log \mathrm{OR_r}(X) - \theta_Y(0) - \frac{1}{1-Q(X)}\frac{T-P_0(X)}{P_0(X)\{1-P_0(X)\}}, \end{aligned}$$

where $Q(X) := \mathbb{P}(Y=1|X)$. Therefore, by Lemma A.1, F_r is clearly in the tangent space, and hence it suffices by Lemma A.2 to show that

$$\mathbb{E}\{F_r(Y,T,X)S_p(Y,T,X)\} = \theta_Y(1)-\theta_Y(0), \tag{A.5}$$

and

$$\begin{aligned} &\mathbb{E}\{F_r(Y,T,X)S_\gamma(Y,T,X)\} \\ &\qquad = \mathbb{E}\{A_1(X)-A_0(X)\} + p_0\int_{\mathcal{X}} \log \mathrm{OR_r}(x) S_{X|Y}(x\mid 1) f_{X|Y}(x\mid 1)dx \\ &\qquad\qquad + (1-p_0)\int_{\mathcal{X}} \log \mathrm{OR_r}(x) S_{X|Y}(x\mid 0) f_{X|Y}(x\mid 0)dx, \end{aligned} \tag{A.6}$$

where S_p and $S_\gamma = YS_{\gamma,1} + (1-Y)S_{\gamma,0}$ are provided in equations (A.2) and (A.3).

First, by using the fact that $(Y-p_0)Y = (1-p_0)Y$ and $(Y-p_0)(1-Y) = -p_0(1-Y)$, we obtain

$$F_r(Y,T,X)S_p(Y,T,X) = \frac{(Y-p_0)^2}{p_0(1-p_0)}\{\theta_Y(1)-\theta_Y(0)\} + \frac{Y}{p_0}F_{r1}(T,X) - \frac{1-Y}{1-p_0}F_{r0}(T,X),$$

of which the expectation shows that equation (A.5) is satisfied.

Now, consider equation (A.6). Note that

$$\begin{aligned} F_r(Y,T,X)S_\gamma(Y,T,X) &= S_\gamma(Y,T,X)(Y-p_0)\{\theta_Y(1)-\theta_Y(0)\} \\ &\qquad + YF_{r1}(T,X)S_{\gamma,1}(T,X) + (1-Y)F_{r0}(T,X)S_{\gamma,0}(T,X), \end{aligned} \tag{A.7}$$

where the last two terms use the fact that $Y(1-Y)=0$. Here, by using the fact that $(Y-p_0)Y = (1-p_0)Y$ and $(Y-p_0)(1-Y) = -p_0(1-Y)$ again, we obtain

$$\mathbb{E}\{(Y-p_0)S_\gamma(Y,T,X)\} = (1-p_0)\mathbb{E}\{YS_{\gamma,1}(T,X)\} - p_0\mathbb{E}\{(1-Y)S_{\gamma,0}(T,X)\} = 0, \tag{A.8}$$

where the last equality follows from $\mathbb{E}\{S_{\gamma,y}(T,X)\mid Y=y\}=0$. So, the first term on the right-hand side of equation (A.7) has been taken care of.

For the second term on the right-hand side of equation (A.7), note that

$$\begin{aligned} &\mathbb{E}\{YF_{r1}(T,X)S_{\gamma,1}(T,X)\} \\ &= p_0\mathbb{E}\Big[\Big\{\log \mathrm{OR_r}(X) + \frac{1}{Q(X)}\frac{T-P_1(X)}{P_1(X)\{1-P_1(X)\}}\Big\}S_{\gamma,1}(T,X)\ \Big|\ Y=1\Big] \\ &= p_0\mathbb{E}\Big[\log \mathrm{OR_r}(X)S_{X|Y}(X\mid 1) + \frac{1}{Q(X)}\frac{\{T-P_1(X)\}^2}{P_1(X)\{1-P_1(X)\}}A_1(X)\ \Big|\ Y=1\Big] \\ &= p_0\mathbb{E}\big\{\log \mathrm{OR_r}(X)S_{X|Y}(X\mid 1)\ \big|\ Y=1\big\} + \mathbb{E}\Big\{\frac{YA_1(X)}{Q(X)}\Big\} \\ &= p_0\mathbb{E}\big\{\log \mathrm{OR_r}(X)S_{X|Y}(X\mid 1)\ \big|\ Y=1\big\} + \mathbb{E}\big\{A_1(X)\big\}, \end{aligned} \tag{A.9}$$

where the last equality follows from the fact that $\mathbb{E}(Y|X) = Q(X)$. Similarly, the expectation of the third term on the right-hand side of (A.7) is

$$\mathbb{E}\{(1-Y)F_{r0}(T,X)S_{\gamma,0}(T,X)\} = (1-p_0)\mathbb{E}\big\{\log \mathrm{OR}_{\mathrm{r}}(X)S_{X|Y}(X \mid 0) \bigm| Y=0\big\} - \mathbb{E}\big\{A_0(X)\big\}. \tag{A.10}$$

Combining equation (A.7) with (A.8) to (A.10) verifies equation (A.6). So, we are done. □

Proof of Theorem 3.2: In the proof, we focus on the prospective estimating equation. The case of retrospective estimating equation is similar. For simplicity, we suppress subscript p in the notation. Recall that for $\eta = (a,b,c)^\intercal \in \mathcal{G}^3$, and

$$\tilde{F}(\eta)[Y,T,X] = \log\left[\frac{b(X)\{1-a(X)\}}{\{1-b(X)\}a(X)}\right] - \theta_0 + \frac{T}{c(X)}\frac{\{Y-b(X)\}}{b(X)\{1-b(X)\}} - \frac{1-T}{1-c(X)}\frac{\{Y-a(X)\}}{a(X)\{1-a(X)\}}$$

and

$$\eta_0(x) = \big(\mathbb{P}(Y=1|T=0,X=x), \mathbb{P}(Y=1|T=1,X=x), \mathbb{P}(T=1 \mid X)\big)^\intercal.$$

Define

$$\mathrm{OR}(\eta)[X] := \frac{b(X)\{1-a(X)\}}{\{1-b(X)\}a(X)}. \tag{A.11}$$

Now, as in the proof of lemma A.2, we have

$$\partial_\gamma \log \mathrm{OR}\{\eta_0 + \gamma(\eta-\eta_0)\}[X]\big|_{\gamma=0} = \frac{b(X)-b_0(X)}{b_0(x)\{1-b_0(x)\}} - \frac{a(X)-a_0(X)}{a_0(x)\{1-a_0(x)\}}, \tag{A.12}$$

where $a_0(X) := \mathbb{P}(Y=1|T=0,X)$ and $b_0(X) := \mathbb{P}(Y=1|T=1,X)$.

Define

$$\Delta_0(\eta)[Y,T,X] = -\frac{1-T}{1-c(X)}\frac{\{Y-a(X)\}}{a(X)\{1-a(X)\}},$$
$$\Delta_1(\eta)[Y,T,X] = \frac{T}{c(X)}\frac{\{Y-b(X)\}}{b(X)\{1-b(X)\}}.$$

Then, we have that

$$\mathbb{E}\Big(\partial_\gamma \Delta_0\{\eta_0+\gamma(\eta-\eta_0)\}[Y,T,X]\big|_{\gamma=0} \Bigm| X\Big) = \frac{a(X)-a_0(X)}{a_0(X)\{1-a_0(X)\}}, \tag{A.13}$$

$$\mathbb{E}\Big(\partial_\gamma \Delta_1\{\eta_0+\gamma(\eta-\eta_0)\}[Y,T,X]\big|_{\gamma=0} \Bigm| X\Big) = -\frac{b(X)-b_0(X)}{b_0(X)\{1-b_0(X)\}}. \tag{A.14}$$

Therefore, the conclusion follows from (A.12) to (A.14). □

Proof of Theorem 4.1: As in the previous proof, we focus on we focus on the prospective estimating equation. The case of retrospective estimating equation is similar. As before, we suppress subscript p in the notation.

We verify Assumptions 3.1 and 3.2 of Chernozhukov et al. (2018, C-DML hereafter). Following the notation used in C-DML, we have $\psi(W;\theta,\eta) = \tilde{F}_r(\eta)[Y,T,X]$ with $W = (Y,T,X)$: so, our case belongs to that of linear scores, namely

$$\psi(W;\theta,\eta) = \psi^a(W;\eta)\theta + \psi^b(W;\eta),$$

where

$$\psi^a(W;\eta) = -1,$$
$$\psi^b(W;\eta) = \log \mathrm{OR}(\eta)[X] + \frac{T}{c(X)}\frac{Y-b(X)}{b(X)\{1-b(X)\}} - \frac{1-T}{1-c(X)}\frac{Y-a(X)}{a(X)\{1-a(X)\}}$$

and $\mathrm{OR}(\eta)[X]$ is defined in (A.11).

Verification of Assumption 3.1 of C-DML. Under Assumption 3.1, Specifically, Assumption 3.1 (a) of C-DML is satisfied by (3.2); part (b) is by the linearity of the score ψ; part (c) is by Assumption 3.1; part (d) is by Theorem 3.2; part (e) follows because $\mathbb{E}[\psi^a(W;\eta_0)] = -1$.

Verification of Assumption 3.2 (b) of C-DML. It holds trivially that $|\psi^a(W;\eta)|$ is bounded by a constant uniformly in η. Moreover, by Assumption 3.1, there is a constant $c_1 < \infty$ such that

$$|\psi(W;\theta,\eta)| \leq c_1$$

uniformly in η almost surely.

Verification of Assumption 3.2 (d) of C-DML. Let $b_0(X) = \mathbb{P}(Y=1|X,T=1)$ and $c_0 = \mathbb{P}(T=1|X)$. Note that

$$\mathbb{E}[\psi^2(W;\theta,\eta_0)] \geq \mathbb{E}\left[\{\log \mathrm{OR}(X) - \theta_0\}^2 + \frac{1}{c_0(X)}\frac{1}{b_0(X)\{1-b_0(X)\}}\right],$$

which is bounded from below by a constant under Assumption 3.1.

Since Assumption 3.2 (a) of C-DML is the definition of the first stage estimator, Theorem 4.1 follows immediately from Theorems 3.1 and 3.2 of C-DML, provided that we verify the remaining Assumption 3.2 (c) of C-DML.

Verification of Assumption 3.2 (c) of C-DML. Using the notation used in C-DML, define

$$\begin{aligned} r_n &:= \sup_{\eta\in\mathcal{T}_N} |\mathbb{E}[\psi^a(W;\eta) - \psi^a(W;\eta_0)]|, \\ r_n' &:= \sup_{\eta\in\mathcal{T}_N} (\mathbb{E}[|\psi(W;\theta,\eta) - \psi(W;\theta,\eta_0)|^2])^{1/2}, \\ \lambda_n' &:= \sup_{\gamma\in(0,1),\eta\in\mathcal{T}_N} |\partial_\gamma^2 \mathbb{E}[\psi(W;\theta,\eta_0+\gamma(\eta-\eta_0))]|, \end{aligned}$$

where $\mathcal{T}_N \subseteq \mathcal{G}^3$ is a nuisance realization set that is discussed in detail in C-DML.

Step 1. Note that $r_n = 0$ since $\psi^a(W;\eta)$ does not depend on η.

Step 2. Now write that

$$\Big[\mathbb{E}\big\{|\psi(W;\theta,\eta) - \psi(W;\theta,\eta_0)|^2\big\}\Big]^{1/2} = \|\psi(W;\theta,\eta) - \psi(W;\theta,\eta_0)\|_{P,2} \leq \|\mathcal{B}_1\|_{P,2} + \|\mathcal{B}_2\|_{P,2} + \|\mathcal{B}_3\|_{P,2},$$

where

$$\begin{aligned} \mathcal{B}_1 &:= \log \mathrm{OR}(\eta)[X] - \log \mathrm{OR}(X), \\ \mathcal{B}_2 &:= \frac{T}{c(X)}\frac{Y-b(X)}{b(X)\{1-b(X)\}} - \frac{T}{c_0(X)}\frac{Y-b_0(X)}{b_0(X)\{1-b_0(X)\}}, \\ \mathcal{B}_3 &:= \frac{1-T}{1-c(X)}\frac{Y-a(X)}{a(X)\{1-a(X)\}} - \frac{1-T}{1-c_0(X)}\frac{Y-a_0(X)}{a(X)\{1-a_0(X)\}}. \end{aligned}$$

Then, in view of Assumptions 3.1 and 4.1, there exists a sequence $\tilde{\delta}_n \to 0$ such that

$$\Big[\mathbb{E}\big\{|\psi(W;\theta,\eta) - \psi(W;\theta,\eta_0)|^2\big\}\Big]^{1/2} \leq \tilde{\delta}_n$$

holds with probability at least $1-\tau_n$. This implies that we can take $r_n' = \tilde{\delta}_n$.

Step 3. Define $a_\gamma(X) := a_0(X) + \gamma\{a(X) - a_0(X)\}$, $b_\gamma(X) := b_0(X) + \gamma\{b(X) - b_0(X)\}$ and $c_\gamma(X) := c_0(X) + \gamma\{c(X) - c_0(X)\}$. Note that

$$\partial_\gamma \log \mathrm{OR}\{\eta_0 + \gamma(\eta-\eta_0)\}[X] = \frac{b(X)-b_0(X)}{b_\gamma(X)\{1-b_\gamma(X)\}} - \frac{a(X)-a_0(X)}{a_\gamma(X)\{1-a_\gamma(X)\}}.$$

In addition,

$$\begin{aligned}
\partial_\gamma\left[\frac{\{Y-a_\gamma(X)\}}{a_\gamma(X)\{1-a_\gamma(X)\}}\right] &= -\frac{a(X)-a_0(X)}{a_\gamma(X)\{1-a_\gamma(X)\}} - \frac{\{Y-a_\gamma(X)\}\{1-2a_\gamma(X)\}}{a_\gamma^2(X)\{1-a_\gamma(X)\}^2}\{a(X)-a_0(X)\},\\
\partial_\gamma\left[\frac{Y-b_\gamma(X)}{b_\gamma(X)\{1-b_\gamma(X)\}}\right] &= -\frac{b(X)-b_0(X)}{b_\gamma(X)\{1-b_\gamma(X)\}} - \frac{\{Y-b_\gamma(X)\}\{1-2b_\gamma(X)\}}{b_\gamma^2(X)\{1-b_\gamma(X)\}^2}\{b(X)-b_0(X)\},\\
\partial_\gamma\left[\frac{T}{c_\gamma(X)}\right] &= -\frac{T\{c(X)-c_0(X)\}}{\{c_\gamma(X)\}^2},\\
\partial_\gamma\left[\frac{1-T}{1-c_\gamma(X)}\right] &= \frac{(1-T)\{c(X)-c_0(X)\}}{\{1-c_\gamma(X)\}^2}.
\end{aligned}$$

Combining these yields

$$\begin{aligned}
&\partial_\gamma\psi(W;\theta,\eta_0+\gamma(\eta-\eta_0))\\
&\qquad= \frac{b(X)-b_0(X)}{b_\gamma(X)\{1-b_\gamma(X)\}} - \frac{a(X)-a_0(X)}{a_\gamma(X)\{1-a_\gamma(X)\}}\\
&\qquad- \frac{T}{c_\gamma(X)}\left[\frac{b(X)-b_0(X)}{b_\gamma(X)\{1-b_\gamma(X)\}} + \frac{\{Y-b_\gamma(X)\}\{1-2b_\gamma(X)\}}{b_\gamma^2(X)\{1-b_\gamma(X)\}^2}\{b(X)-b_0(X)\}\right]\\
&\qquad+ \frac{1-T}{1-c_\gamma(X)}\left[\frac{a(X)-a_0(X)}{a_\gamma(X)\{1-a_\gamma(X)\}} + \frac{\{Y-a_\gamma(X)\}\{1-2a_\gamma(X)\}}{a_\gamma^2(X)\{1-a_\gamma(X)\}^2}\{a(X)-a_0(X)\}\right]\\
&\qquad- \frac{T\{c(X)-c_0(X)\}}{\{c_\gamma(X)\}^2}\left[\frac{Y-b_\gamma(X)}{b_\gamma(X)\{1-b_\gamma(X)\}}\right] - \frac{(1-T)\{c(X)-c_0(X)\}}{\{1-c_\gamma(X)\}^2}\left[\frac{\{Y-a_\gamma(X)\}}{a_\gamma(X)\{1-a_\gamma(X)\}}\right].
\end{aligned}$$

If we take the second-order derivative in the equation above, we can see that each term of the second-order derivatives on the right-hand side can be bounded in absolute value by

$$\chi(X) := C_1\{a(X)-a_0(X)\}^2 + C_2\{b(X)-b_0(X)\}^2 + C_3\{c(X)-c_0(X)\}^2$$

for some appropriate constants C_1, C_2, and C_3, because η is in $\mathcal{G}^3$ and $|f(X)g(X)| \le \{f^2(X)+g^2(X)\}/2$ for any real-valued functions f and g. Therefore, by averaging X out, we obtain

$$\left|\partial_\gamma^2\mathbb{E}[\psi(W;\theta,\eta_0+\gamma(\eta-\eta_0))]\right| \le \mathbb{E}\{\chi(X)\}.$$

Then, by Assumption 4.1, there exists a sequence $\tilde{\delta}_n' \to 0$ such that

$$\sup_{\gamma\in(0,1),\eta\in\mathcal{T}_N}\left|\partial_\gamma^2\mathbb{E}[\psi(W;\theta,\eta_0+\gamma(\eta-\eta_0))]\right| \le \tilde{\delta}_n' n^{-1/2}$$

holds with probability at least $1-\tau_n$. Therefore, we can take $\lambda_n' = \tilde{\delta}_n' n^{-1/2}$. □

Proof of Theorem 5.1: We focus on the case, where m is evaluated at $\varphi_{p0}, \vartheta_0$, and an arbitrary point φ_r; the other case is similar. Recall that

$$\mathbb{P}(Y=1\mid T=0,X) = \frac{\exp\{\varphi_{p0}(X)\}}{1+\exp\{\varphi_{p0}(X)\}}, \tag{A.15}$$

$$\mathbb{P}(Y=0\mid T=0,X) = \frac{1}{1+\exp\{\varphi_{p0}(X)\}} \tag{A.16}$$

by the definition of φ_{p0}. Therefore,

$$\begin{aligned}
\mathbb{P}(Y=1\mid T,X) &= \frac{\exp\{\varphi_{p0}(X)\}\exp\{\vartheta_0(X)T\}}{1+\exp\{\varphi_{p0}(X)\}\exp\{\vartheta_0(X)T\}} \qquad\qquad (A.17)\\
&= \frac{\mathbb{P}(Y=1\mid T=0,X)\exp\{\vartheta_0(X)T\}}{\mathbb{P}(Y=0\mid T=0,X)+\mathbb{P}(Y=1\mid T=0,X)\exp\{\vartheta_0(X)T\}},
\end{aligned}$$

where the second equality follows by dividing the numerator and the denominator by $1+\exp\{\varphi_{p0}(X)\}$. Also,

$$\begin{aligned}\mathbb{P}(Y=0\mid T,X) &= 1-\mathbb{P}(Y=1\mid T,X) \\ &= \frac{\mathbb{P}(Y=0\mid T=0,X)}{\mathbb{P}(Y=0\mid T=0,X)+\mathbb{P}(Y=1\mid T=0,X)\exp\{\vartheta_0(X)T\}}.\end{aligned} \tag{A.18}$$

So, we can combine equations (A.17) and (A.18) by

$$\mathbb{P}(Y=y\mid T,X)=\mathcal{D}^{-1}\mathbb{P}(Y=y\mid T=0,X)\exp\{\vartheta_0(X)Ty\}, \tag{A.19}$$

where

$$\mathcal{D}:=\mathbb{P}(Y=0\mid T=0,X)+\mathbb{P}(Y=1\mid T=0,X)\exp\{\vartheta_0(X)T\}.$$

Now,

$$\mathbb{E}\{m(\varphi_{p0},\varphi_r,\vartheta_0,Y,T,X)\mid T,X\}=\big\{T-\Lambda_0(\varphi_r,X)\big\}\mathbb{E}\big[\big\{Y-\Lambda_0(\varphi_{p0},X)\big\}\exp\{-\vartheta_0(X)TY\}\mid T,X\big],$$

where the expectation factor on the right-hand side is equal to

$$\begin{aligned}&\sum_{y=0}^{1}\{y-\Lambda_0(\varphi_{p0},X)\}\exp\{-\vartheta_0(X)Ty\}\mathbb{P}(Y=y\mid T,X)\\ &\qquad=\mathcal{D}^{-1}\sum_{y=0}^{1}\{y-\Lambda_0(\varphi_{p0},X)\}\mathbb{P}(Y=y\mid T=0,X)=\mathcal{D}^{-1}\big[\mathbb{E}\{Y\mid T=0,X\}-\Lambda_0(\varphi_{p0},X)\big]=0,\end{aligned}$$

where the first equality is by equation (A.19). Therefore, the first assertion has been shown. For the second assertion, note that

$$\mathbb{E}\{m(\varphi_{p0},\varphi_r,\vartheta,Y,T,X)\mid T=0,X\}=-\Lambda_0(\varphi_r,X)\mathbb{E}\big\{Y-\Lambda_0(\varphi_{p0},X)\mid T=0,X\big\}=0.$$

Further,

$$\mathbb{E}\{m(\varphi_{p0},\varphi_r,\vartheta,Y,T,X)\mid T=1,X\}=\big\{1-\Lambda_0(\varphi_r,X)\big\}\mathbb{E}\big[\{Y-\Lambda_0(\varphi_{p0},X)\}\exp\{-\vartheta(X)Y\}\mid T=1,X\big],$$

where the expectation factor on the right-hand side is equal to

$$\begin{aligned}&\sum_{y=0}^{1}\{y-\Lambda_0(\varphi_{p0},X)\}\exp\{-\vartheta(X)y\}\mathbb{P}(Y=y\mid T=1,X)\\ &\quad=\mathcal{D}^{-1}\sum_{y=0}^{1}\{y-\Lambda_0(\varphi_{p0},X)\}\exp\big[y\{\vartheta_0(X)-\vartheta(X)\}\big]\mathbb{P}(Y=y\mid T=0,X)\\ &\quad=\frac{-\Lambda_0(\varphi_{p0},X)+\{1-\Lambda_0(\varphi_{p0},X)\}\exp\{\varphi_{p0}(X)\}\exp\{\vartheta_0(X)-\vartheta(X)\}}{\mathcal{D}[1+\exp\{\varphi_{p0}(X)\}]}\\ &\quad=\frac{\big[\exp\{\vartheta_0(X)-\vartheta(X)\}-1\big]\exp\{\varphi_{p0}(X)\}}{\mathcal{D}[1+\exp\{\varphi_{p0}(X)\}]^2},\end{aligned}$$

where the first equality is by equation (A.19). Then, we note that this is equal to zero if and only if $\vartheta(X)=\vartheta_0(X)$. □