Working paper. (2023) RESEARCH NOTE

Structure-agnostic Optimality of Doubly Robust Learning for Treatment Effect Estimation

Jikai Jin*,1 and Vasilis Syrgkanis*,2

[1]Institute for Computational & Mathematical Engineering, Stanford University, CA, USA
[2]Department of Management Science & Engineering, Stanford University, CA, USA
*Corresponding author. Email: jkjin@stanford.edu; vsyrgk@stanford.edu

Abstract

Average treatment effect estimation is the most central problem in causal inference with application to numerous disciplines. While many estimation strategies have been proposed in the literature, recently also incorporating generic machine learning estimators, the statistical optimality of these methods has still remained an open area of investigation. In this paper, we adopt the recently introduced structure-agnostic framework of statistical lower bounds, which poses no structural properties on the nuisance functions other than access to black-box estimators that attain small errors; which is particularly appealing when one is only willing to consider estimation strategies that use non-parametric regression and classification oracles as a black-box sub-process. Within this framework, we prove the statistical optimality of the celebrated and widely used doubly robust estimators for both the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATTE), as well as weighted variants of the former, which arise in policy evaluation.

Keywords: Causal inference, machine learning, semi-parametric estimation

1. Introduction

Estimating the average treatment effect is one of the central problems in causal inference and has found important applications in numerous disciplines such as economics (Hirano, Imbens, and Ridder 2003; Imbens 2004), education (Oreopoulos 2006), epidemiology (Little and Rubin 2000; Wood et al. 2008) and political science (Mayer 2011). In view of its practical importance, a large body of work is devoted to developing statistically efficient estimators for the average treatment effect based on regression (Robins, Rotnitzky, and Zhao 1994, 1995; Imbens, Newey, and Ridder 2003), matching (Heckman, Ichimura, and Todd 1998; Rosenbaum 1989; Abadie and Imbens 2006) and propensity scores (Rosenbaum and Rubin 1983; Hirano, Imbens, and Ridder 2003) as well as their combinations. In this paper, we consider estimation of average treatment effects under the assumption that all potential confounders X between a binary treatment D and an outcome Y are observed, albeit potentially of large dimensionality; a setting that has considered substantial attention in the recent literature at the intersection of causal inference and machine learning.

Under this assumption the statistical estimation problem can be formalized as follows. We observe data (X, D, Y) that follow a distribution that satisfies the following non-linear regression equations:

$$\begin{aligned} Y &= g_0(D, X) + U \\ D &= m_0(X) + V \end{aligned} \tag{1}$$

where D is a binary treatment variable, U, V are noise variables that satisfy $\mathbb{E}[U \mid D, X] = \mathbb{E}[V \mid X] = 0$. The conditional mean $g_0(d, x)$ and the propensity score $m_0(x)$ are commonly referred to as *nuisance functions*.

We are interested in the problem of estimating the weighted average treatment effect (WATE) and the average treatment effect of the treated (ATTE). The WATE is defined as

$$\theta^{ATE} = \mathbb{E}\left[w(X)\left(g_0(1, X) - g_0(0, X)\right)\right] \tag{2}$$

where $w(x) \geqslant 0$ is a pre-specified and known bounded weight function defined on $\mathrm{supp}(X)$. In the special case when w is constant, Equation (2) recovers the standard definition of the Average Treatment Effect (ATE). WATE with different choices of the weight function is often used to measure the effect of personalized interventions on a specific targeted population by some pre-specified personalized policy (Tao and Fu 2019; Hirano, Imbens, and Ridder 2003). The ATTE is defined as

$$\theta^{ATTE} = \mathbb{E}\left[g(1, X) - g(0, X) \mid D = 1\right] \tag{3}$$

and measures the effect of treatments on the treated population (Heckman, Ichimura, and Todd 1998).

Since the nuisance functions g_0 and m_0 in Equation (1) are unknown and may have complex structures, and since the dimension K of the covariates X can be large relative to the number of data n in many applications, it is extremely suitable to apply modern machine learning (ML) methods, including Lasso (Belloni, Chernozhukov, and Wang 2014; Geer et al. 2014; Chernozhukov, Newey, and Singh 2022), random forest (Wager and Walther 2015; Syrgkanis and Zampetakis 2020), elastic net (Zou and Hastie 2005) and deep learning (neural networks) (Schmidt-Hieber 2020; Farrell, Liang, and Misra 2021) for nonparametric estimation of nuisance functions. Then, a straightforward approach for estimating our target quantities is to directly plug in the ML estimators that we obtain. Concretely, let $\hat{g}(d, x)$ and $\hat{m}(x)$ be our ML estimators for $g_0(d, x)$ and $m_0(x)$ respectively, then one can estimate θ^{WATE} with

$$\hat{\theta}_{PI}^{WATE} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)\left(\hat{g}(1, X_i) - \hat{g}(0, X_i)\right), \tag{4}$$

and θ^{ATTE} with

$$\hat{\theta}_{PI}^{ATTE} = \left(\sum_{i=1}^{n} \hat{m}(X_i)\right)^{-1} \sum_{i=1}^{n} \left(\hat{g}(1, X_i) - \hat{g}(0, X_i)\right)\hat{m}(X_i). \tag{5}$$

However, in order to avoid overfitting, ML methods typically uses various forms of regularization that may lead to prohibitively large bias in the plug-in estimators. To mitigate this issue, a line of works (Chernozhukov et al. 2017; Chernozhukov et al. 2018; Foster and Syrgkanis 2023; Rotnitzky, Smucler, and Robins 2021; Chernozhukov et al. 2022; Chernozhukov, Newey, and Singh 2023) proposes to employ a two-stage estimation process called *double/debiased machine learning (DML)*, that first obtains an ML estimator using a portion of data, and then use the remaining data to debias this estimator based on the doubly robust estimating equations (Robins and Rotnitzky 1995). Formally, suppose that θ_0 is the parameter of interest, η_0 is a nuisance function, and P is a data distribution such that the *moment condition*

$$M\left(\theta_0, \eta_0\right) := \mathbb{E}_{W\sim P}\left[m\left(W; \theta_0, \eta_0\right)\right] = 0$$

holds for some *moment function* m. After obtaining an ML estimator $\hat{\eta}$, DML constructs an estimator $\hat{\theta}$ of θ_0 by solving the following moment equation:

$$M_n(\theta, \hat{\eta}) := \frac{1}{n}\sum_{i=1}^{n} \psi(W_i; \theta, \hat{\eta}) = 0.$$

Then the following result is known for the DML estimator:

Theorem 1. *(informal version of (Chernozhukov et al. 2018), Theorem 3.1*[1]*) Suppose that* $\psi(W;\theta,\eta)$ *is linear in* θ, i.e., $\psi(W;\theta,\eta) = a(W,\eta)\theta + v(W,\eta)$ *for some functions* a *and* v*. Assuming that*

- *the nuisance estimation is consistent:*

$$\|\hat{\eta} - \eta_0\|_{P,2} = o(1) \quad (n \to +\infty). \tag{6}$$

- *the Neyman orthogonality condition holds:*

$$D_\eta M(\theta_0, \eta_0)\,[\hat{\eta} - \eta_0] = 0 \tag{7}$$

- *second order directional derivative of the moment* M *in the direction of the nuisance error converges to zero faster than* $n^{-\frac{1}{2}}$*:*

$$\sqrt{n} D_{\eta\eta} M\,(\theta_0, \bar{\eta})\,[\hat{\eta} - \eta_0] = o_p(1), \quad \forall \bar{\eta} = \tau\hat{\eta} + (1-\tau)\eta_0, \tau \in [0,1], \tag{8}$$

and some additional regularity conditions, the DML estimator $\hat{\theta}$ *is asymptotically normal:* $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma})$ *for some covariance matrix* $\mathbf{\Sigma}$*.*

The asymptotic normality property stated in Theorem 1 implies a high-probability guarantee on the error rate:

Corollary 1. *In the same setup as Theorem 1, for any given* $\gamma > 0$*, there exists a constant* C_γ *such that* $P[|\hat{\theta} - \theta_0| \leqslant C_\gamma \sqrt{n}] \geqslant 1 - \gamma$ *for sufficiently large* n*.*

For our goal of estimating WATE and ATTE, one can derive explicit expression for the moment function ψ that satisfies the conditions in Theorem 1.

Example 1. *Let* $\theta_0 = \theta^{WATE}$ *be the target parameter,* $W = (X, D, Y)$ *is the data sampled from Equation (1),* $\eta_0 = (m_0, g_0)$ *and*

$$\psi(W;\theta,\eta) = w(X)\left[g_0(1,X) - g_0(0,X) + \left(\frac{D}{p_0(X)} - \frac{1-D}{1-p_0(X)}\right)(Y - g_0(D,X))\right] - \theta.$$

which is partially linear in θ *and satisfies Equation (7). Moreover, Equation (8) holds when*

$$\|m_0 - \hat{m}_0\|_{P,2} \cdot \|g_0 - \hat{g}_0\|_{P,2} = o\left(n^{-\frac{1}{2}}\right). \tag{9}$$

The resulting debiased estimator is

$$\hat{\theta}^{WATE} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)\left[\hat{g}(1,X_i) - \hat{g}(0,X_i) + \left(\frac{D_i}{\hat{m}(X_i)} - \frac{1-D_i}{1-\hat{m}(X_i)}\right)(Y_i - \hat{g}(D_i,X_i))\right]. \tag{10}$$

The condition (9) holds as long as the L^2 estimation errors of all nuisance functions are faster than $\mathcal{O}(n^{-\frac{1}{4}})$, which can be achieved by a broad range of machine learning methods (Bickel, Ritov, and Tsybakov 2009; Belloni and Chernozhukov 2011, 2013; Chen and White 1999; Wager and Athey 2018; Athey, Tibshirani, and Wager 2019). By Theorem 1, we can then deduce that the debiased estimator (10) is $\sqrt{n}$-consistent. In contrast, the plug-in estimator defined in Equation (4) is not $\sqrt{n}$-consistent unless it holds that $\|g_0 - \hat{g}_0\|_{P,2} = \mathcal{O}(n^{-\frac{1}{2}})$ (Chernozhukov et al. 2018). This is a strong requirement to impose; for example, it is shown in (Chernozhukov et al. 2022), Section 5.2 that it is violated by the Lasso estimator.

1. To be concise, here we only list the key assumptions underlying DML; we point the readers to (Chernozhukov et al. 2018) for a rigorous version of the theorem.

Example 2. *Let* $\theta_0 = \theta^{ATTE}$ *be the target parameter,* $W = (X, D, Y)$ *is the data sampled from Equation (1),* $\eta_0 = (m_0, g_0)$ *and*

$$\psi(W;\theta,\eta) = D\left(Y - g_0(0,X)\right) - \frac{m_0(X)}{1-m_0(X)}(1-D)\left(Y - g_0(0,X)\right) - D\theta.$$

which is partially linear in θ *and satisfies Equation (7). Moreover, Equation (8) holds when*

$$\|m_0 - \hat{m}_0\|_{P,2} \cdot \|g_0(0,\cdot) - \hat{g}_0(0,\cdot)\|_{P,2} = o\left(n^{-\frac{1}{2}}\right). \tag{11}$$

The resulting debiased estimator is

$$\hat{\theta}^{ATTE} = \left(\sum_{i=1}^{n} D_i\right)^{-1} \sum_{i=1}^{n} \left[D_i\left(Y_i - \hat{g}(0,X_i)\right) - \frac{\hat{m}(X_i)}{1-\hat{m}(X_i)}(1-D_i)(Y_i - \hat{g}(0,X_i))\right]. \tag{12}$$

The condition (11) holds as long as the L^2 estimation errors of the nuisance functions m_0 and $g_0(0,\cdot)$ are faster than $\mathcal{O}(n^{-\frac{1}{4}})$. In this case, Theorem 1 implies that the debiased estimator (12) is $\sqrt{n}$-consistent. Note that for ATTE, we do not need to estimate $g_0(1,\cdot)$ for constructing the DML estimator.

Given the theoretical benefits of DML as discussed above, one naturally asks whether the error rate guarantee of Theorem 1 can be further improved, especially in regimes where the nuisance function estimates do not converge at $n^{-1/4}$ rates. Indeed, for general nonparametric functional estimation, it has been shown decades ago that if the function possesses certain smoothness properties, then higher-order debiasing schemes can be designed that lead to improved error rates (Bickel and Ritov 1988; Birgé and Massart 1995). Specifically, first-order debiasing methods are suboptimal even when the nuisance function estimators are minimax optimal. Estimators based on higher-order debiasing have also been proposed and analyzed for functionals that arise in causal inference problems (Robins et al. 2008; Vaart 2014; Robins, Li, and Mukherjee 2017; Liu et al. 2017). However, the fast rates of these methods crucially rely on the structure of the underlying function classes.

Unlike first-order debiasing methods, higher-order methods are *not* structure-agnostic, in the sense that their error rates no longer apply to black-box estimators of the nuisance functions and the corresponding estimators are many times cumbersome to deploy in practice. This observation leads to the following question: *does there exist structure-agnostic estimators that can achieve better error rates than first-order debiasing methods?* This question led to the recent introduction of the *structure-agnostic minimax optimality framework* (Kennedy et al. 2022). However, existing structure-agnostic lower bounds do not apply to the central average treatment effect problem.

We give a negative answer to this question: *we show that double/debiased machine learning (DML) methods that rely on first-order debiasing are structure-agnostic optimal for estimating both WATE and ATTE.* More concretely, we prove information-theoretic lower bounds that match the upper bounds achieved by the doubly robust estimators in terms of the sample size and the quality of nuisance function estimators. Our proof of the lower bounds uses the method of fuzzy hypotheses that reduces our estimation problem to the problem of testing a pair of *mixtures* of hypotheses. Such methods are widely adopted in establishing lower bounds for non-parametric functional estimation problems (Tsybakov 2008; Robins et al. 2009; Kennedy et al. 2022; Balakrishnan, Kennedy, and Wasserman 2023). Our main technical contribution is a collection of carefully-designed constructions of the hypotheses that are built on *asymmetric perturbations* in the space of nuisance functions. We note that due to the more complicated relationships between the estimand and the data distribution, existing constructions of composite hypotheses (Robins et al. 2009; Kennedy et al. 2022) do not apply to our setting.

1.1 Related work

Typical debiasing methods, both first-order (Bickel 1982; Schick 1986) and higher order (Bickel and Ritov 1988; Birgé and Massart 1995) variants, often employs sample-splitting schemes that use the first half of the data to construct an initial estimator, and then performs an additional debiasing step using the second half of data. With recent advances in machine learning (ML), the first step is now often performed using ML estimators (Belloni et al. 2012; Belloni, Chernozhukov, and Kato 2015; Farrell 2015; Syrgkanis and Zampetakis 2020) that are especially suitable in the case where the covariate X is high-dimensional. However, these ML methods typically involve model selection/regularization to trade-off bias and variance, which may lead to large bias in model estimation (Belloni, Chernozhukov, and Hansen 2014). To resolve this issue, a line of works (Chernozhukov et al. 2017; Chernozhukov et al. 2018; Foster and Syrgkanis 2023; Rotnitzky, Smucler, and Robins 2021; Chernozhukov et al. 2022; Chernozhukov, Newey, and Singh 2023; Chen, Syrgkanis, and Austern 2022) propose double/debiased machine learning (DML) to debias *any* black-box ML estimators. On the other hand, when the nuisance parameters are assumed to have some level of smoothness, higher-order debiasing methods are proposed (Robins et al. 2008; Robins, Li, and Mukherjee 2017; Kennedy et al. 2022) and are shown to be minimax optimal for smooth classes of nuisance parameters, but these approaches require *ad hoc* design of both the estimator in the first step and the debiasing scheme in the second step, that heavily rely on the smoothness properties. In this sense, DML is structural-agnostic while higher-order debiasing methods are not.

To have a better theoretical understanding of the statistical benefits of DML, one then needs to somehow disentangle the effect of the structural assumptions and the debiasing techniques. The framework proposed by (Balakrishnan and Wasserman 2019) is precisely targeted towards this goal. It assumes that we already have black-box estimators of the nuisance functions, and we have n *i.i.d.* data from the ground-truth model. The goal is to characterize the best-achievable estimation error of the target quantity as a function of the sample size and the estimation error of the nuisance function. Our paper investigates the statistical limit of learning average treatment effect in the structural-agnostic framework proposed by (Balakrishnan, Kennedy, and Wasserman 2023) for functional estimation problems.

In (Balakrishnan, Kennedy, and Wasserman 2023), the authors investigate the estimation problem of three functionals: quadratic functionals in Gaussian sequence models, quadratic integral functionals and the expected conditional covariance $\theta^{Cov} = \mathbb{E}\left[\mathrm{Cov}(D, Y \mid X)\right]$ in Equation (1). The authors of (Balakrishnan, Kennedy, and Wasserman 2023) establish their lower bound by reducing it to the new problem of lower-bounding the error of a hypothesis testing problem. The error is then lower-bounded by constructing priors (mixtures) of the composite null and alternate distribution. The priors they construct are based on adding or subtracting "bumps" on top of a fixed hypothesis in a symmetric manner, which is a standard proof strategy for this type of problems (Ingster 1994; Robins et al. 2009; Arias-Castro, Pelletier, and Saligrama 2018; Balakrishnan and Wasserman 2019). The reason why the proof strategy of (Balakrishnan, Kennedy, and Wasserman 2023) fails for WATE and ATTE is that the functional relationships between the nuisance parameters and these target parameters are in different forms. Specifically, the target parameters that (Balakrishnan and Wasserman 2019) investigates are all in the form of

$$T(f, g) = \langle f, g \rangle_{\mathcal{H}}, \tag{13}$$

where f, g are unknown nuisance parameters that lie in some Hilbert space $\mathcal{H}$. To be concrete, consider the example of the expected conditional covariance θ^{Cov}. Let $\mu_0(x) = \mathbb{E}[Y \mid X = x]$, then we have that

$$\theta^{Cov} = \mathbb{E}[DY] - \int m_0(x)\mu_0(x)\mathrm{d}p_X(x)$$

where p_X is the marginal density of X. The first term, $\mathbb{E}[DY]$, can be estimated with a standard $\mathcal{O}(n^{-\frac{1}{2}})$ rate, so what remains to be estimated is the second term which is exactly in the form of Equation (13). However, this is *not* the case for WATE and ATTE, for which the estimand can be written as

$$\theta^{WATE} = T_1(m_0, g_0) := \mathbb{E}_X\left[w(X)\left(g_0(1,X) - g_0(0,X)\right)\right]$$

and

$$\theta^{ATTE} = T_2(m_0, g_0) := \frac{\mathbb{E}_X\left[\left(g_0(1,X) - g_0(0,X)\right) m_0(X)\right]}{\mathbb{E}_X\left[m_0(X)\right]}.$$

We view this as the major challenge in extending existing approaches of establishing lower bounds to the problem of estimating WATE and ATTE, and it is our main contribution in this paper to address it.

1.2 Notations

We use P_X to denote the marginal distribution of the confounding factors X in the model (1). For any function $f : \mathbb{R}^n \mapsto \mathbb{R}^k$ and distribution P over $\mathbb{R}^n$, we define its L^r-norm as

$$\|f\|_{P,r} = \left(\int \|f\|^r \, \mathrm{d}P\right)^{\frac{1}{r}}, \quad r \in (0, +\infty)$$

and

$$\|f\|_{P,\infty} = \text{ess sup}\left\{f(X) : X \sim P\right\}.$$

We also slightly abuse notation and use $\|f\|_r$ instead when the distribution is clear from context.

For two sequences $(a_n)_{n\geqslant 1}$ and $(b_n)_{n\geqslant 1}$, we write $a_n = \mathcal{O}(b_n)$ if there exists a constant $C > 0$ such that $|a_n| \leqslant C|b_n|, \forall n \geqslant 1$, and we write $a_n = \Omega(b_n)$ if there exists a constant $c > 0$ such that $|a_n| \geqslant c|b_n|, \forall n \geqslant 1$.

2. Structure-agnostic estimation of average treatment effect

As discussed in the previous section, while higher-order debiasing schemes yield minimax optimal estimation errors for functionals within certain smoothness classes, their improved rates crucially rely on the structural properties of the underlying function spaces, thereby losing the desirable property of being structure agnostic. To analyze the statistical limit of estimating treatment effect without making assumptions on regularity properties of nuisance functions, we adopt the structure-agnostic framework introduced by (Balakrishnan, Kennedy, and Wasserman 2023).

Specifically, we assume the existence of black-box estimates $\hat{m}(x)$ and $\hat{g}(d,x)$ of $m(x)$ and $g(d,x)$ that are accurate in the sense of L^2 distance:

$$\begin{aligned} &\left\|g_0(0,X) - \hat{g}(0,X)\right\|_{P_X,2}^2 \leqslant e_n, \left\|g_0(1,X) - \hat{g}(1,X)\right\|_{P_X,2}^2 \leqslant e_n', \\ &\left\|m_0(X) - \hat{m}(X)\right\|_{P_X,2}^2 \leqslant f_n, \end{aligned} \tag{14}$$

where e_n, e_n' and f_n are *unknown* positive numbers that depend on the sample size n used to estimate the nuisance functions. Note that here we assume that the estimators $\hat{m}, \hat{g}$ are already known to the statistician rather than a part of the estimation process. The reason for considering this setup is that we do not want to open the *black box* of how these estimators are obtained. In practice, these estimators can be obtained by leveraging powerful machine learning methods such as Lasso (Bickel, Ritov, and Tsybakov 2009), random forest (Syrgkanis and Zampetakis 2020), deep neural networks (Chen and White 1999; Schmidt-Hieber 2020; Farrell, Liang, and Misra 2021) among others. Ideally, we would like have a guarantee on the final estimation error that only depends on the nuisance estimation error (14) but not on algorithmic-dependent properties. Moreover, while we do not explicitly impose smoothness assumptions on the ground-truth nuisance functions m_0 and g_0, existing works that rely on such assumptions can still be related to our current setup, since the level of smoothness of the nuisance functions directly affects their minimax optimal estimation errors e_n and f_n (Kennedy et al. 2022).

Apart from the nuisance estimators, we also assume access to *i.i.d.* data $\{(X_i, D_i, Y_i)\}_{i=1}^n$ that are also *independent* of the data used to obtain nuisance estimators. In this way, we fully disentangle the complete estimation procedure into a *learning* phase where any machine learning methods can be used to obtain nuisance estimators from a portion of data, and an *improvement/correction* phase that leads to the final estimate. While estimation of WATE/ATTE does not necessarily follows this procedure, we notice that this is what actually being in practice where one performs sample splitting and use the remaining samples to correct the bias of the estimations in the first phase. Moreover, this procedure allows maximal flexibility of choosing the statistician's favorite ML estimator in the first phase.

We are interested in answering the following question: *what is the optimal error rate that we can achieve for estimating WATE and ATTE, given known estimators of nuisance functions and i.i.d. data* $\{(X_i, D_i, Y_i)\}_{i=1}^n$*?*

This question has already been answered in (Kennedy et al. 2022) for estimating the expected conditional covariance $\theta^{Cov} = \mathbb{E}\left[\mathrm{Cov}(D, Y \mid X)\right]$. Specifically, they assume the existence of black-box estimators of nuisance functions $\mu_0(x) = \mathbb{E}[Y \mid X = x]$ and $m_0(x)$ with errors e_n and f_n respectively, and derive a minimax optimal estimation error of $\Theta\left(e_n \cdot f_n + \frac{1}{n}\right)$. However, to the best of our knowledge, no such results are known for estimating WATE or ATTE. Indeed, as we discussed in Section 1.1, existing approaches for establishing minimax optimal error rates cannot be directly adapted to handle these quantities.

To begin with, we first introduce achievable structure-agnostic upper bounds that is quite well-understood in existing literature. We revisit double/debiased machine learning (Chernozhukov et al. 2017; Chernozhukov et al. 2018), a popular technique that performs a first-order bias correction to the naive plug-in estimators, as introduced in Section 1. Focusing on our setting, the following results are known for estimating WATE.

Theorem 2. *Suppose that there is a constant $c \in (0,1)$ such that $c \leqslant \hat{m}(x) \leqslant 1-c, \forall x \in \mathrm{supp}(X)$, then for any $\delta > 0$, there exists a constant C_δ such that the debiased estimator for WATE (defined in Equation (10)) achieves estimation error*

$$\left|\hat{\theta}^{WATE} - \theta^{WATE}\right|^2 \leqslant C_\delta \left(\max\{e_n, e_n'\} \cdot f_n \cdot \|w\|_{P_X,\infty}^2 + \frac{1}{n}\|w\|_{P_X,2}^2\right).$$

with probability $\geqslant 1-\delta$.

Theorem 2 implies that with high probability, the estimation error of the debiased estimator (10) is upper-bounded by the sum of the oracle error which equals $\frac{1}{n}$ multiplied by the L^2 norm of weight function w, and the product of the error in estimating nuisance functions m_0 and g_0, multiplied by the L^∞ norm of w. Similarly, for estimating ATTE, we have the following.

Theorem 3. *Suppose that there is a constant $c \in (0,1)$ such that $c \leqslant \hat{m}(x) \leqslant 1-c, \forall x \in \mathrm{supp}(X)$, then for any $\delta > 0$, there exists a constant C_δ such that the debiased estimator for the average treatment effect of the treated (defined in Equation (12)) achieves estimation error*

$$\left|\hat{\theta}^{ATTE} - \theta^{ATTE}\right|^2 \leqslant C_\delta \left(e_n \cdot f_n + \frac{1}{n}\right) \tag{15}$$

with probability $\geqslant 1-\delta$.

Theorem 3 implies that with high probability, the estimation error of the debiased estimator (12) is upper-bounded by the sum of the oracle error $\frac{1}{n}$ and the product of the error in estimating nuisance functions m_0 and $g_0(0,\cdot)$. The bound for ATTE is similar to that of WATE, except that it does not depend on e_n'.

Given the high-probability upper bounds in Theorem 2 and 3, it is natural to ask whether these structure-agnostic guarantees achieved by DML can be further improved. We will investigate this problem in the subsequent sections.

3. Main results

In this section, we present our main results that lower-bound the estimation errors in the structural-agnostic setting. Our lower bounds match the upper bounds derived in the previous section, implying that double/debiased ML estimators are structure-agnostic optimal in estimating WATE and ATTE.

We restrict ourselves to the case of binary outcomes:

Assumption 1. *The outcome variable Y is binary,* i.e., $Y \in \{0,1\}$.

Given that the black-box nuisance function estimators satisfy Equation (14), we define the following constraint set

$$\begin{aligned}\mathcal{F}_{e_n,e_n',f_n} = \Big\{ &(m,g) \mid \mathrm{supp}(X) = [0,1]^K, P_X = \mathrm{Uniform}\left([0,1]^K\right), \\ &\|g(0,X)-\hat{g}(0,X)\|_{P_X,2}^2 \leqslant e_n, \|g(1,X)-\hat{g}(1,X)\|_{P_X,2}^2 \leqslant e_n', \\ &\|m(X)-\hat{m}(X)\|_{P_X,2}^2 \leqslant f_n, 0 \leqslant m(x), g(d,x) \leqslant 1, \forall x \in [0,1]^K \Big\}\end{aligned} \tag{16}$$

where

$$e_n, e_n', f_n = o(1) \quad (n \to +\infty).$$

Note that introducing Assumption 2 and constraints on P_X in Equation (16) only strengthens the lower bound that we are going to prove, since they provide additional information on the ground-truth model. Moreover, the constraints $0 \leqslant m(x), g(d,x) \leqslant 1$ naturally holds due to the fact that both the treatment and outcome variables are binary. We then define the minimax $(1-\gamma)$-quantile risk of estimating θ^{WATE} over a function space $\mathcal{F}$ as

$$\mathfrak{M}_{n,\gamma}^{WATE}(\mathcal{F}) = \inf_{\hat{\theta}:(\mathcal{X}\times\mathcal{D}\times\mathcal{Y})^n \mapsto \mathbb{R}} \sup_{(m^*,g^*)\in\mathcal{F}} \mathcal{Q}_{P_{m^*,g^*},1-\gamma}\left(\left|\hat{\theta}-\theta^{WATE}\right|^2\right), \tag{17}$$

where $\mathcal{Q}_{P,\gamma}(X) = \inf\{x \in \mathbb{R} : P[X \leqslant x] \geqslant \gamma\}$ denotes the quantile function of a random variable X, and P_{m^*,g^*} is the joint distribution of (X,D,Y) which is uniquely determined by the functions m^* and g^*. Specifically, let μ be the uniform distribution on $\mathcal{X}\times\mathcal{D}\times\mathcal{Y} = [0,1]^K \times \{0,1\} \times \{0,1\}$, then the density $p_{m^*,g^*} = \frac{\mathrm{d}P_{m^*,g^*}}{\mathrm{d}\mu}$ can be expressed as

$$p_{m^*,g^*}(x,d,y) = m^*(x)^d(1-m^*(x))^{1-d} g^*(d,x)^y (1-g^*(d,x))^{1-y}.$$

According to Equation (17), $\mathfrak{M}_{n,\gamma}^{WATE}(\mathcal{F}) \geqslant \rho$ would imply that for any estimator $\hat{\theta}$ of WATE, there must exist some $(m^*,g^*) \in \mathcal{F}$, such that under the induced data distribution, the probability of $\hat{\theta}$ having estimation error $\geqslant \rho$ is at least $1-\gamma$. This provides a stronger form of lower bound compared with the minimax *expected* risk defined in (Balakrishnan, Kennedy, and Wasserman 2023), in the sense that the lower bound $\mathfrak{M}_{n,\gamma}^{WATE}(\mathcal{F}) \geqslant \rho$ implies a lower bound $(1-\gamma)\rho$ of the minimax expected risk, but the converse does not necessarily hold.

Similarly, one can define the minimax quantile risk for estimating ATTE as

$$\mathfrak{M}_{n,\gamma}^{ATTE}(\mathcal{F}) = \inf_{\hat{\theta}:(\mathcal{X}\times\mathcal{D}\times\mathcal{Y})^n \mapsto \mathbb{R}} \sup_{(m^*,g^*)\in\mathcal{F}} \mathcal{Q}_{P_{m^*,g^*},1-\gamma}\left(\left|\hat{\theta}\left(\{(X_i, D_i, Y_i)\}_{i=1}^n\right) - \theta^{ATTE}\right|^2\right). \quad (18)$$

The main objective of this section is to derive lower bounds for $\mathfrak{M}_{n,\gamma}^{WATE}\left(\mathcal{F}_{e_n,e_n',f_n}\right)$ and $\mathfrak{M}_{n,\gamma}^{ATTE}\left(\mathcal{F}_{e_n,e_n',f_n}\right)$ in terms of e_n, e_n', f_n and n.

To derive our lower bound, we also need to assume that the estimators $\hat{m}(x): [0,1]^K \mapsto [0,1]$ and $\hat{g}(d,x): \{0,1\} \times [0,1]^K \mapsto [0,1]$ are bounded away from 0 and 1.

Assumption 2. *There exists a constant c such that $c \leqslant \hat{m}(x), \hat{g}(d,x) \leqslant 1-c$ for all $d \in \{0,1\}$ and $x \in [0,1]^K$.*

The assumption that $c \leqslant \hat{m}(x) \leqslant 1-c$ is common in deriving upper bounds for the error induced by debiased estimators. On the other hand, the assumption that $c \leqslant \hat{g}(d,x) \leqslant 1-c$ is typically not needed for deriving upper bounds, but it is also made in prior works for proving *lower bounds* of estimating the expected conditional covariance $\mathbb{E}\left[\mathrm{Cov}(D,Y) \mid X\right]$ (Robins et al. 2009; Balakrishnan, Kennedy, and Wasserman 2023).

Now we are ready to state our main results.

Theorem 4. *For any constant $\gamma \in \left(\frac{1}{2}, 1\right)$ and estimators $\hat{m}(x)$ and $\hat{g}(d,x)$ that satisfy Assumption 2, for any given weight function w, the minimax risk of estimating the WATE is*

$$\mathfrak{M}_{n,\gamma}^{WATE}\left(\mathcal{F}_{e_n,e_n',f_n}\right) = \Omega\left(\max\{e_n, e_n'\} f_n \cdot \|w\|_{P_X,\infty}^2 + \frac{1}{n}\|w\|_{P_X,2}^2\right)$$

Remark 1. *If we only assume that $c \leqslant \hat{m}(x), \hat{g}(1,x) \leqslant 1-c$ in Assumption 2, then we would have the lower bound*

$$\mathfrak{M}_{n,\gamma}^{WATE}\left(\mathcal{F}_{e_n,e_n',f_n}\right) = \Omega\left(e_n' f_n \cdot \|w\|_{P_X,\infty}^2 + \frac{1}{n}\|w\|_{P_X,2}^2\right).$$

Furthermore, this lower bound still holds in the case where we know the baseline response, i.e., $\hat{g}(0,x) = g_0(0,x) = 0$.

Theorem 5. *For any constant $\gamma \in \left(\frac{1}{2}, 1\right)$ and estimators $\hat{m}(x)$ and $\hat{g}(d,x)$ that satisfy Assumption 2, the minimax risk of estimating the ATTE is given by*

$$\mathfrak{M}_{n,\gamma}^{ATTE}\left(\mathcal{F}_{e_n,e_n',f_n}\right) = \Omega\left(e_n \cdot f_n + \frac{1}{n}\right)$$

Theorems 4 and 5 provide lower bounds of the minimax estimation errors for the WATE and ATTE, in terms of the sample size and the estimation error of the black-box nuisance function estimators. Our lower bounds exactly matches the upper bounds in Theorems 2 and 3 attained by DML estimators, indicating that such estimators are minimax optimal in the structural-agnostic setup that we focus on.

4. Proof of Theorem 4

In this section, we give the detailed proof of our main result, Theorem 4, for the lower bound of estimating WATE. We first introduce some preliminary results that our proof will rely on.

4.1 Preliminaries

In this subsection, we introduce some known results that build the relationship between functional estimation and hypothesis testing, and then prove some preparatory results for the construction of hypotheses in subsequent sections. Let H be the Hellinger distance defined as

$$H(P,Q) = \frac{1}{2}\int \left(\sqrt{P(\mathrm{d}x)} - \sqrt{Q(\mathrm{d}x)}\right)^2$$

for any distributions P, Q. The first result that we will introduce is due to (Robins et al. 2009) and upper-bounds the Hellinger distance between two mixtures of product measures.

Formally, let $\mathcal{X} = \cup_{j=1}^m \mathcal{X}_j$ be a measurable partition of the sample space. Given a vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ in some product measurable space $\Lambda = \Lambda_1 \times \cdots \times \Lambda_m$, let P and Q_λ be probability measures on $\mathcal{X}$ such that the following statements hold:

- $P\left(\mathcal{X}_j\right) = Q_\lambda\left(\mathcal{X}_j\right) = p_j$ for every $\lambda \in \Lambda$, and
- The probability measures P and Q_λ restricted to $\mathcal{X}_j$ depend on the j-th coordinate λ_j of λ only.

Let p and q_λ be the densities of the measures P and Q_λ that are jointly measurable in the parameter λ and the observation x, and π be a probability measure on Λ. Define

$$b = m \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(q_\lambda - p)^2}{p} d\mu,$$

and the mixed density $q = \int q_\lambda d\pi(\lambda)$, then we have the following result.

Theorem 6. *((Robins et al. 2009), Theorem 2.1, simplified version) Suppose that the mixed densities are equal, i.e. that $q = p$, and that $n \max\{1, b\} \max_j p_j \leqslant A$ for all j for some positive constant A, then there exists a constant C that depends only on A such that, for any product probability measure $\pi = \pi_1 \otimes \cdots \otimes \pi_m$,*

$$H\left(P^{\otimes n}, \int Q_\lambda^{\otimes n} d\pi(\lambda)\right) \leqslant \max_j p_j \cdot C n^2 b^2.$$

Remark 2. *Theorem 6 considers a special case of (Robins et al. 2009), Theorem 2.1. The original variant of the theorem considers a more general setting where the measures p are also indexed by λ, i.e. p_λ, P_λ and where p is the mixture density. Here, we only need the special cases where all P_λ's are equal to P. The original version of the theorem also required that all p_λ satisfy that $\underline{B} \leqslant p \leqslant \bar{B}$ for some constants $\underline{B}, \bar{B}$. In our special case, we no longer need to assume that. The only step in the proof of (Robins et al. 2009) that makes use of this assumption is that*

$$\max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{p^2}{p_\lambda} \frac{d\mu}{p_j} \leqslant \frac{\bar{B}}{\underline{B}}$$

(see the arguments following their proof of Lemma 5.2). However, in our setting this term is simply

$$\max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{p^2}{p_\lambda} \frac{d\mu}{p_j} = \max_j p_j^{-1} \int_{\mathcal{X}_j} p d\mu = \max_j p_j^{-1} P(\mathcal{X}_j) = 1.$$

Theorem 7. *((Tsybakov 2008), Theorem 2.15) Let π be a probability distribution on a set (measure space) of distributions $\mathcal{P}$ with common support $\mathcal{X}$, which induce the distribution*

$$Q_1(A) = \int Q^{\otimes n}(A) d\pi(Q), \quad \forall A \subset \mathcal{P}.$$

Suppose that there exists $P \in \mathcal{P}$ and a functional $T : \mathcal{P} \mapsto \mathbb{R}$ which satisfies

$$T(P) \leqslant c, \quad \pi(\{Q : T(Q) \geqslant c + 2s\}) = 1 \tag{19}$$

for some $s > 0$. If $H^2\left(P^{\otimes n}, Q_1\right) \leqslant \delta < 2$, then:

$$\inf_{\hat{T}:\mathcal{X}^n \mapsto \mathbb{R}} \sup_{P \in \mathcal{P}} P\left[\left|\hat{T} - T(P)\right| \geqslant s\right] \geqslant \frac{1 - \sqrt{\delta(1-\delta/4)}}{2}.$$

Equivalently, let $\gamma = \frac{1+\sqrt{\delta(1-\delta/4)}}{2}$, then

$$\inf_{\hat{T}:\mathcal{X}^n \mapsto \mathbb{R}} \sup_{P \in \mathcal{P}} \mathcal{Q}_{P,1-\gamma}\left(\left|\hat{T} - T(P)\right|^2\right) \geqslant s^2.$$

4.2 Partitioning the covariate space

The following lemma states that for an arbitrary weight function $w(x)$, one can always partition the domain into two subsets that have the same amount of weights.

Definition 1. *We say that a set $\mathcal{S} \subseteq \mathbb{R}^K$ is a hyperrectangle collection if it can be partitioned into a finite number of disjoint hyperrectangles in $\mathbb{R}^K$.*

Lemma 1. *Let $\mathcal{S} \subseteq \mathbb{R}^K$ be a hyperrectangle collection and $w(x) : [0,1]^K \mapsto \mathbb{R}$ be a non-negative Lebesgue-integrable function such that*

$$\int_{[0,1]^K} w(x) d\mu_L(x) > 0,$$

then $\mathcal{S}$ can be partitioned into two hyperrectangle collections $\mathcal{S}_1, \mathcal{S}_2$ such that $\mu_L(\mathcal{S}_1) = \frac{1}{2}\mu_L(\mathcal{S})$ and

$$\int_{\mathcal{S}_1} w(x) d\mu_L(x) = \frac{1}{2}\int_{\mathcal{S}} w(x) d\mu_L(x),$$

where μ_L is the Lebesgue measure on $\mathbb{R}^K$.

Proof. Suppose that $\mathcal{S} = \cup_{i=1}^n \mathcal{C}_i$, where $\mathcal{C}_i = \times_{j=1}^K [a_{ij}, b_{ij}]$ are disjoint hyperrectangles. Let

$$\mathcal{T}_\alpha = \bigcup_{i=1}^n \left(\bigtimes_{j=1}^{K-1} [a_{ij}, b_{ij}] \times \left[\left(1 - \frac{\alpha}{2}\right) a_{iK} + \frac{\alpha}{2} b_{iK}, \frac{1-\alpha}{2} a_{iK} + \frac{1+\alpha}{2} b_{iK}\right]\right), \alpha \in [0,1],$$

then it is easy to see that $\mu_L(\mathcal{T}_\alpha) = \frac{1}{2}\mu_L(\mathcal{S})$ and that both $\mathcal{T}_\alpha$ and $\mathcal{S}\backslash\mathcal{T}_\alpha$ are hyperrectangle collections.[2]

For $\alpha_0 \in (0,1)$, dominated convergence theorem implies that $\lim_{\alpha\to\alpha_0} \int_{\mathcal{T}_\alpha} w(x) \mathrm{d}\mu_L(x) = \lim_{\alpha\to\alpha_0} \int_{\mathcal{S}} \mathbb{1}\{x \in \mathcal{T}_\alpha\} w(x) \mathrm{d}\mu_L(x) = \int_{\mathcal{S}} \mathbb{1}\{x \in \mathcal{T}_{\alpha_0}\} w(x) \mathrm{d}\mu_L(x) = \int_{\mathcal{T}_{\alpha_0}} w(x) \mathrm{d}\mu_L(x)$, so the mapping

$$\psi : [0,1] \mapsto \mathbb{R}, \quad \alpha \mapsto \int_{\mathcal{T}_\alpha} w(x) \mathrm{d}\mu_L(x)$$

2. Intuitively, $\mathcal{T}_\alpha$ splits $\mathcal{S}$ along the K-th dimension into two sets: one set that contains an interval of length $\frac{1}{2}(b_{iK} - a_{iK})$ that lies strictly inside the interval $[a_{iK}, b_{iK}]$ and one set that contains two disconnected intervals, one to the left of the aforementioned middle interval and of length $\frac{\alpha}{2}(b_{iK} - a_{iK})$ and one to the right of the aforementioned interval of length $\left(\frac{1}{2} - \frac{\alpha}{2}\right)(b_{iK} - a_{iK})$.

is continuous and satisfies

$$\psi(0) + \psi(1) = \int_{\mathcal{S}} w(x) \mathrm{d}\mu_L(x)$$

because $\mathcal{T}_0 \cup \mathcal{T}_1 = \mathcal{S}$ and $\mu_L(\mathcal{T}_0 \cap \mathcal{T}_1) = 0$, so there must exists some α such that:

$$\psi(\alpha) = \frac{\psi(0) + \psi(1)}{2} = \frac{1}{2} \int_{\mathcal{S}} w(x) \mathrm{d}\mu_L(x).$$

Hence we can choose $\mathcal{S}_1 = \mathcal{T}_\alpha$ and $\mathcal{S}_2 = \mathcal{S} \setminus \mathcal{T}_\alpha$, concluding the proof. $\square$

Let P_X be the uniform distribution on $\mathrm{supp}(X) = [0,1]^K$ and p_X be its density. We define the "truncated" weight function $\hat{w}(x) = w(x)\mathbb{1}\left\{w(x) > \frac{1}{2}\|w\|_{P_X,\infty}\right\}$. Applying Lemma 1 to $\hat{w} \cdot w$, recursively, for m times, with $m \in \mathbb{Z}_+$, we can partition $[0,1]^d$ into $M = 2^m$ hyperrectangle collections $B_1, B_2, \cdots, B_M$, such that $\mu_L(B_j) = \frac{1}{M}$ and

$$\int_{B_j} w(x)\hat{w}(x)\mathrm{d}x = \frac{1}{M}, \quad j = 1, 2, \cdots, M.$$

Since P_X is the uniform distribution on $[0,1]^d$, the above implies that

$$\mathbb{E}_X\left[w(X)\hat{w}(X)\mathbb{1}\{X \in B_j\}\right] = \frac{1}{M}, \quad j = 1, 2, \cdots, M.$$

Let $\lambda_i, i = 1, 2, \cdots, M$ be i.i.d. Rademacher random variables taking values $+1$ and -1 both with probability 0.5. We define

$$\Delta(\lambda, x) = \sum_{j=1}^{M/2} \lambda_j \left(\mathbb{1}\left\{x \in B_{2j}\right\} - \mathbb{1}\left\{x \in B_{2j-1}\right\}\right). \tag{20}$$

The following properties of $\Delta(\lambda, x)$ are straightforward.

Proposition 1. *We have*

$$\mathbb{E}_\lambda \Delta(\lambda, x) = \sum_{j=1}^{M/2} \mathbb{E}\lambda_j \left(\mathbb{1}\left\{x \in B_{2j}\right\} - \mathbb{1}\left\{x \in B_{2j-1}\right\}\right) = 0, \quad \forall x \in [0,1]^K \tag{21a}$$

$$\mathbb{E}_X w(X)\hat{w}(X)\Delta(\lambda, X) = \sum_{j=1}^{M/2} \lambda_j \left(\frac{1}{M} - \frac{1}{M}\right) = 0, \quad \forall \lambda \in \{0,1\}^{M/2} \tag{21b}$$

$$\Delta(\lambda, x)^2 = \sum_{j=1}^{M/2} \left(\mathbb{1}\left\{x \in B_{2j}\right\} - \mathbb{1}\left\{x \in B_{2j-1}\right\}\right)^2 = 1, \quad \forall x \in [0,1]^K, \lambda \in \{0,1\}^{M/2}. \tag{21c}$$

Remark 3. *The construction of "bump" functions $\Delta(\lambda, x)$ in the form of Equation (20) has also been used in a line of prior works for proving minimax lower bounds (Balakrishnan, Kennedy, and Wasserman 2023). However, here we need to carefully construct the partition B_j of the whole domain to handle non-uniform weights. We note that if we only wanted to deal with an Average Treatment Effect and not a Weighted Average Treatment Effect, then we would have simply chosen B_j to be an equi-partition of the $[0,1]^K$ space and the above constructions of the regions B_j, related to balancing the given weights, would not be needed.*

4.3 Core part of lower bound construction

Having completed all preparation steps, we are now ready to present our proof for Theorem 4. The remaining part of Section 4 is organized as follows. In Section 4.4 and 4.5, we first establish our lower bound $\Omega\left(e_n' f_n \|w\|_{P_X,\infty}^2\right)$ under the following weaker version of Assumption 2, as previously mentioned in Remark 1:

Assumption 3. *There exists a constant $c > 0$ such that $c \leqslant \hat{m}(x), \hat{g}(1,x) \leqslant 1 - c$ for all $x \in [0,1]^K$.*

We separately present our proof of this lower bound for the two cases $e_n' \geqslant f_n$ and $e_n' < f_n$. Interestingly, these two cases need to be handled separately using different constructions of the composite hypotheses. In Section 4.6, we show how the lower bound $\Omega\left(e_n f_n \|w\|_{P_X,\infty}^2\right)$ can be derived in a similar fashion. To conclude our proof, it remains to prove the lower bound $\mathcal{O}\left(n^{-1}\|w\|_{P_X,2}^2\right)$, which is the standard oracle error and can be found in the supplementary material.

4.4 Case 1: $e_n \geqslant f_n$

In this case, we define

$$\begin{aligned} g_\lambda(0,x) &= \hat{g}(0,x) \\ m_\lambda(x) &= \hat{m}(x)\left[1 - \frac{\beta}{\hat{g}(1,x)}\hat{w}(x)\Delta(\lambda,x)\right] \\ g_\lambda(1,x) &= \frac{\hat{m}(x)}{m_\lambda(x)}\left[\hat{g}(1,x) + \alpha\hat{w}(x)\Delta(\lambda,x)\right], \end{aligned} \tag{22}$$

where $\alpha, \beta > 0$ are constants that will be specified later in Lemma 5, where we will verify that (m_λ, g_λ) belongs to the constrained set $\mathcal{F}_{e_n,e_n',f_n}$ and thus are valid probabilities in particular. Compared with standard approaches for constructing the composite hypotheses (Ingster 1994; Robins et al. 2009; Arias-Castro, Pelletier, and Saligrama 2018; Balakrishnan and Wasserman 2019), we employ an *asymmetric* construction which means that the nuisance functions are *non-linear* in the Rademacher variables λ (in particular the function g_λ depends non-linearly in λ due to the dependence on m_λ in the denominator). As discussed in Section 1.1, such type of non-standard constructions are necessary since the functional that we need to estimate has a different structure than those handled in previous works.

We first prove some basic properties of our construction.

Proposition 2. *For all $x \in [0,1]^K$, we have*

$$\mathbb{E}_\lambda m_\lambda(x) = \hat{m}(x) - \hat{m}(x)\frac{\beta}{\hat{g}(1,x)}\hat{w}(x)\mathbb{E}_\lambda\Delta(\lambda,x) = \hat{m}(x) \tag{23a}$$

$$\mathbb{E}_\lambda\left[g_\lambda(1,x)m_\lambda(x)\right] = \hat{m}(x)\left(\hat{g}(1,x) + \alpha\hat{w}(x)\mathbb{E}_\lambda\Delta(\lambda,x)\right) = \hat{g}(1,x)\hat{m}(x). \tag{23b}$$

We start by bounding the L^2 distance from g_λ, m_λ to $\hat{g}, \hat{m}$.

Lemma 2. *Assuming that $\beta \leqslant \frac{1}{2}c\|w\|_{P_X,\infty}^{-1}$ where c is the constant introduced in Assumption 2, then the following holds for all $0 < r \leqslant +\infty$:*

$$\begin{aligned} \left\|g_\lambda(1,X) - \hat{g}(1,X)\right\|_{P_X,r} &\leqslant 2(\alpha + c^{-1}\beta)\|\hat{w}(X)\|_{P_X,r}, \\ \left\|m_\lambda(X) - \hat{m}(X)\right\|_{P_X,r} &\leqslant c^{-1}\beta\|\hat{w}(X)\|_{P_X,r}. \end{aligned}$$

Proof. We have

$$\left\|m_\lambda(X) - \hat{m}(X)\right\|_{P_X,r} = \beta \left\| \frac{\hat{w}(X)\Delta(\lambda, X)}{\hat{g}(1,X)} \right\|_{P_X,r} \leqslant c^{-1}\beta \|\hat{w}(X)\|_{P_X,r}$$

and

$$\begin{aligned}\left\|g_\lambda(1,X) - \hat{g}(1,X)\right\|_{P_X,r} &\leqslant \left\| \frac{\hat{m}(X) - m_\lambda(X)}{m_\lambda(X)} \hat{g}(1,X) \right\|_{P_X,r} + \alpha \left\| \frac{\hat{m}(X)}{m_\lambda(X)} \hat{w}(X) \right\|_{P_X,r} \\ &\leqslant 2(c^{-1}\beta + \alpha)\|\hat{w}(X)\|_{P_X,r}.\end{aligned}$$

□

Let Q_λ be the joint distribution of (X, D, Y) induced by g_λ and m_λ and μ be the uniform distribution on $[0,1]^K \times \{0,1\} \times \{0,1\}$. Define $q_\lambda = \frac{\mathrm{d}Q_\lambda}{\mathrm{d}\mu}$. Similarly, let $\hat{P}$ be the joint distribution of (X, D, Y) induced by $\hat{g}$ and $\hat{m}$, and $\hat{p} = \frac{\mathrm{d}\hat{P}}{\mathrm{d}\mu}$. The next lemma states that the mixture of Q_λ with prior $\pi(\lambda)$ is exactly equal to $\hat{P}$.

Lemma 3. *Let* $Q = \int Q_\lambda d\pi(\lambda)$ *and* $q = \frac{dQ}{d\mu} = \int q_\lambda d\pi(\lambda)$*, then* $\hat{p} = q$*.*

Proof. By definition, we have

$$q_\lambda(x,d,y) = m_\lambda(x)^d(1 - m_\lambda(x))^{1-d} g_\lambda(d,x)^y (1 - g_\lambda(d,x))^{1-y}$$

and

$$\hat{p}(x,d,y) = \hat{m}(x)^d(1 - \hat{m}(x))^{1-d}\hat{g}(d,x)^y(1 - \hat{g}(d,x))^{1-y}.$$

The "mixed" joint density q is then given by

$$\begin{aligned} q(x,d,y) &= \int q_\lambda(x,d,y)\mathrm{d}\pi(\lambda) \\ &= \int m_\lambda(x)^d(1 - m_\lambda(x))^{1-d} g_\lambda(d,x)^y(1 - g_\lambda(d,x))^{1-y}\mathrm{d}\pi(\lambda) \end{aligned}$$

When $d = 1$, we have

$$q(x,1,y) = \begin{cases} \int m_\lambda(x) g_\lambda(1,x)\mathrm{d}\pi(\lambda) & \text{if } y = 1 \\ \int m_\lambda(x)\left(1 - g_\lambda(1,x)\right)\mathrm{d}\pi(\lambda) & \text{if } y = 0. \end{cases}$$

By Equation (23), we know that

$$\int m_\lambda(x) g_\lambda(1,x)\mathrm{d}\pi(\lambda) = \hat{m}(x)\hat{g}(1,x) = \hat{p}(x,1,1)$$

and

$$\int m_\lambda(x)\left(1 - g_\lambda(1,x)\right)\mathrm{d}\pi(\lambda) = \hat{m}(x) - \hat{m}(x)\hat{g}(1,x) = \hat{p}(x,1,0),$$

thus $q(x,1,y) = \hat{p}(x,1,y), y \in \{0,1\}$.

When $d = 0$, recall that $\hat{g}(0,x) = g_\lambda(0,x)$ by our construction, so we have

$$\begin{aligned} q(x,0,y) &= \int (1 - m_\lambda(x))\hat{g}(0,x)^y(1 - \hat{g}(0,x))^{1-y}\mathrm{d}\pi(\lambda) \\ &= (1 - \hat{m}(x))\hat{g}(0,x)^y(1 - \hat{g}(0,x))^{1-y} = \hat{p}(x,0,y). \end{aligned}$$

where we again use Equation (23a) in the second equation. Hence $\hat{p} = q$ as desired. □

The following lemma implies that the Hellinger distance between the empirical distribution under $\hat{P}$ and Q_λ with prior $\pi(\lambda)$ can be made arbitrarily small, as long as the domain $\mathrm{supp}(X)$ is partitioned into sufficiently many pieces.

Lemma 4. *For any $\delta > 0$, as long as $M \geqslant \max\{n, \frac{32C}{c^4\delta}n^2\}$ where c is the constant introduced in Assumption 3 and C is the constant implied by Theorem 6 for $A = 4c^{-2}$, we have*

$$H^2\left(\hat{P}^{\otimes n}, \int Q_\lambda^{\otimes n} d\pi(\lambda)\right) \leqslant \delta.$$

Proof. We prove this lemma by applying Theorem 6 to the partition

$$\mathcal{X}_j = \left(B_{2j-1} \cup B_{2j}\right) \times \{0,1\} \times \{0,1\}, \quad j = 1, 2, \cdots, M/2$$

of $[0,1]^K \times \{0,1\} \times \{0,1\}$, $p = \hat{p}$ and q_λ as constructed above, and μ being the uniform distribution over $[0,1]^K \times \{0,1\} \times \{0,1\}$. Recall that B_j's are chosen to satisfy $\mu_L(B_j) = \frac{1}{M}$ where μ_L is the Lebesgue measure, so that

$$p_j := \hat{P}(\mathcal{X}_j) = Q_\lambda(\mathcal{X}_j) = \mu_L(B_{2j-1}) + \mu_L(B_{2j}) = \frac{2}{M} \tag{24}$$

since their marginal distribution P_X is the uniform distribution. Also, since for any $x \in \mathcal{X}_j$ we have $\Delta(\lambda, x) = \lambda_j(\mathbb{1}\{x \in B_{2j-1}\} - \mathbb{1}\{x \in B_{2j}\})$, the distribution Q_λ restricted to $\mathcal{X}_j$ only depends on λ_j. It follows from Equation (24) that

$$\begin{aligned} b &= \frac{M}{2} \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(q_\lambda - \hat{p})^2}{\hat{p}} d\mu \\ &\leqslant \max_j \frac{M}{2} p_j \cdot \sup_{(x,d,y)\in\mathcal{X}_j} \frac{(\hat{p}(x,d,y) - q_\lambda(x,d,y))^2}{\hat{p}(x,d,y)} \\ &\leqslant \frac{4}{c^2}, \end{aligned}$$

where the last step holds since

$$\hat{p}(x,1,y) \geqslant p_X(x) \cdot \min\left\{\hat{m}(x), 1 - \hat{m}(x)\right\} \cdot \min\left\{\hat{g}(1,x), 1 - \hat{g}(1,x)\right\} \geqslant c^2$$

by Assumption 3, which implies that

$$\frac{(\hat{p}(x,1,y) - q_\lambda(x,1,y))^2}{\hat{p}(x,1,y)} \leqslant \frac{4}{c^2},$$

and for all $(x,0,y) \in \mathrm{supp}(\hat{P})$,

$$\begin{aligned} \frac{(\hat{p}(x,0,y) - q_\lambda(x,0,y))^2}{\hat{p}(x,0,y)} &\leqslant \frac{(m_\lambda(x) - \hat{m}(x))^2 \hat{g}(0,x)^{2y}(1 - \hat{g}(0,x))^{2(1-y)}}{(1 - \hat{m}(x))\hat{g}(0,x)^y(1 - \hat{g}(0,x))^{1-y}} \\ &\leqslant \frac{4}{c}. \end{aligned}$$

Hence we have

$$Cn^2\left(\max_j p_j\right) b^2 \leqslant \frac{32Cn^2}{c^4 M} \leqslant \delta$$

Finally, we have $n \max\{1, b\} \max_j p_j \leqslant 4nc^{-2}M^{-1} \leqslant 4c^{-2} = A$ by our choice of M, so all conditions of Theorem 6 hold. By Theorem 6, we can conclude that $H^2(\hat{P}, Q) \leqslant \delta$. $\square$

As the final building block for establishing our lower bound, we prove the following lemma, which implies that with proper choices of α and β, m_λ, g_λ are close (in the sense of L^2-distance) to $\hat{m}$ and $\hat{g}$ respectively, and that the separation condition (19) holds with distance $s = \Omega\left(\sqrt{e_n f_n}\,\|w\|_{P_X,\infty}\right)$.

Lemma 5. *Let*

$$\alpha = \frac{\sqrt{e_n'}}{4\|\hat{w}(X)\|_{P_X,2}}, \quad \beta = \frac{c\sqrt{f_n}}{4\|\hat{w}(X)\|_{P_X,2}},$$

then for sufficiently large n, we have $(m_\lambda, g_\lambda) \in \mathcal{F}_{e_n,e_n',f_n}$ *and*

$$\mathbb{E}_X\left[w(X)g_\lambda(1,X)\right] \geqslant \mathbb{E}\left[w(X)\hat{g}(1,X)\right] + \frac{1}{2}\alpha\beta\mathbb{E}\left[\frac{w(X)\hat{w}(X)^2}{\hat{g}(1,X)}\right], \forall\lambda \in \{0,1\}^{M/2}. \tag{25}$$

Proof. Our assumption that $e_n' \geqslant f_n$ implies that $\alpha \geqslant \beta$. Since $e_n', f_n = o(1)(n \to +\infty)$, for sufficiently large n we must have

$$\max\{\alpha,\beta\} \leqslant \frac{1}{4}c^2\left(1+\|w(X)\|_{P_X,\infty}\right)^{-4}\min\left\{1, \mathbb{E}\left[\frac{w(X)\hat{w}(X)^2}{\hat{g}(1,X)}\right]\right\}, \tag{26}$$

where c is the constant introduced in Assumption 3. In the remaining part of the proof we will assume that Equation (26) holds.

First, by Lemma 2 and our choice of α and β it is easy to see that

$$\begin{aligned}\left\|\hat{m}(X) - m_\lambda(X)\right\|_{P_X,2} &\leqslant c^{-1}\beta\|\hat{w}(X)\|_{P_X,2} \leqslant \sqrt{f_n}\\ \left\|\hat{g}(1,X) - g_\lambda(1,X)\right\|_{P_X,2} &\leqslant (\alpha + c^{-1}\beta)\|\hat{w}(X)\|_{P_X,2} \leqslant \sqrt{e_n'}.\end{aligned}$$

Note that the second inequality above makes use of our assumption that $e_n' \geqslant f_n$. Again applying Lemma 2 with $r = \infty$, we have $\left\|\hat{g}(1,X) - g_\lambda(1,X)\right\|_{P_X,\infty} \leqslant (\alpha + c^{-1}\beta)\|\hat{w}(X)\|_{P_X,\infty} \leqslant \frac{1}{2}c$ which implies that $0 \leqslant g_\lambda \leqslant 1$. Similarly we have $0 \leqslant m_\lambda \leqslant 1$, so $(m_\lambda, g_\lambda) \in \mathcal{F}_{e_n,e_n',f_n}$.

It remains to show that Equation (25) holds. To see this, note that for fixed $\lambda \in \{0,1\}^{M/2}$ we have

$$
\begin{aligned}
&\mathbb{E}\left[w(X)g_\lambda(1,X)\right] && \text{(27a)}\\
&= \mathbb{E}\left[w(X)\frac{\hat{g}(1,X)+\alpha\hat{w}(X)\Delta(\lambda,X)}{1-\frac{\beta}{\hat{g}(1,X)}\hat{w}(X)\Delta(\lambda,X)}\right] && \text{(27b)}\\
&= \mathbb{E}\left[w(X)\left(\hat{g}(1,X)+\alpha\hat{w}(X)\Delta(\lambda,X)\right)\sum_{k=0}^{+\infty}\left(\frac{\beta}{\hat{g}(1,X)}\hat{w}(X)\Delta(\lambda,X)\right)^k\right] && \text{(27c)}\\
&= \mathbb{E}\left[w(X)\left(\hat{g}(1,X)+(\alpha+\beta)\hat{w}(X)\Delta(\lambda,X)+\frac{\alpha\beta+\beta^2}{\hat{g}(1,X)}\hat{w}(X)^2\right)\right]\\
&\quad+\mathbb{E}\left[\frac{\alpha\beta^2}{\hat{g}(1,X)^2}w(X)\hat{w}(X)^3\Delta(\lambda,X)\right] && \text{(27d)}\\
&\quad+\mathbb{E}\left[w(X)\left(\hat{g}(1,X)+\alpha\hat{w}(X)\Delta(\lambda,X)\right)\sum_{k=3}^{+\infty}\left(\frac{\beta}{\hat{g}(1,X)}\hat{w}(X)\Delta(\lambda,X)\right)^k\right] && \text{(27e)}\\
&\geqslant \mathbb{E}\left[w(X)\hat{g}(1,X)\right]+\alpha\beta\mathbb{E}\left[\frac{w(X)\hat{w}(X)^2}{\hat{g}(1,X)}\right]\\
&\quad-c^{-2}\alpha\beta^2\|w\|^4_{P_X,\infty}-2c\|w\|_{P_X,\infty}\sum_{k=3}^{+\infty}\left(c^{-1}\beta\|w\|_{P_X,\infty}\right)^k && \text{(27f)}\\
&\geqslant \mathbb{E}\left[w(X)\hat{g}(1,X)\right]+\alpha\beta\mathbb{E}\left[\frac{w(X)\hat{w}(X)^2}{\hat{g}(1,X)}\right]-C_0(\alpha\beta^2+\beta^3), && \text{(27g)}
\end{aligned}
$$

where Equation (27b) follows from our construction in Equation (22), Equation (27c) uses a Taylor expansion which is valid since Equation (26) implies that $\left|\frac{\beta}{\hat{g}(1,X)}\hat{w}(X)\Delta(\lambda,X)\right| \leqslant c^{-1}\|w\|_{P_X,\infty}\beta \leqslant \frac{1}{2}$, Equation (27e) follows from a direct expansion of Equation (27c) up to the second-order term, Equation (27f) is deduced by noticing that $\mathbb{E}_X w(X)\hat{w}(X)\Delta(\lambda,X)=0$ (by Proposition 1) and $\frac{\hat{w}(X)^2}{\hat{g}(1,X)} \geqslant 0$ and $\hat{g}(1,X)\geqslant c$ and using the upper bound on α by Equation (26). Finally, Equation (27g) holds for $C_0 = 4c^{-2}\|w\|^4_{P_X,\infty}$, invoking also the identity $\sum_{k=3}^{\infty} t^k = t^3/(1-t)$ for $t = c^{-1}\beta|0w|0_{P_X,\infty} \leqslant 1/2$. Here, it is important to note that our construction in Equation (22) exactly ensures that the first-order terms (in α and β) cancel out. Finally, Equation (26) and $\alpha \geqslant \beta$ together imply that $C_0(\alpha\beta^2+\beta^3) \leqslant 2C_0\alpha\beta^2 \leqslant \frac{1}{2}\mathbb{E}\left[\frac{w(X)\hat{w}(X)^2}{\hat{g}(1,X)}\right]\alpha\beta$, so Equation (25) immediately follows from Equation (27), concluding the proof. □

We are now ready to prove Theorem 4 in the case when $e_n' \geqslant f_n$. For any $\gamma > \frac{1}{2}$, there exists some $\delta \in (0,2)$ such that $\frac{1+\sqrt{\delta(1-\delta/4)}}{2} = \gamma$. We choose $M \geqslant \max\{n, \frac{32C}{c^4\delta}n^2\}$ and $\mathcal{P} = \{\hat{P}\}\cup\left\{Q_\lambda : \lambda\in\{0,1\}^{M/2}\right\}$, $P=\hat{P}$, π be the discrete uniform distribution on $\left\{Q_\lambda : \lambda\in\{0,1\}^{M/2}\right\}$, $s = \frac{1}{4}\alpha\beta\mathbb{E}\left[\frac{w(X)\hat{w}(X)^2}{\hat{g}(1,X)}\right]$ in the context of Theorem 7. Then Lemma 4 and 5 imply that all the listed conditions are satisfied for the WATE functional

$$T(P) = \theta^{WATE}(P) = \mathbb{E}_P\left[w(X)\left(g(1,X)-g(0,X)\right)\right].$$

Therefore, by Theorem 7, we have

$$
\begin{aligned}
&\inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathcal{Q}_{P,1-\gamma}\left(\left|\hat{\theta}\left(\{(X_i,D_i,Y_i)\}_{i=1}^N\right)-\theta^{WATE}\right|^2\right)\\
&=\Omega\left(\alpha\beta\mathbb{E}\left[\frac{w(X)\hat{w}(X)^2}{\hat{g}(1,X)}\right]\right)\\
&=\Omega\left(\frac{\sqrt{e_n' f_n}}{\|\hat{w}(X)\|_{P_X,2}^2}\cdot\mathbb{E}\left[\frac{w(X)\hat{w}(X)^2}{\hat{g}(1,X)}\right]\right)\\
&=\Omega\left(\sqrt{e_n' f_n}\cdot\frac{\mathbb{E}\left[w(X)\hat{w}(X)^2\right]}{\|\hat{w}(X)\|_{P_X,2}^2}\right)\\
&=\Omega\left(\sqrt{e_n' f_n}\cdot\frac{\mathbb{E}\left[w(X)^3\mathbb{1}\left\{w(X)>\frac{1}{2}\|w\|_{P_X,\infty}\right\}\right]}{\mathbb{E}\left[w(X)^2\mathbb{1}\left\{w(X)>\frac{1}{2}\|w\|_{P_X,\infty}\right\}\right]}\right)\\
&=\Omega\left(\|w\|_{P_X,\infty}\sqrt{e_n' f_n}\right).
\end{aligned}
$$

4.5 Case 2: $f_n > e_n'$

In this case, we consider a different construction as follows:

$$
\begin{aligned}
g_\lambda(0,x) &= \hat{g}(0,x)\\
g_\lambda(1,x) &= \frac{\hat{g}(1,x)}{1+\frac{\beta}{\hat{g}(1,x)}\hat{w}(x)\Delta(\lambda,x)-\alpha\beta\hat{w}(x)^2}\\
m_\lambda(x) &= \frac{\hat{g}(1,x)}{g_\lambda(1,x)}\left(\hat{m}(x)+\alpha\hat{m}(x)\hat{g}(1,x)\hat{w}(x)\Delta(\lambda,x)\right)
\end{aligned}
\tag{28}
$$

where $\Delta(\lambda,x)$ is defined in Equation (20) and $\alpha,\beta>0$ are constants that will be specified later. Parallel to Proposition 2 and Lemma 2, we first prove some basic properties of our construction.

Proposition 3. *We have*

$$
\mathbb{E}_\lambda\left[m_\lambda(x)g_\lambda(1,x)\right]=\hat{m}(x)\hat{g}(1,x)\quad \textit{and}
$$
$$
\mathbb{E}_\lambda\left[m_\lambda(x)\right]=\hat{m}(x)
$$

Proof. By Proposition 1, we have

$$
\mathbb{E}_\lambda\left[m_\lambda(x)g_\lambda(1,x)\right]=\hat{m}(x)\hat{g}(1,x)+\alpha\hat{m}(x)\hat{g}(1,x)^2\hat{w}(x)\mathbb{E}_\lambda\Delta(\lambda,x)=\hat{m}(x)\hat{g}(1,x)
$$

$$
\begin{aligned}
\mathbb{E}_\lambda\left[m_\lambda(x)\right]&=\mathbb{E}_\lambda\left[\left(\hat{m}(x)+\alpha\hat{m}(x)\hat{g}(1,x)\hat{w}(x)\Delta(\lambda,x)\right)\left(1+\frac{\beta}{\hat{g}(1,x)}\hat{w}(x)\Delta(\lambda,x)-\alpha\beta\hat{w}(x)^2\right)\right]\\
&=\hat{m}(x)+\left[\alpha(1-\alpha\beta\hat{w}(x)^2)\hat{m}(x)\hat{g}(1,x)\hat{w}(x)+\beta\frac{\hat{m}(x)}{\hat{g}(1,x)}\hat{w}(x)\right]\mathbb{E}_\lambda\Delta(\lambda,x)\\
&\quad-\alpha\beta\hat{m}(x)\left(1-\mathbb{E}_\lambda\Delta(\lambda,x)^2\right)\hat{w}(x)^2\\
&=\hat{m}(x).
\end{aligned}
$$

□

Lemma 6. *Assuming that* $\alpha \leqslant \max\{1, \|w\|_{P_X,\infty}\}^{-1}$ *and* $\beta \leqslant \frac{1}{4}c \cdot \max\{1, \|w\|_{P_X,\infty}\}^{-2}$ *where* c *is a constant introduced in Assumption 2, then the following holds for all* $0 < r \leqslant +\infty$*:*

$$\begin{aligned}\left\|g_\lambda(1,X) - \hat{g}(1,X)\right\|_{P_X,r} &\leqslant 2\beta\|\hat{w}(X)\|_{P_X,r}\\ \left\|m_\lambda(X) - \hat{m}(X)\right\|_{P_X,r} &\leqslant 2(\alpha + c^{-1}\beta)\|\hat{w}(X)\|_{P_X,r}.\end{aligned}$$

Proof. From our assumptions on α and β, one can see that $\left|\frac{\beta}{\hat{g}(1,x)}\hat{w}(x)\Delta(\lambda,x) - \alpha\beta\hat{w}(x)^2\right| \leqslant c^{-1}\beta|0w|0_{P_X,\infty} + \beta|0w|0^2_{P_X,\infty} \leqslant \frac{1}{2}$. Thus it follows that

$$\begin{aligned}\left\|g_\lambda(1,X) - \hat{g}(1,X)\right\|_{P_X,r} &\leqslant 2\left\|\hat{g}(1,X)\left(\frac{\beta}{\hat{g}(1,X)}\hat{w}(X)\Delta(\lambda,X) - \alpha\beta\hat{w}(X)^2\right)\right\|_{P_X,r}\\ &\leqslant (\beta + \alpha\beta|0w|0_{P_X,\infty})|0w(X)|0_{P_X,r} \leqslant 2|0w(X)|0_{P_X,r}\end{aligned}$$

and

$$\begin{aligned}\left\|m_\lambda(X) - \hat{m}(X)\right\|_{P_X,r} &\leqslant 2c^{-1}\beta|0w|0_{P_X,\infty} + \alpha|0w|0_{P_X,\infty} + 2c^{-1}\alpha\beta|0w|0^2_{P_X,\infty}\\ &\leqslant 2(\alpha + c^{-1}\beta)\|\hat{w}(X)\|_{P_X,r}.\end{aligned}$$

□

Note that a key difference between Lemma 6 and Lemma 2 is that in the former lemma, the deviations of g_λ and m_λ are $\mathcal{O}(\beta)$ and $\mathcal{O}(\alpha+\beta)$ respectively, while the converse is true in the latter one. This difference is intentional, since here we assume that $f_n > e'_n$.

Let Q_λ be the joint distribution of (X, D, Y) induced by g_λ and m_λ and μ be the uniform distribution on $[0,1]^K \times \{0,1\} \times \{0,1\}$. Define $q_\lambda = \frac{dQ_\lambda}{d\mu}$. Similarly, let $\hat{P}$ be the joint distribution of (X, D, Y) induced by $\hat{g}$ and $\hat{m}$, and $\hat{p} = \frac{d\hat{P}}{d\mu}$. Using exactly the same arguments as we did in Lemma 3 and 4, one can prove the following lemmas.

Lemma 7. *Let* $Q = \int Q_\lambda d\pi(\lambda)$ *and* $q = \frac{dQ}{d\mu} = \int q_\lambda d\pi(\lambda)$*, then* $\hat{p} = q$*.*

Lemma 8. *For any* $\delta > 0$*, as long as* $M \geqslant \max\{n, \frac{32C}{c^4\delta}n^2\}$ *where* c *is the constant introduced in Assumption 3 and* C *is the constant implied by Theorem 6 for* $A = 4c^{-2}$*, we have*

$$H^2\left(\hat{P}^{\otimes n}, \int Q_\lambda^{\otimes n} d\pi(\lambda)\right) \leqslant \delta.$$

Finally, we prove the analogue of Lemma 5 for the different construction that we are now considering.

Lemma 9. *Let*

$$\alpha = \frac{\sqrt{f_n}}{4\|\hat{w}(X)\|_{P_X,2}}, \quad \beta = \frac{c\sqrt{e'_n}}{4\|\hat{w}(X)\|_{P_X,2}},$$

then for sufficiently large n*, we have* $(m_\lambda, g_\lambda) \in \mathcal{F}_{e_n,e'_n,f_n}$*, and* $\forall \lambda \in \{0,1\}^{M/2}$*:*

$$\mathbb{E}_X\left[w(X)g_\lambda(1,X)\right] \geqslant \mathbb{E}\left[w(X)\hat{g}(1,X)\right] + \frac{1}{2}\alpha\beta\mathbb{E}_X\left[\hat{g}(1,X)w(X)\hat{w}(X)^2\right] \tag{29}$$

Proof. Since $e_n', f_n = o(1)(n \to +\infty)$, for sufficiently large n we must have

$$\max\{\alpha, \beta\} < \frac{1}{4}c^2(1 + \|w\|_{P_X,\infty})^{-4} \min\left\{1, \mathbb{E}_X\left[\hat{g}(1,X)w(X)\hat{w}(X)^2\right]\right\}, \tag{30}$$

where c is the constant introduced in Assumption 3. First, by Lemma 6 our choice of α and β it is easy to see that

$$\begin{aligned}
\left\|\hat{m}(X) - m_\lambda(X)\right\|_{P_X,2} &\leqslant 2(\alpha + c^{-1}\beta)\|\hat{w}(X)\|_{P_X,2} \leqslant \sqrt{f_n} \\
\left\|\hat{g}(1,X) - g_\lambda(1,X)\right\|_{P_X,2} &\leqslant 2\beta\|\hat{w}(X)\|_{P_X,2} \leqslant \sqrt{e_n'}.
\end{aligned}$$

Note that the first inequality above makes use of our assumption that $f_n > e_n'$. Applying Lemma 6 with $r = \infty$, Equation (30) implies that $0 \leqslant m_\lambda, g_\lambda \leqslant 1$. Hence $(m_\lambda, g_\lambda) \in \mathcal{F}_{e_n, e_n', f_n}$.

It remains to show that Equation (29) holds. Note that for fixed $\lambda \in \{0,1\}^{M/2}$ we have

$$\begin{aligned}
&\mathbb{E}_X\left[w(X)g_\lambda(1,X)\right] \\
&= \mathbb{E}_X\left[w(X)\frac{\hat{g}(1,X)}{1 + \frac{\beta}{\hat{g}(1,X)}\hat{w}(X)\Delta(\lambda,X) - \alpha\beta\hat{w}(X)^2}\right] \\
&= \mathbb{E}_X\left[w(X)\hat{g}(1,X)\left(1 + \sum_{k=1}^{+\infty}\beta^k\left(\alpha\hat{w}(X)^2 - \frac{1}{\hat{g}(1,X)}\hat{w}(X)\Delta(\lambda,X)\right)^k\right)\right] && \text{(31a)} \\
&= \mathbb{E}_X w(X)\hat{g}(1,X) + \alpha\beta\mathbb{E}_X\left[\hat{g}(1,X)w(X)\hat{w}(X)^2\right] - \beta\mathbb{E}_X\left[w(X)\hat{w}(X)\Delta(\lambda,X)\right] \\
&\quad + \mathbb{E}_X\left[w(X)\hat{g}(1,X)\sum_{k=2}^{+\infty}\beta^k\left(\alpha\hat{w}(X)^2 - \frac{1}{\hat{g}(1,X)}\hat{w}(X)\Delta(\lambda,X)\right)^k\right] && \text{(31b)} \\
&\geqslant \mathbb{E}_X\left[w(X)\hat{g}(1,X)\right] + \alpha\beta\mathbb{E}_X\left[\hat{g}(1,X)w(X)\hat{w}(X)^2\right] - C_0\beta^3, && \text{(31c)}
\end{aligned}$$

where Equation (31a) uses Taylor expansion which holds since

$$\left|\beta\hat{w}(X)\left(\alpha\hat{w}(X) - \frac{1}{\hat{g}(1,X)}\Delta(\lambda,X)\right)\right| \leqslant \frac{1}{4}c \cdot \left(1 + \frac{1}{c}\right) \leqslant \frac{1}{2}$$

by Equation (30), Equation (31b) follows from directly expanding Equation (31a), and Equation (31c) holds with $C_0 = 2c^{-2}\|w\|_{P_X,\infty}^4$ where we use the fact that $\mathbb{E}_X\left[w(X)\hat{w}(X)\Delta(\lambda,X)\right] = 0$ (by Proposition 1) and that for any $|t| \leqslant 1/2$, $\sum_{k=2}^{\infty} t^k \geqslant \sum_{k=3}^{\infty} t^k = t^3/(1-t)$ (applied for $t := \beta\left(\alpha\hat{w}(X)^2 - \frac{1}{\hat{g}(1,X)}\hat{w}(X)\Delta(\lambda,X)\right)$, which also satisfies that $t^3 \geqslant -\beta^3 |0w|0_{P_X,\infty}^3/\hat{g}(1,X)^3$). Moreover, Equation (30) and $f_n > e_n'$ together imply that $C_0\beta^3 \leqslant C_0\beta^2 c\alpha \leqslant \frac{1}{2}c\alpha\beta\mathbb{E}_X\left[\hat{g}(1,X)w(X)\hat{w}(X)^2\right]$, so Equation (29) immediately follows from Equation (31), concluding the proof. □

We are now ready to prove Theorem 4 in the case when $f_n > e_n$. We choose $M \geqslant \max\{n, \frac{32C}{c^4\delta}n^2\}$ and $\mathcal{P} = \{\hat{P}\} \cup \left\{Q_\lambda : \lambda \in \{0,1\}^{M/2}\right\}$, $P = \hat{P}$, π be the discrete uniform distribution on $\left\{Q_\lambda : \lambda \in \{0,1\}^{M/2}\right\}$, $s = \frac{1}{4}\alpha\beta\mathbb{E}_X\left[\hat{g}(1,X)w(X)\hat{w}(X)^2\right]$ in the context of Theorem 7. Then all the listed conditions are satisfied for the WATE functional

$$T(P) = \theta^{WATE}(P) = \mathbb{E}_P\left[w(x)\left(g(1,X) - g(0,X)\right)\right].$$

Therefore, by Theorem 7, we obtain a lower bound

$$
\begin{aligned}
&\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathcal{Q}_{P,1-\gamma}\left(\left|\hat{\theta}\left(\{(X_i, D_i, Y_i)\}_{i=1}^N\right) - \theta^{WATE}\right|^2\right) \\
&= \Omega\left(\alpha\beta\mathbb{E}_X\left[\hat{g}(1,X)w(X)\hat{w}(X)^2\right]\right) \\
&= \Omega\left(\sqrt{e_n' f_n} \cdot \frac{\mathbb{E}_X\left[\hat{g}(1,X)w(X)\hat{w}(X)^2\right]}{\|\hat{w}(x)\|_{P,2}^2}\right) \\
&= \Omega\left(\sqrt{e_n' f_n} \cdot \frac{\mathbb{E}_X\left[w(X)^3 \mathbb{1}\left(w(X) \geqslant \frac{1}{2}\|w\|_{P_X,\infty}\right)\right]}{\mathbb{E}_X\left[w(X)^2 \mathbb{1}\left(w(X) \geqslant \frac{1}{2}\|w\|_{P_X,\infty}\right)\right]}\right) \\
&= \Omega\left(\|w\|_{P_X,\infty} \cdot \sqrt{e_n' f_n}\right).
\end{aligned}
$$

4.6 Proof of the lower bound $e_n f_n \|w\|_{P_X,\infty}^2$

Combining the derivations in Section 4.4 and 4.5, we have shown that

$$
\mathfrak{M}_n^{WATE}\left(\mathcal{F}_{e_n, e_n', f_n}\right) = \Omega\left(e_n' \cdot f_n \cdot \|w\|_{P_X,\infty}^2\right).
$$

In this section, we illustrate how the lower bound $\Omega\left(e_n f_n \|w\|_{P_X,\infty}^2\right)$ can be derived in a completely symmetric fashion. Parallel to the proofs in Section 4.4 and 4.5, we also consider two cases: $e_n \geqslant f_n$ and $e_n < f_n$.

In the first case, we define

$$
\begin{aligned}
g_\lambda(0,x) &= \frac{1 - \hat{m}(x)}{1 - m_\lambda(x)}\left[\hat{g}(0,x) - \alpha\hat{w}(x)\Delta(\lambda,x)\right] \\
m_\lambda(x) &= \hat{m}(x) + (1 - \hat{m}(x))\frac{\beta}{\hat{g}(0,x)}\hat{w}(x)\Delta(\lambda,x) \\
g_\lambda(1,x) &= \hat{g}(1,x).
\end{aligned}
\tag{32}
$$

In the second case, we define

$$
\begin{aligned}
g_\lambda(0,x) &= \frac{\hat{g}(0,x)}{1 + \frac{\beta}{\hat{g}(0,x)}\hat{w}(x)\Delta(\lambda,x) - \alpha\beta\hat{w}(x)^2} \\
g_\lambda(1,x) &= \hat{g}(1,x) \\
m_\lambda(x) &= 1 - \frac{\hat{g}(0,x)}{g_\lambda(0,x)}(1 - \hat{m}(x))\left(1 - \alpha\hat{g}(0,x)\hat{w}(x)\Delta(\lambda,x)\right)
\end{aligned}
\tag{33}
$$

Then we have the following result.

Lemma 10. *Let Q_λ be the joint distribution of (X, D, Y) induced by g_λ and m_λ and μ be the uniform distribution on $[0,1]^K \times \{0,1\} \times \{0,1\}$. Define $q_\lambda = \frac{dQ_\lambda}{d\mu}$. Then $\int q_\lambda d\pi(\lambda) = \hat{p}$. Moreover, there exists* *constants $c_\alpha, c_\beta > 0$, such that by choosing*

$$(\alpha, \beta) = \begin{cases} \left(c_\alpha \frac{\sqrt{e_n}}{\|\hat{w}(X)\|_{P_X,2}}, c_\beta \frac{\sqrt{f_n}}{\|\hat{w}(X)\|_{P_X,2}}\right) & \text{for the first case;} \\ \left(c_\alpha \frac{\sqrt{f_n}}{\|\hat{w}(X)\|_{P_X,2}}, c_\beta \frac{\sqrt{e_n}}{\|\hat{w}(X)\|_{P_X,2}}\right) & \text{for the second case,} \end{cases}$$

the following inequalities hold for sufficiently large n:

$$\|\hat{m}(X) - m_\lambda(X)\|_{P_X,2}^2 \leqslant f_n, \tag{34a}$$

$$\|\hat{g}(0,X) - g_\lambda(0,X)\|_{P_X,2}^2 \leqslant e_n, \tag{34b}$$

$$\mathbb{E}_X\left[w(X)g_\lambda(0,X)\right] \leqslant \mathbb{E}\left[w(X)\hat{g}(0,X)\right] - \Omega\left(\sqrt{e_n f_n}\|w\|_{P_X,\infty}\right). \tag{34c}$$

The proof of Lemma 10 follows the exactly same route as the proofs in Section 4.4 and 4.5, so we do not repeat it here. Finally, we can directly apply Theorem 7 to obtain the lower bound $\Omega\left(e_n f_n \|w\|_{P_X,\infty}^2\right)$.

5. Proof of Theorem 5

In this section, we give the detailed proof of our main result, Theorem 5, for the lower bound of estimating ATTE. The idea of the proof is similar to that of Theorem 4, but additional effort needs to be made to guarantee that the separation condition (19) holds.

Let P_X be the uniform distribution on $\mathrm{supp}(X) = [0,1]^K$, and $[0,1]^K$ be partitioned into M cubes $B_1, B_2, \cdots, B_M$, each with volume $\frac{1}{M}$. Let $\lambda_i, i = 1, 2, \cdots, \frac{M}{2}$ be i.i.d. variables taking values $+1$ and -1 both with probability 0.5.

Define

$$\theta_{ML}^{ATTE} = \left(\mathbb{E}_X\left[\hat{m}(X)\right]\right)^{-1} \mathbb{E}_X\left[\hat{m}(X)\left(\hat{g}(1,X) - \hat{g}(0,X)\right)\right]$$

and let p_X be the uniform distribution on $[0,1]^K$. We first prove the following lemma:

Lemma 11. *There exist constants $C_u, c_u > 0$ that only depend on $\hat{m}$ and $\hat{g}$, such that for all sufficiently large integer M, there exists a function $u : [0,1]^K \to \mathbb{R}_{\geqslant 0}$ satisfying $\|u\|_\infty \leqslant C_u$ and a partition $[0,1]^K = \cup_{j=1}^M B_j$ into Lebesgue-measurable sets B_j each with measure $\frac{1}{M}$, such that*

$$\mathbb{E}_X\left[u(X)\left(\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE}\right)\Delta(\lambda, X)\right] = 0, \quad \forall \lambda \in \{-1,+1\}^{M/2} \tag{35}$$

and

$$\mathbb{E}_X\left[\frac{u(X)}{\hat{m}(X)\left(1 - \hat{m}(X)\right)}\right] \geqslant c_u, \tag{36}$$

where we recall that

$$\Delta(\lambda, x) := \sum_{j=1}^{M/2} \lambda_j \left(\mathbb{1}\left\{x \in B_{2j-1}\right\} - \mathbb{1}\left\{x \in B_{2j}\right\}\right).$$

Proof. Let $\alpha = \mathbb{P}\left[\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE} = 0\right]$. If $\alpha = 1$, then we can simply choose $u = 1$ and $c_u = 1$. Thus we can assume that $\alpha < 1$. In this case either

$$\mathbb{P}\left[\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE} > 0\right] \geqslant \frac{1-\alpha}{2}$$

or

$$\mathbb{P}\left[\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE} < 0\right] \geqslant \frac{1-\alpha}{2}.$$

We proceed by assuming that the former holds; the case when the latter holds can be handled in exactly the same way.

Define the event $\mathcal{E}_\delta = \left\{\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE} > \delta\right\}$, then

$$\lim_{\delta\to 0} \mathbb{P}\left[\mathcal{E}_\delta\right] \geqslant \frac{1-\alpha}{2},$$

so there exists $\delta_0 > 0$ such that

$$\mathbb{P}\left[\mathcal{E}_{\delta_0}\right] \geqslant \frac{1-\alpha}{3}.$$

Let $M_\alpha = 2\lfloor \frac{1-\alpha}{6} M \rfloor$ and let $B_j, 1 \leqslant j \leqslant M$ be chosen in a way such that $B_j, 1 \leqslant j \leqslant M_\alpha$ are (disjoint) measurable subsets of $\mathcal{E}_{\delta_0}$ with measure $\frac{1}{M}$; the remaining B_j's can be chosen arbitrarily. Then we define

$$u(x) = \begin{cases} 1 & x \in B_{2j-1}, 1 \leqslant j \leqslant M_\alpha/2 \\ \dfrac{\mathbb{E}_X\left[\left(\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE}\right) \mathbb{1}\left\{X \in B_{2j-1}\right\}\right]}{\mathbb{E}_X\left[\left(\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE}\right) \mathbb{1}\left\{X \in B_{2j}\right\}\right]} & x \in B_{2j}, 1 \leqslant j \leqslant M_\alpha/2 \\ 0 & \text{otherwise.} \end{cases}$$

Specifically, $u(x)$ is constant in each B_j. Moreover, note that the denominator in the second case is bounded away from zero, since these regions are subsets of $\mathcal{E}_{\delta_0}$. First, it is easy to see that this choice of u guarantees that

$$\mathbb{E}_X\left[u(X)\left(\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE}\right)\left(\mathbb{1}\left\{X \in B_{2j-1}\right\} - \mathbb{1}\left\{X \in B_{2j}\right\}\right)\right] = 0$$

for all j, so that Equation (35) holds.

Second, let

$$C_u = \delta_0^{-1}\left(2 + \left|\theta_{ML}^{ATTE}\right|\right).$$

Our choice of B_j implies that for $1 \leqslant j \leqslant M_\alpha/2$, we have

$$\mathbb{E}_X\left[\left(\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE}\right) \mathbb{1}\left\{X \in B_{2j}\right\}\right] \geqslant \delta_0 \cdot \mathbb{P}\left[X \in B_{2j}\right] = \frac{\delta_0}{M}$$

and

$$\begin{aligned} \mathbb{E}_X\left[\left(\hat{g}(1,X) - \hat{g}(0,X) - \theta_{ML}^{ATTE}\right) \mathbb{1}\left\{X \in B_{2j-1}\right\}\right] &\leqslant \left(2\sup_{d,x} \hat{g}(d,x) + \left|\theta_{ML}^{ATTE}\right|\right) P[X \in B_{2j-1}] \\ &\leqslant \left(2 + \left|\theta_{ML}^{ATTE}\right|\right)\frac{1}{M}. \end{aligned}$$

As a consequence, we have

$$u(x) \leqslant C_u, \quad \forall x \in [0,1]^K.$$

Finally, since $\mathbb{P}\left[u(X)=1\right] = \frac{M_\alpha}{2M}$ and $u(x) \geqslant 0$ for all x, we can deduce that

$$\mathbb{E}_X\left[\frac{u(X)}{\hat{m}(X)\left(1-\hat{m}(X)\right)}\right] \geqslant \frac{M_\alpha}{2M} \geqslant 0.1(1-\alpha).$$

Hence, the $u(x)$ that we choose satisfies all the required conditions, concluding the proof. □

Returning to our proof of Theorem 5, let $u(x)$ and $\Delta(\lambda, x)$ be the function chosen in Lemma 11 and let

$$v(x) = \frac{1-\hat{m}(x)}{\hat{m}(x)}. \tag{37}$$

We define

$$\begin{aligned} g_\lambda(0,x) &= \hat{g}(0,x) + \alpha\frac{v(x)}{1-m_\lambda(x)}\Delta(\lambda,x) \\ g_\lambda(1,x) &= \hat{g}(1,x) \\ m_\lambda(x) &= \hat{m}(x) - \beta u(x)\Delta(\lambda,x). \end{aligned} \tag{38}$$

where α, β are constants that will be specified later. Then one can easily derive the following results:

Proposition 4. *We have*

$$\begin{aligned} \mathbb{E}_\lambda\left[m_\lambda(x)\right] &= \hat{m}(x) \\ \mathbb{E}_\lambda\left[\left(1-m_\lambda(x)\right)g_\lambda(0,x)\right] &= \hat{g}(0,x)\left(1-\hat{m}(x)\right), \end{aligned}$$

Proof. By Proposition 1, we have

$$\begin{aligned} \mathbb{E}_\lambda\left[m_\lambda(x)\right] &= \hat{m}(x) - \beta u(x)\mathbb{E}_\lambda\left[\Delta(\lambda,x)\right] = \hat{m}(x) \\ \mathbb{E}_\lambda\left[\left(1-m_\lambda(x)\right)g_\lambda(0,x)\right] &= \hat{g}(0,x)\mathbb{E}_\lambda\left[1-m_\lambda(X)\right] + \alpha v(x)\mathbb{E}_\lambda\left[\Delta(\lambda,x)\right] \\ &= \hat{g}(0,x)\left(1-\hat{m}(x)\right). \end{aligned}$$

□

As in Section 4, we can bound the L_2 distance between g_λ, m_λ and $\hat{g}, \hat{m}$ respectively.

Lemma 12. *Suppose that $\alpha \leqslant 1, \beta \leqslant \frac{1}{4}C_u^{-1}$ (where C_u is defined in Lemma 11), then the following holds for all $0 < r \leqslant +\infty$:*

$$\left\|g_\lambda(0,X)-\hat{g}(0,X)\right\|_{P_X,r} \leqslant 2c^{-1}\alpha, \quad \left\|m_\lambda(X)-\hat{m}(X)\right\|_{P_X,r} \leqslant c^{-1}\beta.$$

Remark 4. *Due to the difference in construction, the bounds in the lemma above are in the forms of $\mathcal{O}(\alpha)$ and $\mathcal{O}(\beta)$ rather than $\mathcal{O}(\alpha+\beta)$ and $\mathcal{O}(\beta)$ that we encountered in the case of the WATE. This is the reason why we don't need to consider the two cases $e_n \geqslant f_n$ and $e_n < f_n$ separately for ATTE.*

Let Q_λ be the joint distribution of (X, D, Y) induced by g_λ and m_λ and μ be the uniform distribution on $[0,1]^K \times \{0,1\} \times \{0,1\}$. Define $q_\lambda = \frac{\mathrm{d}Q_\lambda}{\mathrm{d}\mu}$. Similarly, let $\hat{P}$ be the joint distribution of (X, D, Y) induced by $\hat{g}$ and $\hat{m}$, and $\hat{p} = \frac{\mathrm{d}\hat{P}}{\mathrm{d}\mu}$. Using exactly the same arguments as we did in Lemma 3 and 4, one can prove the following lemmas.

Lemma 13. *Let $Q = \int Q_\lambda d\pi(\lambda)$ and $q = \frac{dQ}{d\mu} = \int q_\lambda d\pi(\lambda)$, then $\hat{p} = q$.*

Lemma 14. *For any $\delta > 0$, as long as $M \geqslant \max\{n, \frac{32C}{c^4\delta} n^2\}$ where c is the constant introduced in Assumption 3 and C is the constant implied by Theorem 6 for $A = 4c^{-2}$, we have*

$$H^2\left(\hat{P}^{\otimes n}, \int Q_\lambda^{\otimes n} d\pi(\lambda)\right) \leqslant \delta.$$

Lemma 15. *Let*

$$\alpha = \frac{c}{4}\sqrt{e_n}, \quad \beta = \frac{1}{4}\min\{c, c_u\}\sqrt{f_n},$$

then for sufficiently large n, we have $(m_\lambda, g_\lambda) \in \mathcal{F}_{e_n, e_n', f_n}$ and

$$\theta_\lambda^{ATTE} \leqslant \theta_{ML}^{ATTE} - \frac{1}{2}c_u\alpha\beta, \quad \forall \lambda \in \{0,1\}^{M/2} \tag{40}$$

Proof. Since $e_n, f_n = o(1)(n \to +\infty)$, we have $\alpha \leqslant \frac{1}{4}c$ and $\beta \leqslant \frac{1}{4}c_u c^3 C_u^{-2}$ for sufficiently large n. In the remaining part of the proof, we assume that this inequality holds.

First, by Lemma 12 it is easy to see that

$$\begin{aligned}\|\hat{m}(X) - m_\lambda(X)\|_{P_X,2} &\leqslant 2c^{-1}\beta \leqslant \sqrt{f_n} \\ \|\hat{g}(1,X) - g_\lambda(1,X)\|_{P_X,2} &\leqslant 2c^{-1}\alpha \leqslant \sqrt{e_n}\end{aligned}$$

and $0 \leqslant m_\lambda, g_\lambda \leqslant 1$, so that $(m_\lambda, g_\lambda) \in \mathcal{F}_{e_n, e_n', f_n}$.

It remains to prove Equation (40). For fixed λ, we have

$$\begin{aligned}
\theta_\lambda^{ATTE} &:= \mathbb{E}_X\left[g_\lambda(1,X) - g_\lambda(0,X) \mid D = 1\right] \\
&= \mathbb{E}_X\left[\left(g_\lambda(1,X) - g_\lambda(0,X)\right)\frac{m_\lambda(X)}{\mathbb{P}_\lambda[D=1]}\right] \\
&= \frac{\mathbb{E}_X\left[\left(\hat{g}(1,X) - \hat{g}(0,X)\right)m_\lambda(X) - \frac{\alpha v(x) m_\lambda(X)}{1-m_\lambda(x)}\Delta(\lambda,x)\right]}{\mathbb{E}_X\left[m_\lambda(X)\right]} \\
&= \frac{\mathbb{E}_X\left[\left(\hat{g}(1,X) - \hat{g}(0,X)\right)\left(\hat{m}(X) - \beta u(X)\Delta(\lambda,X)\right) - \frac{\alpha v(x) m_\lambda(X)}{1-m_\lambda(x)}\Delta(\lambda,x)\right]}{\mathbb{E}_X\left[m_\lambda(X)\right]} \\
&= \frac{\mathbb{E}_X\left[\left(\hat{g}(1,X) - \hat{g}(0,X)\right)\hat{m}(X)\right] - \beta\mathbb{E}_X\left[u(X)\left(\hat{g}(1,X) - \hat{g}(0,X)\right)\Delta(\lambda,X)\right]}{\mathbb{E}_X\left[\hat{m}(X)\right] - \beta\mathbb{E}_X\left[u(X)\Delta(\lambda,X)\right]} \\
&\quad - \left(\mathbb{E}_X\left[m_\lambda(X)\right]\right)^{-1}\alpha\mathbb{E}_X\left[\left(1 + \frac{m_\lambda(X) - \hat{m}(X)}{\hat{m}(X)\left(1 - m_\lambda(X)\right)}\right)\Delta(\lambda,X)\right] \\
&=: A - B
\end{aligned} \tag{41}$$

where the third line follows from the fact that $g_\lambda(0,x) - \hat{g}(0,x) = \frac{\alpha v(x)}{1-m_\lambda(x)}\Delta(\lambda,x)$ and the fourth line from the fact that $\hat{m}(X) - \beta u(X)\Delta(\lambda,X) = m_\lambda(x)$, according to Equation (38).

Recall that $\theta_{ML}^{ATTE} = \frac{\mathbb{E}_X\left[\left(\hat{g}(1,X) - \hat{g}(0,X)\right)\hat{m}(X)\right]}{\mathbb{E}_X\left[\hat{m}(X)\right]}$ and

$$\mathbb{E}_X\left[u(X)\left(\hat{g}(1,X) - \hat{g}(0,X)\right)\Delta(\lambda,X)\right] - \theta_{ML}^{ATTE}\mathbb{E}_X\left[u(X)\Delta(\lambda,X)\right] = 0$$

by our choice of u in Lemma 11, so the first term A in Equation (41) equals θ_{ML}^{ATTE}, since:

$$\begin{aligned} A &= \frac{\theta_{ML}^{ATTE}\mathbb{E}_X[\hat{m}(X)] - \beta\mathbb{E}_X\left[u(X)\left(\hat{g}(1,X)-\hat{g}(0,X)\right)\Delta(\lambda,X)\right]}{\mathbb{E}_X\left[\hat{m}(X)\right] - \beta\mathbb{E}_X\left[u(X)\Delta(\lambda,X)\right]} \\ &= \frac{\theta_{ML}^{ATTE}\mathbb{E}_X[\hat{m}(X)] - \beta\theta_{ML}^{ATTE}\mathbb{E}_X\left[u(X)\Delta(\lambda,X)\right]}{\mathbb{E}_X\left[\hat{m}(X)\right] - \beta\mathbb{E}_X\left[u(X)\Delta(\lambda,X)\right]} = \theta_{ML}^{ATTE} \end{aligned}$$

The second term can be further simplified as follows:

$$\begin{aligned} B &= \left(\mathbb{E}_X\left[m_\lambda(X)\right]\right)^{-1}\alpha\mathbb{E}_X\left[\frac{m_\lambda(X)-\hat{m}(X)}{\hat{m}(X)\left(1-m_\lambda(X)\right)}\Delta(\lambda,X)\right] \\ &= -\alpha\beta\left(\mathbb{E}_X\left[m_\lambda(X)\right]\right)^{-1}\mathbb{E}_X\left[\frac{u(X)}{\hat{m}(X)\left(1-m_\lambda(X)\right)}\Delta(\lambda,X)^2\right] \\ &\leqslant -\alpha\beta\mathbb{E}_X\left[\frac{u(X)}{\hat{m}(X)(1-m_\lambda(X))}\right] \quad (42a) \\ &= -\alpha\beta\mathbb{E}_X\left[\frac{u(X)}{\hat{m}(X)(1-\hat{m}(X))}\right] - \alpha\beta\mathbb{E}_X\left[\frac{u(X)(m_\lambda(X)-\hat{m}(X))}{\hat{m}(X)(1-\hat{m}(X))(1-m_\lambda(X))}\right] \\ &= -\alpha\beta\mathbb{E}_X\left[\frac{u(X)}{\hat{m}(X)(1-\hat{m}(X))}\right] + \alpha\beta^2\mathbb{E}_X\left[\frac{u(X)^2\Delta(\lambda,X)}{\hat{m}(X)(1-\hat{m}(X))(1-m_\lambda(X))}\right] \\ &\leqslant -c_u\alpha\beta + 2c^{-3}C_u^2\alpha\beta^2 \leqslant -\frac{1}{2}c_u\alpha\beta \quad (42b) \end{aligned}$$

where Equation (42a) follows from $0 < m_\lambda(X) < 1$ and $u(X) \geqslant 0$, and Equation (42b) follows from

$$\left|m_\lambda(x) - \hat{m}(x)\right| \leqslant \beta C_u \leqslant \frac{1}{2}c \quad \Rightarrow \quad \frac{1}{1-m_\lambda(x)} \leqslant 2c^{-1}$$

and $\beta \leqslant \frac{1}{4}c_u c^3 C_u^{-2}$. Hence, for all $\lambda \in \{-1,+1\}^{M/2}$ we have

$$\theta_\lambda^{ATTE} \leqslant \theta_{ML}^{ATTE} - \frac{1}{2}c_u\alpha\beta.$$

□

We are now ready to prove Theorem 5. We choose M sufficiently large according to Lemma 14, $\mathcal{P} = \{\hat{P}\}\cup\left\{Q_\lambda : \lambda \in \{0,1\}^{M/2}\right\}$, $P = \hat{P}$, π be the discrete uniform distribution on $\left\{Q_\lambda : \lambda \in \{0,1\}^{M/2}\right\}$, $s = \frac{1}{2}c_u\alpha\beta = \Omega(\sqrt{e_n f_n})$ in the context of Theorem 7. Then all the listed conditions are satisfied for the ATE functional

$$T(P) = -\theta^{ATTE}(P) = -\mathbb{E}_P\left[g(1,X) - g(0,X) \mid D = 1\right].$$

Therefore, by Theorem 7, we obtain a lower bound

$$\inf_{\hat{\theta}}\sup_{P\in\mathcal{P}} \mathcal{Q}_{P,1-\gamma}\left(\left|\hat{\theta}\left(\{(X_i,D_i,Y_i)\}_{i=1}^N\right) - \theta^{WATE}\right|^2\right) = \Omega(\alpha^2\beta^2) = \Omega\left(e_n f_n\right).$$

6. Conclusion

We investigate the statistical limit of treatment effect estimation in the structural-agnostic regime, which is an appropriate lower bound technique when one wants to only consider estimation strategies that use generic black-box machine learning estimators for the various nuisance functions involved in the estimation of treatment effects. We establish the minimax optimality of the celebrated and widely used in practice doubly robust learning strategies via reducing the estimation problem to a hypothesis testing problem, and lower-bound its error by non-standard constructions of the fuzzy hypotheses. Our results show that these estimators are optimal, in the structure agnostic sense, even in the slow rate regimes, where the implied rate for the functional of interest is slower than root-n. Hence, any improvements upon these estimation strategies need to incroporate elements of the structure of the nuisance functions and cannot simply invoke generic adaptive regression approaches as black-box sub-processes. While the focus of this paper is on treatment effect estimation, we believe that our techniques can be extended to address structure agnostic minimax lower bounds of more general functional estimation problems.

References

Abadie, Alberto, and Guido W Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *econometrica* 74 (1): 235–267.

Arias-Castro, Ery, Bruno Pelletier, and Venkatesh Saligrama. 2018. Remember the curse of dimensionality: the case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics* 30 (2): 448–471.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. Generalized random forests. *The Annals of Statistics* 47 (2): 1148.

Balakrishnan, S, and L Wasserman. 2019. Hypothesis testing for densities and high-dimensional multinomials: sharp local minimax rates. *Annals of Statistics* 47 (4): 1893–1927.

Balakrishnan, Sivaraman, Edward H Kennedy, and Larry Wasserman. 2023. The fundamental limits of structure-agnostic functional estimation. *arXiv preprint arXiv:2305.04116.*

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80 (6): 2369–2429.

Belloni, Alexandre, and Victor Chernozhukov. 2011. L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* 39 (1): 82.

———. 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli* 19 (2): 521–547.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81 (2): 608–650.

Belloni, Alexandre, Victor Chernozhukov, and Kengo Kato. 2015. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika* 102 (1): 77–94.

Belloni, Alexandre, Victor Chernozhukov, and Lie Wang. 2014. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics* 42 (2): 757.

Bickel, Peter J, Ya'acov Ritov, and Alexandre B Tsybakov. 2009. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics,* 1705–1732.

Bickel, Peter J, and Yaacov Ritov. 1988. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A,* 381–393.

Bickel, PJ. 1982. On adaptive estimation. *The Annals of Statistics* 10 (3): 647.

Birgé, Lucien, and Pascal Massart. 1995. Estimation of integral functionals of a density. *The Annals of Statistics* 23 (1): 11–29.

Chen, Qizhao, Vasilis Syrgkanis, and Morgane Austern. 2022. Debiased machine learning without sample-splitting for stable estimators. *Advances in Neural Information Processing Systems* 35:3096–3109.

Chen, Xiaohong, and Halbert White. 1999. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory* 45 (2): 682–691.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. Double/debiased/neyman machine learning of treatment effects. *American Economic Review* 107 (5): 261–265.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters: double/debiased machine learning. *The Econometrics Journal* 21 (1).

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. 2022. Locally robust semiparametric estimation. *Econometrica* 90 (4): 1501–1535.

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh. 2022. Automatic debiased machine learning of causal and structural effects. *Econometrica* 90 (3): 967–1027.

———. 2023. A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika* 110 (1): 257–264.

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

Farrell, Max H, Tengyuan Liang, and Sanjog Misra. 2021. Deep neural networks for estimation and inference. *Econometrica* 89 (1): 181–213.

Foster, Dylan J, and Vasilis Syrgkanis. 2023. Orthogonal statistical learning. *The Annals of Statistics* 51 (3): 879–908.

Geer, Sara van de, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42 (3).

Heckman, James J, Hidehiko Ichimura, and Petra Todd. 1998. Matching as an econometric evaluation estimator. *The review of economic studies* 65 (2): 261–294.

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

Imbens, Guido W. 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and statistics* 86 (1): 4–29.

Imbens, GW, W Newey, and G Ridder. 2003. *Mean-squared-error calculations for average treatment effects. department of economics, uc berkeley.*

Ingster, Yu I. 1994. Minimax detection of a signal in ℓ_p metrics. *Journal of Mathematical Sciences* 68:503–515.

Kennedy, Edward H, Sivaraman Balakrishnan, James M Robins, and Larry Wasserman. 2022. Minimax rates for heterogeneous causal effect estimation. *arXiv preprint arXiv:2203.00837.*

Little, Roderick J, and Donald B Rubin. 2000. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health* 21 (1): 121–145.

Liu, Lin, Rajarshi Mukherjee, Whitney K Newey, and James M Robins. 2017. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577.*

Mayer, Alexander K. 2011. Does education increase political participation? *The Journal of Politics* 73 (3): 633–645.

Oreopoulos, Philip. 2006. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review* 96 (1): 152–175.

Robins, James, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. 2008. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of david a. freedman,* 2:335–422. Institute of Mathematical Statistics.

Robins, James, Eric Tchetgen Tchetgen, Lingling Li, and Aad van der Vaart. 2009. Semiparametric minimax rates. *Electronic journal of statistics* 3:1305.

Robins, James M, Lingling Li, and Rajarshi Mukherjee. 2017. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics* 45 (5): 1951–1987.

Robins, James M, and Andrea Rotnitzky. 1995. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90 (429): 122–129.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89 (427): 846–866.

———. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association* 90 (429): 106–121.

Rosenbaum, Paul R. 1989. Optimal matching for observational studies. *Journal of the American Statistical Association* 84 (408): 1024–1032.

Rosenbaum, Paul R, and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1): 41–55.

Rotnitzky, Andrea, Ezequiel Smucler, and James M Robins. 2021. Characterization of parameters with a mixed bias property. *Biometrika* 108 (1): 231–238.

Schick, Anton. 1986. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics* 14 (3): 1139–1151.

Schmidt-Hieber, Anselm Johannes. 2020. Nonparametric regression using deep neural networks with relu activation function. *Annals of statistics* 48 (4): 1875–1897.

Syrgkanis, Vasilis, and Manolis Zampetakis. 2020. Estimation and inference with trees and forests in high dimensions. In *Conference on learning theory,* 3453–3454. PMLR.

Tao, Yebin, and Haoda Fu. 2019. Doubly robust estimation of the weighted average treatment effect for a target population. *Statistics in medicine* 38 (3): 315–325.

Tsybakov, Alexandre B. 2008. *Introduction to nonparametric estimation.* Springer Science & Business Media.

Vaart, Aad van der. 2014. Higher order tangent spaces and influence functions. *Statistical science* 29 (4): 679–686.

Wager, Stefan, and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113 (523): 1228–1242.

Wager, Stefan, and Guenther Walther. 2015. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388.*

Wood, Lesley, Matthias Egger, Lise Lotte Gluud, Kenneth F Schulz, Peter Jüni, Douglas G Altman, Christian Gluud, Richard M Martin, Anthony JG Wood, and Jonathan AC Sterne. 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *bmj* 336 (7644): 601–605.

Zou, Hui, and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (2): 301–320.

Appendices

In the appendix we provide the proofs of Theorem 2, Theorem 3 and the $\mathcal{O}(1/n)$ lower bound in Theorem 4 and Theorem 5. The proofs of these results are all relatively standard and are implicit in prior works.

A. Proof of Theorem 2

We define

$$\bar{\theta}^{WATE} = \mathbb{E}w(X)\left[\hat{g}(1,X) - \hat{g}(0,X) + \left(\frac{D}{\hat{m}(X)} - \frac{1-D}{1-\hat{m}(X)}\right)(Y - \hat{g}(D,X))\right],$$

then $\mathbb{E}\hat{\theta}^{WATE} = \bar{\theta}^{WATE}$, which implies that

$$\mathbb{E}\left(\hat{\theta}^{WATE} - \bar{\theta}^{WATE}\right)^2 \leqslant \frac{1}{n}\mathrm{Var}\left(\hat{\theta}^{WATE}\right) \lesssim \frac{1}{n}\|w\|_{P_X,2}^2.$$

On the other hand,

$$\begin{aligned}
&\left|\hat{\theta}^{WATE} - \bar{\theta}^{WATE}\right| \\
&\leqslant \mathbb{E}w(X)\left|1 - \frac{m_0(X)}{\hat{m}(X)}\right|\left|g_0(1,X) - \hat{g}(1,X)\right| + \mathbb{E}w(X)\left|1 - \frac{1-m_0(X)}{1-\hat{m}(X)}\right|\left|g_0(0,X) - \hat{g}(0,X)\right| \\
&\leqslant \|w\|_\infty \cdot \left(\mathbb{E}\left|1 - \frac{m_0(X)}{\hat{m}(X)}\right|\left|g_0(1,X) - \hat{g}(1,X)\right| + \mathbb{E}\left|1 - \frac{1-m_0(X)}{1-\hat{m}(X)}\right|\left|g_0(0,X) - \hat{g}(0,X)\right|\right) \\
&\leqslant c^{-1}\|w\|_\infty\|m_0(X) - \hat{m}(X)\|_{P_X,2} \cdot \left(\left|\|g_0(0,X) - \hat{g}(0,X)\right|\|_{P_X,2} + \left|\|g_0(1,X) - \hat{g}(1,X)\right|\|_{P_X,2}\right) \\
&= \mathcal{O}\left(\|w\|_\infty\sqrt{\max\{r_n, r_n'\}s_n}\right).
\end{aligned}$$

Combining the above inequalities, we have

$$\mathbb{E}\left(\hat{\theta}^{WATE} - \theta^{WATE}\right)^2 = \mathcal{O}\left(\max\{r_n, r_n'\} \cdot s_n \cdot \|w\|_\infty^2 + \frac{1}{n}\right)$$

and the desired high-probability bound follows directly from Markov's inequality.

B. Proof of Theorem 3

Since $\mathbb{E}[D] = \mathbb{E}_X[m_0(X)]$ and $D_i, i = 1, 2, \cdots, n$ are i.i.d. Bernoulli variables, by central limit theorem there exists constant $\tilde{C}_{\delta,1} > 0$ such that

$$\left|\frac{1}{n}\sum_{i=1}^n D_i - \mathbb{E}[D_1]\right| \leqslant C_{\delta,1}\sqrt{\frac{\mathrm{Var}(D_1)}{n}} \quad \text{with probability } \geqslant 1 - \frac{1}{2}\delta. \tag{43}$$

Hence with probability $\geqslant 1-\delta$, we have

$$
\begin{aligned}
&\left|\hat{\theta}^{ATTE}-\theta^{ATTE}\right| \\
\lesssim\ & \frac{1}{\sqrt{n}} + (\mathbb{E}[m_0(X)])^{-1}\left|(\mathbb{E}_n-\mathbb{E})\left(D\left(Y-\hat{g}(0,X)\right)-\frac{\hat{m}(X)}{1-\hat{m}(X)}(1-D)(Y-\hat{g}(0,X))\right)\right| \\
&+\mathbb{E}\Big| m_0(X)\left(g_0(1,X)-g_0(0,X)\right)-m_0(X)\left(g_0(1,X)-\hat{g}(0,X)\right) \\
&-\hat{m}(X)\frac{1-m_0(X)}{1-\hat{m}(X)}\left(g_0(0,X)-\hat{g}(0,X)\right)\Big| \\
\lesssim\ & \frac{1}{\sqrt{n}}+\mathbb{E}\left|\frac{(m_0(X)-\hat{m}(X))(g_0(0,X)-\hat{g}(0,X))}{1-\hat{m}(X)}\right| \\
\lesssim\ & \frac{1}{\sqrt{n}}+\sqrt{r_n s_n},
\end{aligned}
$$

where $\mathbb{E}_n$ denotes the empirical average in the second term of the second line, and this term is bounded by $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ with high probability. This concludes the proof.

C. Proof of the $\Omega(n^{-1})$ lower bound in Theorem 4

We define

$$
\begin{aligned}
g(0,x) &= \hat{g}(0,x) \\
g(1,x) &= \hat{g}(1,x)+\xi w(x) \\
m(x) &= \hat{m}(x)
\end{aligned}
\tag{44}
$$

where ξ is a constant that will be specified later.

Let Q be the joint distribution of (X,D,Y) induced by g and m defined above, then its density (w.r.t uniform measure) can be written as

$$
q(x,d,y)=m(x)^d(1-m(x))^{1-d}g(d,x)^y(1-g(d,x))^{1-y}.
$$

From Equation (44) one can deduce that

$$
\mathbb{E}_X\left[w(x)\left(g(1,x)-g(0,x)\right)\right]=\mathbb{E}_X\left[w(x)\left(\hat{g}(1,x)-\hat{g}(0,x)\right)\right]+\xi\|w\|_{P_X,2}^2
$$

and

$$
\left|q(x,d,y)-\hat{p}(x,d,y)\right|\leqslant \xi|w(x)|.
$$

Moreover, by assumption we know that $\hat{p}(x,d,y)\geqslant c^2$, so we have that

$$
H^2(\hat{P},Q)\lesssim \xi^2\|w\|_{P_X,2}^2.
$$

By choosing $\xi\lesssim\frac{1}{\sqrt{n}\|w\|_{P_X,2}}$, one can guarantee that

$$
H^2(\hat{P}^{\otimes n},Q^{\otimes n})\leqslant nH^2(\hat{P},Q)\leqslant\alpha,
$$

so that the lower bound immediately follows from Theorem 7.

D. Proof of the $\Omega(n^{-1})$ lower bound in Theorem 5

We consider the construction in Equation (44). For the ATTE, one can check that

$$\frac{\mathbb{E}_X\left[(g(1,X)-g(0,X))m(X)\right]}{\mathbb{E}_X[m(X)]} = \frac{\mathbb{E}_X\left[(\hat{g}(1,X)-\hat{g}(0,X)+\xi)\hat{m}(X)\right]}{\mathbb{E}_X[\hat{m}(X)]}$$
$$= \frac{\mathbb{E}_X\left[(\hat{g}(1,X)-\hat{g}(0,X))\hat{m}(X)\right]}{\mathbb{E}_X[\hat{m}(X)]} + \xi.$$

The lower bound then directly follows from repeating the remaining steps in Section C.