HHS Public Access

Author manuscript

Econom J. Author manuscript; available in PMC 2024 January 12.

Published in final edited form as:

Econom J. 2021 September ; 24(3): 559–588. doi:10.1093/ectj/utab019.

Double/debiased machine learning for logistic partially linear model

MOLEI LIU[†], **YI ZHANG**[‡], **DOUDOU ZHOU**[§]

[†]Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA.

[‡]Department of Statistics, Harvard University, One Oxford Street, Cambridge, MA 02138-2901, USA.

[§]Department of Statistics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA.

Summary:

We propose double/debiased machine learning approaches to infer a parametric component of a logistic partially linear model. Our framework is based on a Neyman orthogonal score equation consisting of two nuisance models for the nonparametric component of the logistic model and conditional mean of the exposure with the control group. To estimate the nuisance models, we separately consider the use of high dimensional (HD) sparse regression and (nonparametric) machine learning (ML) methods. In the HD case, we derive certain moment equations to calibrate the first order bias of the nuisance models, which preserves the model double robustness property. In the ML case, we handle the nonlinearity of the logit link through a novel and easy-to-implement 'full model refitting' procedure. We evaluate our methods through simulation and apply them in assessing the effect of the emergency contraceptive pill on early gestation and new births based on a 2008 policy reform in Chile.

Keywords

Logistic partially linear model; double machine learning; double robustness; regularized regression; calibration; C14

1. INTRODUCTION

Consider a logistic partially linear model. Let $\{(Y_i, A_i, \boldsymbol{X}_i): i = 1, 2, \ldots, n\}$ be independent and identically distributed samples of $Y \in \{0, 1\}$, $A \in \mathbb{R}$, and $\boldsymbol{X} \in \mathbb{R}^p$. Assume that

molei_liu@g.harvard.edu .

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication Package

Co-editor Victor Chernozhukov handled this manuscript.

$$\mathbb{P}(Y = 1 \mid A, \boldsymbol{X}) = \text{expit}\{\beta_0 A + r_0(\boldsymbol{X})\}, \tag{1.1}$$

where $\text{expit}(\cdot) = \text{logit}^{-1}(\cdot)$, $\text{logit}(a) = \log\{a/(1-a)\}$, and $r_0(\cdot)$ are unknown nuisance functions of $\boldsymbol{X}$. In an experimental or observational study with A taken as the exposure variable, Y being the binary response of interest and $\boldsymbol{X}$ representing the observed confounding variables, parameter β_0 is of particular interest as it measures the conditional effect of A on Y on the scale of the logarithmic odds ratio. As the most common and natural way to characterise the conditional model of a binary outcome against some exposure, model (1.1) has been extensively used in economics and policy science studies.

Our goal is to estimate and asymptotically infer β_0 at the rate $n^{-1/2}$. When $\boldsymbol{X}$ is a scalar and $r_0(\cdot)$ is smooth, classic semiparametric kernel or sieve regression work well for this purpose (see Severini and Staniswalis, 1994; Lin and Carroll, 2006). When $\boldsymbol{X}$ is of high dimensionality, these approaches can have poor performance due to the curse of dimensionality. Accordingly, it would be more desirable to estimate $r_0(\cdot)$ with modern high dimensional (HD) (parametric) or machine learning (ML) (nonparametric)[1] methods that are much more resistant to the growing dimensionality and complexity of $\boldsymbol{X}$. However, unlike the partially linear model scenario (see Chernozhukov, Chetverikov, et al., 2018; Dukes and Vansteelandt, 2020), robust and efficient inference of β_0 in (1.1) with HD or ML nuisance models has not yet been extensively studied.

In recent literature, Tan (2019a) proposed a simple and flexible doubly robust estimator to enhance the robustness to the potential misspecification of $r(\boldsymbol{x})$ specified as a fixed-dimensional parametric function: $r(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\gamma}$. The authors introduced a parametric model $m(\boldsymbol{x}) = g\left(\boldsymbol{x}^\top \boldsymbol{\alpha}\right)$ with a (known) link function $g(\cdot)$ for the conditional mean model $m_0(\boldsymbol{x}) = \mathbb{E}(A \mid Y = 0, \boldsymbol{X} = \boldsymbol{x})$ and proposed a doubly robust estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\phi}(\boldsymbol{X}_i)\left\{Y_i e^{-\beta A_i - \boldsymbol{X}_i^\top \hat{\gamma}} - (1 - Y_i)\right\}\left\{A_i - g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right)\right\} = 0, \tag{1.2}$$

where $\hat{\phi}(\boldsymbol{x})$ is an estimation of some scalar nuisance function $\phi(\boldsymbol{x})$ affecting the asymptotic efficiency of the estimator, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ are two fixed-dimensional nuisance model estimators. Estimator $\hat{\beta}$ solved from (1.2) is doubly robust in the sense that it is valid when either $r(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\gamma}$ is correctly specified for the logistic model nonparametric component, or when $m(\boldsymbol{x}) = g\left(\boldsymbol{x}^\top \boldsymbol{\alpha}\right)$ is correctly specified for the conditional mean model $m_0(\boldsymbol{x}) = \mathbb{E}(A \mid Y = 0, \boldsymbol{X} = \boldsymbol{x})$. Prior to this, the doubly robust semiparametric estimation of the

[1]Our HD setting refers to the HD parametric (linear or generalised linear) model and the ML setting refers to ML models of conditional mean estimation (prediction/classification) that is black box and usually nonparametric.

odds ratio was built on $p(A \mid \boldsymbol{X}, Y = 0)$, the conditional density of A given $\boldsymbol{X}$ and $Y = 0$ (e.g., Chen, 2007; Tchetgen Tchetgen et al., 2010).[2]

Nevertheless, Tan (2019a) focused on fixed-dimensional parametric nuisance models that were still prone to model misspecification. Their proposed approach is not readily applicable to the recently developed and exploited HD ($p \gg n$ and the two nuisance components are specified as parametric models with sparse coefficients) or ML (the nuisance functions are estimated by arbitrary black box learning algorithm of condition mean) approaches. For the HD framework, this is because simply using regularised nuisance estimators in (1.2) would typically incur excessive bias and would not guarantee the parametric rate of convergence. In this paper, we realise bias reduction with respect to the regularisation errors by constructing certain Dantzig moment equations to estimate the nuisance parameters. With ultra-sparse nuisance parameters, i.e., sparsity level is $o\left(n^{1/2}/\log p\right)$, our proposed estimator preserves the model double robustness property such that it approaches β_0 at the rate $O_p\left(n^{-1/2}\right)$ when either $r(\cdot)$ or $m(\cdot)$ is correctly specified. Under the ML framework, the nonlinearity and unextractability of the logit link makes it impossible to naturally estimater $r_0(\cdot)$ with a learning algorithm of conditional mean as the partially linear setting. We handle this challenge through an easy-to-implement 'full model refitting' (FMR) procedure that facilitates flexible implementation of arbitrary learning algorithms in our framework. Our double machine learning (DML) estimator for β_0 is rated doubly robust in the same sense as Chernozhukov, Chetverikov, et al. (2018), i.e., asymptotically normal at rate $n^{-1/2}$ when the two nuisance ML estimators are consistent for the true models and their root mean squared errors are controlled by $o_p\left(n^{-1/4}\right)$.

In recent years, there has been a large body of literature developed for semiparametric inference with HD and ML nuisance models, which has garnered increased attention and applications in economics and policy sciences (e.g., Athey and Imbens, 2017; Knaus, 2018, 2020; Yang et al., 2020. As reflected in the organisation of our paper, there are two different frameworks in the literature, the HD and ML settings, both often referred to as 'machine learning' approaches. The main difference between them is that the HD setting imposes parametric assumptions on the nuisance models and may allow for their potential misspecification, while the ML setting uses nonparametric ML nuisance estimators that are supposed to approach the true functions at a certain geometric rate.

To estimate low dimensional parameters such as the average treatment effect (ATE) and conditional treatment effect in linear or log-linear models in the HD setting with potentially misspecified nuisance models, recent studies, including Smucler et al.(2019),Tan (2020a,2020b), Ning et al. (2020), and Dukes and Vansteelandt (2020), constructed ℓ_1-regularised estimating equations with certain ℓ_∞-constraints to simultaneously estimate the nuisance parameters and calibrate their first order bias. In comparison, Bradic et al. (2019) proposed a more sparsity-rate robust ATE estimator that requires substantially weaker sparsity assumptions but needs both HD parametric nuisance models to be correctly

[2]Chen (2007) and Tchetgen Tchetgen et al. (2010) estimate the conditional distribution of $\boldsymbol{A}$ given $\boldsymbol{X}$ while Tan (2019a) only needs to specify a conditional mean model of $\boldsymbol{A}$.

specified. Our work is the first to consider the logistic partially linear model under a similar HD model robustness regime. Existing approaches including debiased (desparsified) LASSO (see Van de Geer et al., 2014; Janková and Van de Geer, 2016) and regularised Riesz representer (see Chernozhukov, Newey and Robins, 2018; Belloni et al., 2018) used the empirical inverse of the information matrix obtained with ℓ_1-regularised regression to correct for the bias of the logistic LASSO estimator. They imposed a technical ultra-sparsity condition on the inverse of the information matrix, an approach which has been criticised as unreasonable and unverifiable (Xia et al., 2020). In contrast, our sparsity assumption is model-specific and therefore more reasonable and explainable.

We note that near the end of finalizing our paper, a parallel paper by Ghosh and Tan (2020) was published on arXiv.org. Our HD framework studies the same problem but uses a different doubly robust estimating equation and calibrated procedures for the nuisance models. While the two proposals have similar theoretical properties and numerical implementation strategies (see Sections 3.1, 4.1, and Appendix A4), our method under the ML setting introduced below has no overlap with their work. In parallel, Nekipelov et al. (2018) published a preprint (the latest version in October 2020) about DML estimation of the generalised partially linear model using an adjusted Neyman orthogonal estimating equation different from ours. As a comparison, their approach works for more general target models with arbitrary (smooth) link functions and types of outcome while our method is restricted to the logit link and binary outcome.[3] However, their construction cannot achieve the model double robustness property as we can with either fixed or high dimensional parametric nuisance models. In particular, when the nuisance model $r_0(\boldsymbol{x})$ is misspecified, their estimator is invalid while ours is still valid as long as $m_0(\boldsymbol{x})$ is correctly specified. This is due to them using $\mathbb{E}(A \mid \boldsymbol{X} = \boldsymbol{x})$ as the nuisance model while we use $\mathbb{E}(A \mid Y = 0, \boldsymbol{X} = \boldsymbol{x})$. See Tan (2019a) for more discussion on this.

Under the nonparametric ML setting, Chernozhukov, Chetverikov, et al. (2018) established a DML framework utilising Neyman orthogonal scores and cross-fitting to construct a parametrically efficient ML-based casual estimator. Their framework has played a central role in semiparametric inference with ML. To complement their approach, recent work, including Chernozhukov, Newey, Robins and Singh (2018), Zimmert and Lechner (2019), and Colangelo and Lee (2020), localised the orthogonal score function to estimate conditional average treatment effect; Semenova and Chernozhukov (2020) constructed the best linear approximation of a structural function with ML; Farrell et al. (2018) used deep neural networks for DML estimation; Wager and Athey (2018) and Oprescu et al. (2019) proposed and studied the tree-based ML approaches for causal inference; and Cui and Tchetgen Tchetgen (2019) proposed a minimax data-driven model selection approach to choose the ML nuisance models with the lowest bias on the DML estimator. The above-mentioned work elaborated on specific inference problems including partially linear models, ATE, and heterogeneous treatment effects with nuisance models which can be directly estimated with arbitrary (supervised) ML algorithms. As mentioned above, $r_0(\cdot)$ in our case cannot be estimated with general ML algorithms due to the nonlinear structure of (1.1). To

[3]It is still an open problem to generalise our approach to general link functions in M-estimation.

the best of our knowledge, this paper is the first to solve this nontrivial technical problem through our proposed FMR procedure.

The rest of this paper is organised as follows. In Section 2, we define the Neyman orthogonal (doubly robust) score equation for logistic partially linear models. In Section 3, we introduce the realisation of debiased and DML inference for β_0 under the HD and ML settings respectively. In Section 4, we present and justify the asymptotic property of our HD and ML estimators. In Sections 5 and 6, we conduct simulations to study the empirical performance of our method and apply it to assess the effects of the emergency contraceptive (EC) pill on early gestation fetal and new births.

2. NEYMAN ORTHOGONAL SCORE

Before coming to the specific approaches in Section 3, we introduce a Neyman orthogonal (doubly robust) score function for logistic partially linear models and derive its first order bias, which plays a central role in motivating and guiding our methods and theoretical analysis. Let observation $\boldsymbol{D}_i = \{Y_i, A_i, \boldsymbol{X}_i\}$ for $i = 1, \ldots, n$ and $\boldsymbol{D} = \{Y, A, \boldsymbol{X}\}$ be a realisation of $\boldsymbol{D}_i$. Motivated by equation (1.2) and Tan (2019a), we define the Neyman orthogonal score as

$$h(\boldsymbol{D}; \beta, \eta) = \psi(\boldsymbol{X})\left\{Ye^{-\beta A} - (1-Y)e^{r(\boldsymbol{X})}\right\}\left\{A - m(\boldsymbol{X})\right\},$$

where $\eta = \{r(\cdot), m(\cdot), \psi(\cdot)\}$ represents the whole set of nuisance functions. Similar to (1.2), $r(\cdot)$ and $m(\cdot)$ correspond to the nonparametric component $r_0(\cdot)$ defined in (1.1) and $m_0(\boldsymbol{x}) = \mathbb{E}(A|Y = 0, \boldsymbol{X} = \boldsymbol{x})$, respectively. $\psi(\boldsymbol{x})$ is a nuisance function affecting the asymptotic variance of the estimator that may depend on $r(\boldsymbol{x})$ and $m(\boldsymbol{x})$ and actually corresponds to $\phi(\boldsymbol{x})e^{-r(\boldsymbol{x})}$ with $\phi(\boldsymbol{x})$ defined by (1.2).[4]

REMARK 2.1.

The score function $h(\boldsymbol{D}; \beta, \eta)$ is doubly robust in the sense that when $r(\cdot) = r_0(\cdot)$ or $m(\cdot) = m_0(\cdot)$, β_0 solves $\mathbb{E}h(\boldsymbol{D}; \beta, \eta) = 0$. We shortly demonstrate this as follows. When either $r(\cdot) = r_0(\cdot)$ or $m(\cdot) = m_0(\cdot)$ holds, we have

$$\begin{aligned}&\mathbb{E}\psi(\boldsymbol{X})(1-Y)\left\{e^{r(\boldsymbol{X})} - e^{r_0(\boldsymbol{X})}\right\}\left\{A - m(\boldsymbol{X})\right\}\\ &= \mathbb{E}\left[\psi(\boldsymbol{X})\left\{e^{r(\boldsymbol{X})} - e^{r_0(\boldsymbol{X})}\right\}\left\{A - m(\boldsymbol{X})\right\} \mid Y = 0, \boldsymbol{X}\right] = 0,\end{aligned}$$

which combined with (1.1) leads to that

$$\begin{aligned}\mathbb{E}h(\boldsymbol{D}; \beta_0, \eta) &= \mathbb{E}\psi(\boldsymbol{X})\left\{Ye^{-\beta_0 A} - (1-Y)e^{r(\boldsymbol{X})}\right\}\left\{A - m(\boldsymbol{X})\right\}\\ &= \mathbb{E}\psi(\boldsymbol{X})e^{r_0(\boldsymbol{X})}\left\{Ye^{-\beta_0 A - r_0(\boldsymbol{X})} - (1-Y)\right\}\left\{A - m(\boldsymbol{X})\right\}\\ &= \mathbb{E}\psi(\boldsymbol{X})e^{r_0(\boldsymbol{X})}\left\{A - m(\boldsymbol{X})\right\}\frac{Y - \mathbb{P}(Y = 1 \mid A, \boldsymbol{X})}{\mathbb{P}(Y = 1 \mid A, \boldsymbol{X})} = 0.\end{aligned}$$

[4]We rewrite (1.2) with $\psi(\boldsymbol{x}) = \phi(\boldsymbol{x})e^{-r(\boldsymbol{x})}$ to form the score function so that one could find both its partial derivatives on r and ψ are Neyman orthogonal in a more explicit way.

Suppose the nuisance models $r_0(\boldsymbol{x})$ and $m_0(\boldsymbol{x})$ are estimated by $\hat{r}(\boldsymbol{x})$ and $\hat{m}(\boldsymbol{x})$ converging to $\bar{r}(\boldsymbol{x})$ and $\bar{m}(\boldsymbol{x})$ respectively, and let $\hat{\psi}(\boldsymbol{x})$ represent the estimator for $\psi(\boldsymbol{x})$ approaching $\bar{\psi}(\boldsymbol{x})$. Denote by $\bar{\eta} = \{\bar{r}(\cdot), \bar{m}(\cdot), \bar{\psi}(\cdot)\}$, and $\hat{\eta} = \{\hat{r}(\cdot), \hat{m}(\cdot), \hat{\psi}(\cdot)\}$. We then write the Gateaux (partial) derivative of the score function $h(\boldsymbol{D}; \beta_0, \bar{\eta})$ as

$$\begin{aligned}
&\partial_\eta h(\boldsymbol{D}; \beta_0, \bar{\eta})[\eta - \bar{\eta}] \\
&\quad = \partial_\psi h(\boldsymbol{D}; \beta_0, \bar{\eta})[\psi - \bar{\psi}] + \partial_r h(\boldsymbol{D}; \beta_0, \bar{\eta})[r - \bar{r}] + \partial_m h(\boldsymbol{D}; \beta_0, \bar{\eta})[m - \bar{m}] \\
&\quad =: \left\{ Y e^{-\beta_0 A} - (1 - Y) e^{\bar{r}(\boldsymbol{X})} \right\} \left\{ A - \bar{m}(\boldsymbol{X}) \right\} \left\{ \psi(\boldsymbol{X}) - \bar{\psi}(\boldsymbol{X}) \right\} \\
&\qquad - (1 - Y) \bar{\psi}(\boldsymbol{X}) e^{\bar{r}(\boldsymbol{X})} \left\{ A - \bar{m}(\boldsymbol{X}) \right\} \left\{ r(\boldsymbol{X}) - \bar{r}(\boldsymbol{X}) \right\} \\
&\qquad - \bar{\psi}(\boldsymbol{X}) \left\{ Y e^{-\beta_0 A} - (1 - Y) e^{\bar{r}(\boldsymbol{X})} \right\} \left\{ m(\boldsymbol{X}) - \bar{m}(\boldsymbol{X}) \right\}.
\end{aligned} \tag{2.1}$$

REMARK 2.2.

We evaluate the Neyman orthogonal score on some limiting parameters $\bar{r}(\cdot)$ and $\bar{m}(\cdot)$ instead of on $r_0(\cdot)$ and $m_0(\cdot)$ as in Chernozhukov, Chetverikov, et al.(2018). This is because, different from their ML framework (and ours), assuming both nuisance estimators converge to the true models, i.e., $\bar{r}(\cdot) = r_0(\cdot)$ and $\bar{m}(\cdot) = m_0(\cdot)$, our HD realisation allows at most one nuisance model to be wrongly specified.

Inspired by our deduction in Remark 2.1, $\mathbb{E}\,\partial_\psi h(\boldsymbol{D}; \beta_0, \bar{\eta})[\psi - \bar{\psi}] = 0$ for any ψ whenever $\bar{r}(\cdot) = r_0(\cdot)$ or $\bar{m}(\cdot) = m_0(\cdot)$. Also, $\mathbb{E}\,\partial_r h(\boldsymbol{D}; \beta_0, \bar{\eta})[r - \bar{r}] = 0$ when $\bar{m}(\cdot) = m_0(\cdot)$ and $\mathbb{E}\,\partial_m h(\boldsymbol{D}; \beta_0, \bar{\eta})[m - \bar{m}] = 0$ when $\bar{r}(\cdot) = r_0(\cdot)$. Thus, under the ML setting, $h(\boldsymbol{D}; \beta_0, \bar{\eta})$ satisfies the Neyman orthogonality condition, $\partial_\eta h(\boldsymbol{D}; \beta_0, \bar{\eta})[\eta - \bar{\eta}] = 0$, as defined in Chernozhukov, Chetverikov, et al. (2018) and the first order (over-fitting) bias of $n^{-1} \sum_{i=1}^{n} h(\boldsymbol{D}_i; \beta, \hat{\eta}))$ can be removed through cross-fitting (introduced in Section 3.2) and concentration. While under the HD parametric setting with $\bar{r}(\cdot) \neq r_0(\cdot)$ or $\bar{m}(\cdot) \neq m_0(\cdot)$, the moment equations for $\bar{r}(\cdot)$ and $\bar{m}(\cdot)$ must be carefully constructed to ensure the orthogonality conditions. Similar to existing literature, such as Chernozhukov, Chetverikov, et al. (2018) and Tan (2020a), removal of the second order (and beyond) bias relies on the assumption on quality of the nuisance estimators $\hat{r}(\cdot)$ and $\hat{m}(\cdot)$ (see Section 4).

3. METHOD

In this section, we separately present our specific construction procedures for HD parametric and ML nonparametric realisation of the debiased/DML estimator for β_0, based on the derivation and discussion in Section 2.

3.1. High dimensional parametric model realisation

Consider the setting with $p \gg n$, where each $\boldsymbol{X}_i$ has its first element being 1, and $r(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\gamma}$ and $m(\boldsymbol{x}) = g\left(\boldsymbol{x}^\top \boldsymbol{\alpha}\right)$, where $g(\cdot)$ is a monotone and smooth link function with derivative $g'(\cdot)$. Inspired by Smucler et al. (2019), Tan (2020a), and Dukes and Vansteelandt (2020), we construct Dantzig moment equations to ensure the Neyman orthogonality empirically: $\partial_r h(\boldsymbol{D}; \beta_0, \bar{\eta})[r - \bar{r}] = 0$ and $\partial_m h(\boldsymbol{D}; \beta_0, \bar{\eta})[m - \bar{m}] = 0$, under potential misspecification of one the nuisance models.

First, we obtain $\tilde{\boldsymbol{\gamma}}$ as some initial estimator for $\boldsymbol{\gamma}$ that converges to some limiting parameter $\boldsymbol{\gamma}^*$ equal to the true model parameter $\boldsymbol{\gamma}_0$ when $r(\boldsymbol{x})$ is correctly specified as $r_0(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\gamma}_0$. Let $\widehat{\psi}(\boldsymbol{x})$ be some estimator of the nuisance function $\psi(\boldsymbol{x})$ depending on $\tilde{\boldsymbol{\gamma}}$ with its limiting function being $\overline{\boldsymbol{\psi}}(\boldsymbol{x})$, whose choice will be discussed in Section 3.3. According to (2.1), we obtain $\hat{\boldsymbol{\alpha}}$ through the Dantzig moment equation:

$$\min_{\alpha \in \mathbb{R}^p} \| \boldsymbol{\alpha} \|_1 \quad \text{s.t} \quad \left\| n^{-1} \sum_{i=1}^{n} (1 - Y_i) \widehat{\psi}(\boldsymbol{X}_i) e^{\boldsymbol{X}_i^\top \tilde{\boldsymbol{\gamma}}} \left\{ A_i - g\left(\boldsymbol{X}_i^\top \boldsymbol{\alpha}\right) \right\} \boldsymbol{X}_i \right\|_\infty \leq \lambda_\alpha, \tag{3.1}$$

where λ_α is a tuning parameter controlling the regularisation bias. Finally, we obtain the nuisance estimator $\hat{\boldsymbol{\gamma}}$ and the targeted HD estimator $\widehat{\beta}_{\mathrm{HD}}$ simultaneously from:

$$\begin{aligned} \min_{\beta \in \mathbb{R}, \gamma \in \mathbb{R}^p} \| \boldsymbol{\gamma} \|_1 \quad \text{s.t} \quad & \left\| n^{-1} \sum_{i=1}^{n} \widehat{\psi}(\boldsymbol{X}_i) \left\{ Y_i e^{-\beta A_i} - (1 - Y_i) e^{\boldsymbol{X}_i^\top \boldsymbol{\gamma}} \right\} g'\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right) \boldsymbol{X}_i \right\|_\infty \leq \lambda_\gamma; \\ & n^{-1} \sum_{i=1}^{n} \widehat{\psi}(\boldsymbol{X}_i) \left\{ Y_i e^{-\beta A_i} - (1 - Y_i) e^{\boldsymbol{X}_i^\top \boldsymbol{\gamma}} \right\} \left\{ A_i - g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right) \right\} = 0. \end{aligned} \tag{3.2}$$

Let the limits of $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}\}$ be $\{\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\gamma}}\}$ and $\bar{\eta} = \{\bar{r}(\cdot), \bar{m}(\cdot), \overline{\psi}(\cdot)\}$ where $\bar{r}(\boldsymbol{x}) = \boldsymbol{x}^\top \bar{\boldsymbol{\gamma}}$, $\bar{m}(\boldsymbol{x}) = g\left(\boldsymbol{x}^\top \bar{\boldsymbol{\alpha}}\right)$ and $\overline{\psi}(\boldsymbol{x})$ as given in Section 3.3. We will comment on the orthogonality (moment) conditions of our proposal in Remark 3.1, compare our method with Dukes and Vansteelandt (2020) in Remark 3.2, and discuss its numerical implementation with a weighted LASSO formation in Remark 3.3.

Remark 3.1.—Neglect the second order error terms for now. When $r(\boldsymbol{x})$ is correct (see Assumption HD1), i.e., $r_0(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\gamma}_0$, it naturally holds that $\mathbb{E}\, \partial_m h(\boldsymbol{D}; \beta_0, \bar{\eta})[\hat{m} - m_0] = 0$ and $\boldsymbol{\gamma}^* = \bar{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0$. Then the ℓ_∞-constraint in (3.1) imposes that

$$\mathbb{E}\, \partial_r h(\boldsymbol{D}; \beta_0, \bar{\eta})[\hat{r} - r_0] \approx \mathbb{E}(1 - Y) \overline{\psi}(\boldsymbol{X}) e^{\boldsymbol{X}^\top \boldsymbol{\gamma}_0} \left\{ A - g\left(\boldsymbol{X}^\top \bar{\boldsymbol{\alpha}}\right) \right\} \boldsymbol{X}^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = \mathbf{0}^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0).$$

When $\bar{m}(\boldsymbol{x}) = m_0(\boldsymbol{x}) = g\left(\boldsymbol{x}^\top \boldsymbol{\alpha}_0\right)$, we have $\mathbb{E}\, \partial_r h(\boldsymbol{D}; \beta_0, \bar{\eta})[\hat{r} - r_0]$ and $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0$ in turn. And the ℓ_∞-constraint of (3.2) results in

$$\begin{aligned} & \mathbb{E}\, \partial_m h(\boldsymbol{D}; \beta_0, \bar{\eta})[\hat{m} - m_0] \approx \\ mathbb{E} \overline{\psi}(\boldsymbol{X}) & \left\{ Y e^{-\beta_0 A} - (1 - Y) e^{\boldsymbol{X}^\top \bar{\boldsymbol{\gamma}}} \right\} g'\left(\boldsymbol{X}_i^\top \boldsymbol{\alpha}_0\right) \boldsymbol{X}^\top (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \\ & = \mathbf{0}^\top (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \end{aligned}$$

Thus, the Neyman orthogonality condition $\partial_\eta h(\boldsymbol{D}; \beta_0, \bar{\eta})[\eta - \bar{\eta}] = 0$ as introduced in Section 2 is satisfied under our construction when either $r(\cdot)$ or $m(\cdot)$ is correctly specified.

Remark 3.2.—Similar to the HD partially linear (or log-linear) setting studied in Dukes and Vansteelandt (2020), estimating an equation for the nuisance parameter $\boldsymbol{\gamma}$ involves the unknown β. Unlike their construction procedure that plugs in β_0 as every $\beta \in \mathbb{R}$ to estimate $\boldsymbol{\gamma}$ and invert the resulted score-test p-values for interval estimation of β_0, we solve for $\hat{\beta}$ and $\hat{\boldsymbol{\gamma}}$ jointly from (3.2), the Dukes and Vansteelandt method, our approach is more friendly in terms of computation and with a doubly robust moment equation for $\hat{\boldsymbol{\beta}}$, as demonstrated in Remark 2.1. Compared with implementation, directly provides a point estimator, and preserves a similar theoretical guarantee (see Section 4.1).

Remark 3.3.—As is detailed in Appendix A4, one can construct LASSO problems with the same Karus–Kuhn–Tucker (KKT) conditions as the ℓ_∞-norm constraints in (3.1) and (3.2) to obtain the estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$, which have equivalent theoretical properties as (3.2), numerical solution of the LASSO counterpart of (3.2) cannot be obtained with existing software, such as the ***R*** packages 'glmnet' (Friedman et al., 2010) and 'RCAL' (Tan, 2019b). A direct solution to this is programming an optimisation procedure such as the Fisher scoring descent algorithm used by Tan (2020b). We also found a convenient way to moderately modify the construction procedure to make the regularised estimating equations solvable with ***R*** package 'RCAL', and use it for numerical implementation. In Appendix A4, we outline this modification and demonstrate its theoretical guarantee.

3.2. Machine learning realisation

We turn now to a (nonparametric) ML setting under which any learning algorithms of conditional mean could potentially be applied to estimate the nuisance functions. Similar to Chernozhukov, Chetverikov, et al. (2018), we randomly split the n samples into K folds: $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_K$ of equal size, to assist with removing the first order (over-fitting) bias through concentration. Then the cross-fitted estimating equation for β is constructed as

$$n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} h\left(\boldsymbol{D}_i; \beta, \hat{\eta}^{[-k]}\right) = 0, \tag{3.3}$$

where $\hat{\eta}^{[-k]} = \left\{\hat{r}^{[-k]}(\cdot), m^{[-k]}(\cdot), \psi^{[-k]}(\cdot)\right\}$, representing ML estimators converging to $\bar{r}(\cdot) = r_0(\cdot)$, $\bar{m}(\cdot) = m_0(\cdot)$, and $\bar{\psi}(\cdot)$ obtained with the samples in $\mathcal{I}_{-k} = \{1, \ldots, n\} \setminus \mathcal{I}_k$ and independent of the samples in $\mathcal{I}_k$. Now we present the specific construction procedures for $\hat{r}^{[-k]}(\cdot)$ and $m^{[-k]}(\cdot)$ with the choice of $\psi^{[-k]}(\cdot)$ again discussed in Section 3.3.

Suppose there is a black box learning algorithm $\mathcal{L}(\boldsymbol{R}_i, \boldsymbol{C}_i; \mathcal{I})$ that inputs samples $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ with some response $\boldsymbol{R}_i$ and covariates $\boldsymbol{C}_i$, and outputs an estimation of $\mathbb{E}[\boldsymbol{R}_i \mid \boldsymbol{C}_i, i \in \mathcal{I}]$. We outline our approach utilising $\mathcal{L}$ to estimate the nuisance functions. Corresponding to the definition of $m_0(\cdot)$, it can be estimated by: $\hat{m}^{[-k]}(\cdot) = \mathcal{L}(\boldsymbol{A}_i, \boldsymbol{X}_i; \mathcal{I}_{-k} \cap \{i : Y_i = 0\})$. Compared to the partially linear setting in Chernozhukov, Chetverikov, et al. (2018), estimation of $r_0(\cdot)$ with $\mathcal{L}$ is more sophisticated

as it is defined through a nonlinear form: $\mathbb{P}(Y = 1 \mid A, \boldsymbol{X}) = \text{expit}\{\beta_0 A + r_0(\boldsymbol{X})\}$. One could modify some ML approaches, e.g., neural network[5] to accommodate this form. However, such modification is not readily available in general and typically requires additional human efforts for its implementation and validation. So, alternatively, we propose a 'full model refitting' (FMR) procedure that can use an arbitrary $\mathscr{L}$ to estimate $r_0(\cdot)$. Our method is motivated by a simple proposition:

Proposition 3.1.—*Let* $M_0(A, \boldsymbol{X}) = \mathbb{P}(Y = 1 \mid A, \boldsymbol{X}) = \text{expit}\{\beta_0 A + r_0(\boldsymbol{X})\}$. *We have:*

$$\beta_0 = \text{argmin}_{\beta \in \mathbb{R}} \mathbb{E}[\text{logit}\{M_0(A, \boldsymbol{X})\} - \beta(A - \mathbb{E}[A \mid \boldsymbol{X}])]^2.$$

Proof.: For any $\beta \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}[\text{logit}\{M_0(A, \boldsymbol{X})\} - \beta(A - \mathbb{E}[A \mid \boldsymbol{X}])]^2 &= \mathbb{E}\{\beta_0 A + r_0(\boldsymbol{X}) - \beta(A - \mathbb{E}[A \mid \boldsymbol{X}])\}^2 \\ &= \mathbb{E}\{(\beta_0 - \beta)(A - \mathbb{E}[A \mid \boldsymbol{X}]) + \tau(\boldsymbol{X})\}^2 = (\beta_0 - \beta)^2 \mathbb{E}(A - \mathbb{E}[A \mid \boldsymbol{X}])^2 + \mathbb{E}\{\tau(\boldsymbol{X})\}^2,\end{aligned}$$

where $\tau(\boldsymbol{X}) = r_0(\boldsymbol{X}) + \beta_0 \mathbb{E}[A \mid \boldsymbol{X}]$. Thus, β_0 minimises

$$\mathbb{E}[\text{logit}\{M_0(A, \boldsymbol{X})\} - \beta(A - \mathbb{E}[A \mid \boldsymbol{X}])]^2.$$

□

Further, randomly split $\mathscr{I}_{-k}$ into K folds $\mathscr{I}_{-k,1}, \ldots \mathscr{I}_{-k,K}$, of equal size and denote by $\mathscr{I}_{-k,-j} = \mathscr{I}_{-k} \setminus \mathscr{I}_{-k,j}$. Motivated by Proposition 3.1, we first estimate the 'full' model $M_0(A, \boldsymbol{X})$ with $\mathscr{I}_{-k,-j}$ as:

$$\widehat{M}^{[-k,-j]}(\cdot) = \mathscr{L}\left(Y_i, \left(A_i, \boldsymbol{X}_i^\top\right)^\top; \mathscr{I}_{-k,-j}\right),$$

and learn $a_0(\boldsymbol{x}) = \mathbb{E}[A \mid \boldsymbol{X} = \boldsymbol{x}]$ by $\hat{a}^{[-k,-j]}(\cdot) = \mathscr{L}(A_i, \boldsymbol{X}_i; \mathscr{I}_{-k,-j})$. Then we fit the (cross-fitted) least square regression to obtain:

$$\breve{\beta}^{[-k]} = \text{argmin}_{\beta \in \mathbb{R}} \frac{1}{|\mathscr{I}_{-k}|} \sum_{j=1}^{K} \sum_{i \in \mathscr{I}_{-k,j}} \left[\text{logit}\left\{\widehat{M}^{[-k,-j]}(A_i, \boldsymbol{X}_i)\right\} - \beta\left\{A_i - \hat{a}^{[-k,-j]}(\boldsymbol{X}_i)\right\}\right]^2, \tag{3.4}$$

as an estimator approaching β_0 at certain rate typically larger than $n^{-1/2}$. Then $r_0(\cdot)$ could be identified through $r_0(\boldsymbol{X}_i) = \text{logit}\{M_0(A_i, \boldsymbol{X}_i)\} - \beta_0 A_i$. Note that the empirically estimated version of $\text{logit}\{M_0(A_i, \boldsymbol{X}_i)\} - \beta_0 A_i$ typically involves A_i due to the discrepancy of β_0 and $M_0(\cdot)$ from their empirical estimation. This can essentially impede removal of the over-

[5]By setting the last layer of the neural network to be the combination of a complex network of $\boldsymbol{X}$ and a linear function of $\boldsymbol{A}$ and linking it with the outcome through an expit link.

fitting bias given that $\partial_r h(\boldsymbol{D}; \beta_0, \bar{\eta})$ is not orthogonal to the error functions depending on A. So we further estimate the conditional mean of $\text{logit}\{M_0(A, \boldsymbol{X})\} - \beta_0 A$ on $\boldsymbol{X}$ to obtain the final estimator for $r_0(\cdot)$. Denote by $W_i = \text{logit}\left\{\widehat{M}^{[-k,-j]}(A_i, \boldsymbol{X}_i)\right\}$ for each $i \in \mathcal{I}_{-k,j}$ and get $\hat{t}^{[-k]}(\cdot) = \mathcal{L}(W_i, \boldsymbol{X}_i; \mathcal{I}_{-k})$ to estimate $t_0(\boldsymbol{x}) = :\mathbb{E}[\text{logit}\{M_0(A, \boldsymbol{X})\} \mid \boldsymbol{X} = \boldsymbol{x}]$. Then the estimator of $r_0(\cdot)$ is given by:

$$\hat{r}^{[-k]}(\boldsymbol{x}) = \hat{t}^{[-k]}(\boldsymbol{x}) - \breve{\beta}^{[-k]}\hat{a}^{[-k]}(\boldsymbol{x}), \quad \text{where} \quad \hat{a}^{[-k]}(\boldsymbol{x}) = \frac{1}{K}\sum_{j=1}^{K}\hat{a}^{[-k,-j]}(\boldsymbol{x}). \tag{3.5}$$

Alternatively, one can estimate $r_0(\cdot)$ through

$$\hat{r}^{[-k]}(\cdot) = \log\left(\frac{\mathcal{L}\left(e^{-\breve{\beta}^{[-k]}A_i}, \boldsymbol{X}_i; \mathcal{I}_{-k} \cap \{i: Y_i = 1\}\right)}{\mathcal{L}(1 - Y_i, \boldsymbol{X}_i; \mathcal{I}_{-k})}\right),$$

motivated by the moment condition that is sufficient to identify $r_0(\cdot)$:

$$\mathbb{E}\left[Ye^{-\beta_0 A} - (1 - Y)e^{r_0(\boldsymbol{X})} \mid \boldsymbol{X}\right] = \mathbb{E}\left[e^{-\beta_0 A} \mid \boldsymbol{X}, Y = 1\right] - e^{r_0(\boldsymbol{X})}\mathbb{E}[(1 - Y) \mid \boldsymbol{X}] = 0.$$

We refer to the estimation step for $\breve{\beta}^{[-k]}$ and $\hat{r}^{[-k]}(\cdot)$ introduced above as 'refitting', and the whole procedure as FMR considering that we 'refit' the least square problem (3.4) and ML models $\mathcal{L}$ to estimate $r_0(\cdot)$ with the initially estimated full model $\text{logit}\{M_0(A_i, \boldsymbol{X}_i)\}$ as a pseudo-outcome. Finally, we solve (3.3) based on $\hat{\eta}^{[-k]}$ to obtain the DML estimator $\hat{\beta}_{\text{ML}}$.

Remark 3.4.—We further use cross-fitting in FMR to avoid over-fitting of the models $\widehat{M}^{[-k,-j]}(\cdot)$ and $a^{[-k,-j]}(\cdot)$ when they are used to obtain the estimators $\breve{\beta}^{[-k]}$, $\hat{t}^{[-k]}(\boldsymbol{x})$ and $\hat{r}^{[-k]}(\boldsymbol{x})$. This is supposed to show empirical improvement of the FMR procedure.

Remark 3.5.—The FMR implicitly assumes that $\mathcal{L}$ should perform similarly well on different learning objects with the covariates set as either $\boldsymbol{X}$ or $\left(A, \boldsymbol{X}^\top\right)^\top$. Classic nonparametric approaches like kernel smoothing or sieve may not satisfy this assumption whereas including one more covariate A in addition to the very low dimensional $\boldsymbol{X}$ can have substantial impact on estimation performance. Thus, we recommend using more dimensionality-robust modern ML approaches, such as random forest and neural networks, in our ML framework. The classic 'plug-in' sieve or kernel method has been well-studied in existing literature (e.g., Severini and Staniswalis (1994); Lin and Carroll (2006).

3.3. Efficiency considerations

The nuisance function $\psi(\cdot)$ in our framework is included and chosen in consideration of estimation efficiency. Tan (2019a) proposed and studied two options for $\phi(\cdot)$ used and defined in (1.2), with the corresponding function $\psi(\cdot)$ taken as:

$$\psi_{\text{opt}}(\boldsymbol{x}) = \frac{e^{-r(\boldsymbol{x})}\mathbb{E}\left[\{A - m(\boldsymbol{X})\}^2 \mid \boldsymbol{X} = \boldsymbol{x}, Y = 0\right]}{\mathbb{E}\left[\{A - m(\boldsymbol{X})\}^2 \text{expit}\{\beta_0 A + r(\boldsymbol{X})\} \mid \boldsymbol{X} = \boldsymbol{x}, Y = 0\right]},$$

and $\psi_{\text{simp}}(\boldsymbol{x}) = \text{expit}\{-r(\boldsymbol{x})\}$.

Remark 3.6.—It was shown in Tan (2019a) that when both nuisance models are correctly specified, the estimator solved with the weight $\psi_{\text{opt}}(\cdot)$ achieves the minimum asymptotic variance among all the doubly robust estimators obtained through our estimating equation. Given that the weighting function $\psi(\cdot)$ is independent of the exposure A, our estimating equations form a strict subset of all the score equations for the logistic partial model defined by (1.1). Thus, our estimator is generally less efficient than the semiparametric efficient estimator obtained through the maximum likelihood approach like Tchetgen Tchetgen et al. (2010).

Though $\psi_{\text{opt}}(\cdot)$ is the optimal choice in consideration of the efficiency under our construction, computation of $\psi_{\text{opt}}(\cdot)$ involves numerical integration with respect to $\boldsymbol{X}$ given $Y = 0$, making it sometimes inconvenient to implement. So Tan (2019a) proposed a simplified but reasonable choice $\psi_{\text{simp}}(\boldsymbol{x})$ obtained by evaluating $\psi_{\text{opt}}(\boldsymbol{x})$ at $\beta_0 = 0$. In the following theoretical and numerical studies, we stick to $\psi(\boldsymbol{x}) = \psi_{\text{simp}}(\boldsymbol{x}), \widehat{\psi}(\boldsymbol{x}) = \text{expit}\left(-\boldsymbol{x}^\top \tilde{\boldsymbol{\gamma}}\right)$ and correspondingly $\overline{\psi}(\boldsymbol{x}) = \text{expit}\left(-\boldsymbol{x}^\top \boldsymbol{\gamma}^*\right)$ under the HD setting, and $\widehat{\psi}^{[-k]}(\boldsymbol{x}) = \text{expit}\left\{-\hat{r}^{[-k]}(\boldsymbol{x})\right\}$ and $\overline{\psi}(\boldsymbol{x}) = \text{expit}\{-r_0(\boldsymbol{x})\}$ under the ML setting. Our theoretical framework allows for other choices on $\psi(\cdot)$ as will be discussed in Section 4.2.

4. ASYMPTOTIC ANALYSIS

Let $o(\alpha_n)$, $O(\alpha_n)$, $\omega(\alpha_n)$, $\Omega(\alpha_n)$, and $\Theta(\boldsymbol{\alpha}_n)$ represent the sequences growing at a smaller, equal/smaller, larger, equal/larger, and equal rate of α_n, respectively. Let $o_{\mathbb{P}}$, $O_{\mathbb{P}}$, $\omega_{\mathbb{P}}$, $\Omega_{\mathbb{P}}$, and $\Theta_{\mathbb{P}}$ be the corresponding rates with probability approaching 1 as $n \to \infty$. Let $\mathcal{X} \subseteq \mathbb{R}^p$ be the domain of $\boldsymbol{X}$. First, we introduce the regularity condition for β and its estimating equation used under both HD and ML settings as Assumption REG, which is standard and can be commonly found in literature of the asymptotic analysis of M-estimator (see Van der Vaart, 2000, ch. 5). We will then study the asymptotic properties of $\widehat{\beta}_{\text{HD}}$ and $\widehat{\beta}_{\text{ML}}$ in Sections 4.1 and 4.2.

Assumption REG (Regularity of estimating equation).

Parameter β *belongs to a compact set* $\mathscr{B} \subseteq \mathbb{R}$ *and there exists* $\delta_n = \Omega\left(n^{-1/2}\log n\right)$ *such that* $(\beta_0 - \delta_n, \beta_0 + \delta_n) \subseteq \mathscr{B}$. *Exposure* A *belongs to a compact set* $\mathscr{A}$ *and* $\sup_{x \in \mathscr{X}} |\mathbb{E}[A \mid \boldsymbol{X} = \boldsymbol{x}, Y = y]| = O(1)$ *for* $y = 0, 1$. *In addition,* [6]

$$\mathbb{E}\overline{\psi}(\boldsymbol{X})Ye^{-\beta_0 A}A\left\{A - \overline{m}(\boldsymbol{X})\right\} = \Theta(1) \quad \textit{and} \quad \mathbb{E}h^2(\boldsymbol{D}; \beta_0, \overline{\eta}) = \Theta(1).$$

4.1. High dimensional (parametric) setting

Let $\mathrm{expit}'(\cdot)$ be the derivative function of $\mathrm{expit}(\cdot)$, $\|\cdot\|_0$ represents the number of nonzero elements in a vector and $s = \max\{\|\boldsymbol{\gamma}^*\|_0, \|\bar{\boldsymbol{\gamma}}\|_0, \|\bar{\boldsymbol{\alpha}}\|_0\}$. We introduce the following assumptions to regularise the covariates and nuisance estimators.

Assumption HD1 (Model double robustness).—*At least one of the following conditions hold: (a) there exists* $\boldsymbol{\gamma}_0 \in \mathbb{R}^p$ *such that* $r_0(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\gamma}_0$ *and* $\boldsymbol{\gamma}^* = \bar{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0$; *(b) there exists* $\boldsymbol{\alpha}_0 \in \mathbb{R}^p$ *such that* $m_0(\boldsymbol{x}) = g\left(\boldsymbol{x}^\top \boldsymbol{\alpha}_0\right)$ *and* $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0$.

Assumption HD2 (Concentration rate).—*It holds that*

$$\left\| n^{-1} \sum_{i=1}^{n} (1 - Y_i)\mathrm{expit}\left(-\boldsymbol{X}_i^\top \boldsymbol{\gamma}^*\right)e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}}\left\{A_i - g\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right)\right\}\boldsymbol{X}_i \right\|_\infty = O_{\mathbb{P}}\left\{(\log p/n)^{1/2}\right\};$$

$$\left\| n^{-1} \sum_{i=1}^{n} \mathrm{expit}\left(-\boldsymbol{X}_i^\top \boldsymbol{\gamma}^*\right)g'\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right)\Psi_i \boldsymbol{X}_i \right\|_\infty = O_{\mathbb{P}}\left\{(\log p/n)^{1/2}\right\};$$

$$\left\| n^{-1} \sum_{i=1}^{n} \mathrm{expit}'\left(-\boldsymbol{X}_i^\top \boldsymbol{\gamma}^*\right)\left\{A_i - g\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right)\right\}\Psi_i \boldsymbol{X}_i \right\|_\infty = O_{\mathbb{P}}\left\{(\log p/n)^{1/2}\right\},$$

where $\Psi_i = Y_i e^{-\beta_0 A_i} - (1 - Y_i)e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}}$.

Assumption HD3 (smooth link function).—*There exists* $L = \Theta(1)$ *that for any* $u, v \in \mathbb{R}$,

$$|g'(u) - g'(v)| \leq L|u - v|.$$

Assumption HD4 (Risk of the L1-regularised estimators).$_1$-regularised estimators).—*There exists tuning parameters* $\lambda_\alpha, \lambda_\gamma = \Theta\left\{(\log p/n)^{1/2}\right\}$ *such that (3.1) and (3.2) have feasible solutions with probability approaching 1 and*

$$\sup_{i \in \{1, \ldots, n\}} \left|g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right)\right| = O_{\mathbb{P}}(1); \quad \|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 + \|\hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}\|_1 + \|\hat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\|_1 = O_{\mathbb{P}}\left\{s(\log p/n)^{1/2}\right\};$$

[6] To accommodate the notations of both HD and ML, we use $\overline{m}(\cdot)$ and $\overline{r}(\cdot)$ to represent the limiting models defined as $g\left(\boldsymbol{x}^\top \bar{\boldsymbol{\alpha}}\right)$ and $\boldsymbol{x}^\top \bar{\boldsymbol{\gamma}}$ under the HD setting and just the true models $m_0(\cdot)$ and $r_0(\cdot)$ under ML.

$$n^{-1}\sum_{i=1}^{n}\left\{1+e^{\boldsymbol{X}_i^\top\bar{\boldsymbol{\gamma}}}\right\}\left[\left\{\boldsymbol{X}_i^\top(\hat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\right\}^2+\left\{\boldsymbol{X}_i^\top(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\right\}^2\right]+\left(\hat{\beta}_{\mathrm{HD}}-\beta_0\right)^2=O_{\mathbb{P}}(s\log p/n);$$

$$n^{-1}\sum_{i=1}^{n}\left\{1+e^{\boldsymbol{X}_i^\top\bar{\boldsymbol{\gamma}}}\right\}\left[\left\{\boldsymbol{X}_i^\top(\hat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}})\right\}^2+\left\{g\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\alpha}}\right)-g\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)\right\}^2\right]=O_{\mathbb{P}}(s\log p/n).$$

Assumption HD5 (Ultra-sparsity).—*It holds that* $s=o\left(n^{1/2}/\log p\right)$.

Remark 4.1.—Under Assumption HD1 and our constructions (3.1) and (3.2) (or the one introduced in Appendix A4), the expectations of the terms to be concentrated in Assumption HD2 are $\mathbf{0}$ by Remark 3.1. Then their maximum norms can be controlled by $O_{\mathbb{P}}\left\{(\log p/n)^{1/2}\right\}$ as assumed in HD2, when the covariates $\boldsymbol{X}_i$ are bounded, sub-exponential, or beyond (Kuchibhotla and Chakrabortty, 2018), using the concentration results derived from the existing literature (Giné and Nickl, 2016).

Remark 4.2.—Rates of the prediction and estimation risk of the nuisance estimators in Assumption HD4 can be derived following the general theoretical framework for ℓ_1-regularised estimation introduced in Candes et al. (2007), Bickel et al. (2009), Bühlmann and Van de Geer (2011), and Negahban et al. (2012). The same rate properties have been used for analysing doubly robust estimators with HD nuisance models in existing literature (Smucler et al., 2019; Tan, 2020a; Dukes and Vansteelandt, 2020). Note that (3.1) and (3.2) involve the estimators $\tilde{\boldsymbol{\gamma}}$ or $\hat{\boldsymbol{\alpha}}$ being obtained beforehand. This will require some additional effort on removing the 'plug-in' errors of $\tilde{\boldsymbol{\gamma}}$ or $\hat{\boldsymbol{\alpha}}$ when deriving the risk rates for $\hat{\boldsymbol{\alpha}}$ or $\tilde{\boldsymbol{\gamma}}$ compared to the standard analysis procedures. See Tan (2020a) for a similar issue and the relevant technical details used to handle it. In addition, $\sup_{i\in\{1,\ldots,n\}}\left|g\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\alpha}}\right)\right|=O_{\mathbb{P}}(1)$ imposed in HD4 is not a standard assumption but is rather mild. This is because $\sup_{\boldsymbol{x}\in\mathcal{X}}\left|g\left(\boldsymbol{x}^\top\bar{\boldsymbol{\alpha}}\right)\right|=O(1)$ by Assumption REG and we only need $g\left(\boldsymbol{x}^\top\hat{\boldsymbol{\alpha}}\right)-g\left(\boldsymbol{x}^\top\bar{\boldsymbol{\alpha}}\right)$ to be $O_{\mathbb{P}}(1)$ uniformly.

Remark 4.3.—The ultra-sparsity assumption HD5 was also imposed in existing literature, including Tan (2020a) and Dukes and Vansteelandt (2020), to control the rate of bias incurred by the HD estimators: $O_{\mathbb{P}}(s\log p/n)$ below the parametric rate. For the linear nuisance model, existingworklikeZhuetal. (2018)andDukesandVansteelandt(2020)suggestedtoaddadditional moment (KKT) constraints to relax the ultra-sparsity assumption. However, their approach has not yet been shown to be feasible for nonlinear models, so, while promising, remains unclear for our framework.

We present the asymptotic property of $\hat{\beta}_{\mathrm{HD}}$ in Theorem 4.2 and its proof in Appendix A2.

Theorem 4.1.—*Denote by* $\bar{I}=\mathbb{E}\bar{\psi}(\boldsymbol{X})Ye^{-\beta_0 A}A\left\{A-\bar{m}(\boldsymbol{X})\right\}$ *and* $\bar{\sigma}^2=\bar{I}^{-2}\mathbb{E}h^2(\boldsymbol{D};\beta_0,\bar{\eta})$. *Under Assumptions REG and HD1–HD5, we have*

$$\sqrt{n}\bar{\sigma}^{-1}\left(\hat{\beta}_{\mathrm{HD}}-\beta_0\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\bar{\sigma}\bar{I})^{-1}h(\boldsymbol{D}_i;\beta_0,\bar{\eta})+o_{\mathbb{P}}(1),$$

which weakly converge to $\mathrm{N}(0,1)$.

As an earlier proposed and commonly used approach, logistic debiased LASSO (Van de Geer et al., 2014; Jankova and Van de Geer,´ 2016) has been criticised because its sparse inverse information matrix condition is not explainable and justifiable, leading to a subpar performance theoretically and numerically (see Xia et al., 2020). Interestingly, we find the model sparsity assumption of our method is more reasonable than debiased LASSO and present a simple comparison of these two approaches in Remark 4.4.

Remark 4.4.—*Assume* $\mathbb{P}(Y=1\mid A,\boldsymbol{X})=\mathrm{expit}\left\{\beta_0 A+\boldsymbol{X}^{\top}\boldsymbol{\gamma}_0\right\}$ is correctly specified. As is argued by Xia et al. (2020), assuming the information matrix of the logistic model has an ultra-sparse[7] inverse, it is crucial to ensure the desirable properties of the debiased LASSO estimator for β_0. However, this assumption is not explainable or convincing for the common Gaussian design with sparse precision matrix, due to the presence of the logistic canonical link. In comparison, we require that $\mathbb{E}(A\mid Y=0,\boldsymbol{X}=\boldsymbol{x})=g\left(\boldsymbol{X}_i^{\top}\boldsymbol{\alpha}_0\right)$ with $\|\boldsymbol{\alpha}_0\|_0=o\left(n^{1/2}/\log p\right)$, which has two advantages over debiased LASSO. First, it accommodates nonlinear link function $g(\cdot)$ and can be more reasonable for a categorical A. Second, it is imposed on a conditional model directly and is thus more explainable. For example, consider a conditional Gaussian model: $\left(A,\boldsymbol{X}^{\top}\right)^{\top}\mid\left\{Y=j\right\}\sim\mathcal{N}(\boldsymbol{\mu}_j,\boldsymbol{\Sigma})$ for $j=0,1$. Then we have $r_0(\boldsymbol{x})=\boldsymbol{x}^{\top}\boldsymbol{\gamma}_0$ where $\left(\beta_0,\boldsymbol{\gamma}_0^{\top}\right)^{\top}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)$, and $A\mid\boldsymbol{X},Y=0$ follows a Gaussian linear model with the coefficient $\boldsymbol{\alpha}_0$ determined by $\boldsymbol{\Sigma}^{-1}$. Therefore, our sparsity assumptions on $\boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_0$ actually assume the data generation parameters $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\mu}_1-\boldsymbol{\mu}_0$ to be sparse, which seems more explainable and verifiable in practice.

4.2. Machine learning (nonparametric) setting

Define $\|f(\cdot)\|_{Q,q}=:\|f(\boldsymbol{U})\|_{Q,q}=:\left\{\int|f(u)|^q dQ(u)\right\}^{1/q}$ for any real number $q>0$, function $f(\cdot)$, random variables $\boldsymbol{U}$ and probability measure Q. Let P denote the probability measure of the observed $\boldsymbol{D}$. We assume that $K=\Theta(1)$ and introduce the following assumption.

Assumption ML1 (Quality of the ML nuisance estimators).—*For each* k,

$$\begin{aligned}&\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{r}^{[-k]}(\boldsymbol{x})-r_0(\boldsymbol{x})\right|+\left|\hat{m}^{[-k]}(\boldsymbol{x})-m_0(\boldsymbol{x})\right|=o_{\mathbb{P}}(1);\\&\left\|\hat{r}^{-k]}(\cdot)-r_0(\cdot)\right\|_{P,2}+\left\|\hat{m}^{[-k]}(\cdot)-m_0(\cdot)\right\|_{P,2}=o_{\mathbb{P}}\left(n^{-1/4}\right).\end{aligned}$$

[7]Or approximately sparse (see recent work, e.g., Belloni et al., 2018; Ma et al., 2020; Liu et al., 2020).

Remark 4.5.—Similar to Assumptions 3.2 and 3.4 of Chernozhukov, Chetverikov, et al. (2018), our Assumption ML1 requires that the ML estimators for $r_0(\cdot)$ and $m_0(\cdot)$ are uniformly consistent and their mean squared errors (MSE) achieve the rate $O_{\mathbb{P}}\left(n^{-1/4}\right)$. This assumption is also referred to as rate double robustness in Smucler et al. (2019) as it requires production of the MSEs of $\hat{r}^{[-k]}(\cdot)$ and $\hat{m}^{[-k]}(\cdot)$ to be $o_{\mathbb{P}}\left(n^{-1/2}\right)$. In Appendix A1, we provide justification for our proposed FMR procedure to derive that the resulted $\hat{r}^{[-k]}(\cdot)$ satisfies Assumption ML1 as long as the learning algorithm $\mathscr{L}$ satisfies the same strong convergence properties as assumed in ML1 on all the learning tasks in FMR. Thus, FMR does not actually clip the wings of the ML algorithms being used in our framework.

We present the asymptotic property of $\hat{\beta}_{\mathrm{ML}}$ in Theorem 4.2 with its proof found in Appendix A3.

Theorem 4.2.—*Denote by* $I_0 = \mathbb{E}\overline{\psi}(\boldsymbol{X})Ye^{-\beta_0 A}A\{A - m_0(\boldsymbol{X})\}$; $\sigma_0^2 = I_0^{-2}\mathbb{E}h^2(\boldsymbol{D}; \beta_0, \eta_0)$. *Under Assumptions REG and ML1, we have*

$$\sqrt{n}\sigma_0^{-1}\left(\hat{\beta}_{\mathrm{ML}} - \beta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\sigma_0 I_0)^{-1}h(\boldsymbol{D}_i; \beta_0, \eta_0) + o_{\mathbb{P}}(1),$$

which weakly converge to $\mathrm{N}(0, 1)$.

Remark 4.6.—Given that $\overline{\psi}(\boldsymbol{x}) = \mathrm{expit}\{-r_0(\boldsymbol{x})\}$ and $\hat{\psi}^{[-k]}(\boldsymbol{x}) = \mathrm{expit}\left\{-\hat{r}^{[-k]}(\boldsymbol{x})\right\}$ in our ML case, one could show that $\hat{\psi}^{[-k]}(\boldsymbol{x})$ achieves the same strong convergence and rate properties as $\hat{r}^{[-k]}(\cdot)$ under Assumption ML1. While generally speaking, uniform consistency of $\hat{\psi}^{[-k]}(\cdot)$ is sufficient for the desirable conclusion in Theorem 4.2 so our framework accommodates more flexible choices on $\psi(\boldsymbol{x})$, for example, $\psi_{\mathrm{opt}}(\boldsymbol{x})$ as introduced in Section 3.3. We demonstrate this point during the proof of Theorem 4.2 in Appendix A3.

5. SIMULATION STUDY

We conduct simulation studies for our HD and ML settings separately in Sections 5.1 and 5.2, to study the point and interval estimation performance of our method.

5.1. High dimensional (parametric) setting

For the HD parametric setting, we design three data generation configurations introduced as follows to simulate different scenarios of model specification:

i. First, generate Y following $P(Y = 1) = 1/2$. Then generate $\left(A, \boldsymbol{X}^\top\right)^\top \mid \left\{Y = j\right\} \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. Specification of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}$ are presented in Appendix A5 such that $\beta_0 = 0.5$, $r_0(\boldsymbol{X}) = -0.22(X_1 + X_2) + 0.08(X_3 + X_4)$, and $m_0(\boldsymbol{X}) = -0.13(X_1 + X_2 + X_3 + X_4)$.

ii. Generate Y following $P(Y=1)=1/2$ and $\left(A, \boldsymbol{X}^{\top}\right)^{\top} \mid \left\{Y=j\right\} \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j=0,1$. Specification of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are presented in Appendix A5 such that $\beta_0=0.5$, $r_0(\boldsymbol{X})=-0.22(X_1+X_2)+0.08(X_3+X_4)-0.15(X_1X_2+X_1X_3+X_2X_3)$, and $m_0(\boldsymbol{X})=-0.13(X_1+X_2+X_3+X_4)$.

iii. First generate $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ given in Appendix A5. Then we generate A given $\boldsymbol{X}$ from a Gaussian linear model with unit variance and conditional mean

$$E(A \mid \boldsymbol{X}) = 0.15(X_1+X_2+X_3+X_4)+0.075(X_1X_2+X_1X_3+X_2X_3).$$

Finally, generate Y by

$$\mathbb{P}(Y=1 \mid A, \boldsymbol{X}) = \operatorname{expit}(0.5A+0.25X_1+0.25X_2+0.1X_3+0.1X_4).$$

We realise configurations (i)–(iii) with the sample size $n=1,000$, $1,500$ or $2,000$ separately and the dimension of $\boldsymbol{X}$ fixed as $p=200$. Under all these settings, we specify the nuisance models as: $r(\boldsymbol{X})=\boldsymbol{X}^{\top}\boldsymbol{\gamma}$ and $m(\boldsymbol{X})=\boldsymbol{X}^{\top}\boldsymbol{\alpha}$. Then both nuisance models are correctly specified under (i), only $m(\boldsymbol{X})$ is correctly specified under (ii), and only $r(\boldsymbol{X})$ is correct under (iii). Note that we cannot extract the explicit form of $m_0(\boldsymbol{x})$ under (iii) because A is generated conditional on $\boldsymbol{X}$ without fixing $Y=0$. But we still expect the linear model $m(\boldsymbol{X})=\boldsymbol{X}^{\top}\boldsymbol{\alpha}$ to be misspecified under (iii) due to the nonlinear terms in $E(A \mid \boldsymbol{X})$. Implementing details of our HD approach are presented in Appendix A4. Specifically, all the tuning parameters in ℓ_1-regularised regression are selected using cross-validation among the range $\left[0.2(\log p/n)^{1/2}, 2(\log p/n)^{1/2}\right]$. We conducted the regression in each setting with 300 repeated simulations.

Table 1 evaluates the performance of our estimator $\widehat{\beta}_{\mathrm{HD}}$ under all settings on its mean square error (MSE), absolute bias, and coverage probability (CP) of the 95% confidence interval (CI) estimated using bootstrap. Under all the settings, our method outputs low root-MSE and bias respectively being at most 18% and 7% of the magnitude of the true $\beta_0(=0.5)$ when $n=1,000$, and at most 12% and 4% of the β_0 when $n=2,000$. As the sample size n grows, one could see a trend of decaying on the MSEs and bias of our estimator. In addition, under all the settings, our interval estimation has proper CP locating in ±0.03 range of the nominal level 0.95. Thus, our HD estimator performs steadily well under different model specification scenarios as long as at least one nuisance model is correctly specified.

5.2. Machine learning (nonparametric) setting

To study our proposed method under the ML setting, welet $\boldsymbol{\Sigma} \in R^{p \times p}$ with $\Sigma_{ii}=1$, $\Sigma_{ij}=0.2$ for $i \neq j$, and truncate the randomly generated $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ vectors by $(-2, 2)$ to obtain the covariates $\boldsymbol{X}$. We then generate A from the Gaussian model given $\boldsymbol{X}$ with unit variance and conditional mean $a_0(\boldsymbol{X})=\zeta_a^{\top} f_a(\boldsymbol{X})$ where $f_a(\boldsymbol{x})$ is a nonlinear basis function of $\boldsymbol{x}$ including various types of effects (interaction, indicator, and trigonometric function, etc.), as defined

in Appendix A5 and ζ_a represents its loading coefficients also given in Appendix A5. Then $A = a_0(x) + e$ where e is a standard normal variable truncated into $(-2, 2)$. Finally, we set $\beta_0 = 1$, $r_0(\boldsymbol{X}) = \boldsymbol{\zeta}_r^\top f_r(\boldsymbol{X})$ with the nonlinear basis $f_r(\boldsymbol{X})$ (see Appendix A5) and generate Y following model (1.1). We fix $p = 20$ and set $n = 1,000$ and $2,000$ separately.

To estimate the nuisance function $r_0(\boldsymbol{x})$, we use the FMR procedure with its last step being (3.5). The number of fold for cross-fitting is set as $K = 5$. For choice of the learning algorithms $\mathscr{L}$, we consider four ML methods and a hybrid method of the ML estimators introduced as follows.

- **a.** Gradient boosted machines (GBM): an ensemble approach of classification and regressiontree (CART) using gradient boosting. Implemented by **R** package 'gbm' (Greenwell et al., 2020).
- **b.** Random forest (RF): ensemble of CART with bagging. Implemented with **R** package 'RandomForest' (Liaw and Wiener, 2002).
- **c.** Support vector machine (SVM): with linear kernel and implemented using **R** package 'e1071' (Dimitriadou et al., 2004).
- **d.** Neural network (NN): single hidden layer neural network implemented with **R** package 'nnet' (Ripley and Venables, 2016).
- **e.** Best nuisance models (Best): similar to Chernozhukov, Chetverikov, et al. (2018), for each nuisance component, we use a simple hybrid method choosing the ML estimator among (a)–(d) as the one with best prediction performance evaluated by the cross-validated sum-squared loss.

All the above-mentioned ML algorithms have been commonly used in recent years and considered in the literature of DML; for example, see Chernozhukov, Chetverikov, et al. (2018) and Cui and Tchetgen Tchetgen (2019). Tuning parameters of the ML models including the number of trees of GBM and RF, the margin of SVM, and the number of units and the weight decay of NN are selected using the resampling approach of **R** package 'caret' (Kuhn et al., 2020). We conducted the regression in each setting with 300 repeated simulations.

Table 2 presents the resulted average MSE, absolute bias and CP of 95% CI of $\hat{\beta}_{\mathrm{ML}}$ obtained with the five ML modelling strategies for $n = 1,000$ and $n = 2,000$ separately. The five approaches have relatively consistent performance in terms of MSE, bias, and CP under both settings, with the variation of their MSEs smaller than 0.015. This demonstrates that performance of our framework is robust to the choice of ML algorithms. While, to a certain degree, 'NN' has the best performance (with the lowest bias and MSE) when $n = 1,000$ and 'GBM' and 'Best' have the best performance when $n = 2,000$. Also, interval estimations of all the approaches achieve proper coverage rates, all of which are between 0.90 and 0.95.

6. A REAL EXAMPLE: EFFECT OF EC PILL ON EARLY FETAL GESTATION

In this section, we implement our proposed HD and ML methods to study the effect of the emergency contraceptive (EC) pill on the rate of new birth and early gestation fetal death (abortion), by revisiting and exploring the data of a quasi-experimental study based on the policy reform on the EC pill in Chile (Bentancor and Clarke, 2017). In the original study, the authors collected all records of birth and fetal deaths in Chile, as well as a number of municipality-level features (education, salary, and healthcare, etc.) of women at the reproductive age (15–34), in the years around 2008, during which the country was experiencing a reform relating to the legislation of EC pills. As a consequence of this, about half of the municipalities in Chile started to provide EC pill freely in 2009, while in the remaining half, the EC pill was not available or extremely restricted in use during that period. This policy was mostly dependent on the political, economic, and public health factors characterised by a total of 16 features (denoted as $\boldsymbol{Z}$) such as education spending, public health spending, condom use, and political conservativeness. Thus, the treatment of EC pill ($A = 1$ for EC pill accessible; $A = 0$ for EC pill not accessible) can be regarded as exogenous for the individuals.

Let $Y^{(1)}$ denote the indicator for the status of early gestation fetal death of each individual record and $Y^{(2)}$ indicate new births (pregnant and did not incur fetal death). Assume that

$$\mathbb{P}\left(Y^{(1)} = 1 \mid A, \boldsymbol{Z}\right) = \operatorname{expit}\left\{\beta_0^{(1)} A + r_0^{(1)}(\boldsymbol{Z})\right\};$$

$$\mathbb{P}\left(Y^{(2)} = 1 \mid A, \boldsymbol{Z}\right) = \operatorname{expit}\left\{\beta_0^{(2)} A + r_0^{(2)}(\boldsymbol{Z})\right\},$$

where $r_0^{(1)}(\cdot)$ and $r_0^{(2)}(\cdot)$ are two unknown functions. We are interested in inferring the two parameters $\beta_0^{(1)}$ and $\beta_0^{(2)}$ characterising the log odd ratios (log-OR) of abortion (among the pregnant individuals) and birth (among all individuals) to the treatment of the EC pill respectively. To investigate $\beta_0^{(1)}$ as the effect of the EC pill on abortion, we follow a similar strategy as Bentancor and Clarke (2017) that focuses on the individual records at the stage between 15 and 25, on which early gestation fetal death can be viewed as a reasonable proxy for illegal abortion. Note that the prevalence of $Y^{(1)}$ and $Y^{(2)}$ in their corresponding populations are both less than 5%, which could cause the logistic model to be unstable to fit. We randomly downsample the zeros in both analysis to make the prevalence of $Y^{(1)}$ and $Y^{(2)}$ 0.25 and 0.4, respectively. This procedure only changes intercepts of the logistic models and does not affect the target parameters. We find our results are not sensitive to the prevalence set for $Y^{(1)}$ and $Y^{(2)}$ as long as they are in a proper range, say 0.2 to 0.5. The resulting data set for analysing $\beta_0^{(1)}$ (abortion) has $n^{(1)} = 5,824$ samples. We take a subset with $n^{(2)} = 10,000$ samples for $\beta_0^{(2)}$ so that our algorithms will not require excessive computation time.

For our HD approach, we let $\boldsymbol{X}$ be the $p = 175$ dimensional basis joining $\boldsymbol{Z}$, all the interaction terms of $\boldsymbol{Z}$ and the three-dimensional natural splines of all the continuous

variables in $\boldsymbol{Z}$. We specify the nuisance functions as $m(\boldsymbol{x}) = \text{expit}\left(\boldsymbol{x}^\top \boldsymbol{\alpha}^{(\ell)}\right)$, and $r(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\gamma}^{(\ell)}$ for $\ell = 1, 2$. For the ML approach, we take $\boldsymbol{X} = \boldsymbol{Z}$ as the input covariates of the nuisance models. The choice and implementation of the ML algorithms are the same as in Section 5.2, except that we also use dropout[8] to avoid over-fitting due to the fact that most covariates are at the municipality level while the records are at the individual level.

Tables 3 and 4 present the point estimation, 95% CI and (two-side) p-values of our approaches $\beta_0^{(1)}$ and $\beta_0^{(2)}$, respectively. For $\beta_0^{(1)}$, point estimations of all methods are negative and around -0.18 ± 0.04. Their interval estimations are also internally consistent except that SVM outputs a relatively narrow CI and NN includes 0 near its CI upper bound. Correspondingly, all methods but NN reject the null '$\beta_0^{(1)} = 0$' at level 0.05. We note that NN outputs a slightly worse prediction model for $A \mid \boldsymbol{X}, Y = 0$, which causes it to produce a relatively wider CI. The result of our hybrid method 'Best' is very consistent with HD, indicating that our methods under both settings lead to basically the same conclusion. A similar situation occurs to the estimators of $\beta_0^{(2)}$. All methods reject '$\beta_0^{(2)} = 0$' at level 0.05 and their estimations are all negative values, and internally consistent on the magnitudes and CIs (SVM shows a moderate variation from other methods).

Our results reveal that distribution of the EC pill could significantly reduce the rate of illegal abortion (in the age group 15–25) and new births. This is consistent with the results of Bentancor and Clarke (2017) obtained through their municipality-level analysis. Although the estimated effect sizes are at different scales[9] and thus incomparable between the two studies, our p-values appear to show more significance in that nearly all of them are below 0.05 while their estimated p-values are between 0.05 and 0.1. This is because we use more complex and robust nuisance models to adjust for the confounding effects of $\boldsymbol{Z}$ and perform our analysis at the individual level, which enables us to have larger sample sizes.

ACKNOWLEDGEMENTS

The authors thank their adviser Tianxi Cai for helpful discussion and comments on this paper.

APPENDIX A1.: JUSTIFICATION OF THE FMR PROCEDURE

In this section, we derive error rates for the ML estimator $\hat{r}^{[-k]}(\cdot)$. obtained with the FMR procedure introduced in Section 3.2. Assume that the learning algorithm $\mathscr{L}$ attains the same strong convergence and rate properties as those for $\hat{m}^{[-k]}(\cdot)$ in Assumption ML1, i.e., for each $j \in \{1, 2, \ldots, K\}$ and $k \in \{1, 2, \ldots, K\}$:

[8] A common and flexible technique in ML research used for regularisation and avoiding overfitting. Here we randomly and independently set each entry of the training covariates matrix as $\mathrm{N}(0, 1)$ variable with probability 0.4 and 0.3, for the first and second study, respectively.

[9] Their effect is defined in a partially linear model of the abortion/birth rate against the treatment and control variables. While we are measuring the effect of the EC pill in a logistic model at the individual level.

$$\sup_{\boldsymbol{x}\in\mathcal{X},a\in\mathcal{A}}\left|\widehat{M}^{[-k,-j]}(a,\boldsymbol{x})-M_0(a,\boldsymbol{x})\right|=o_{\mathbb{P}}(1);\quad \left\|\widehat{M}^{[-k,-j]}(\cdot)-M_0(\cdot)\right\|_{P,2}=o_{\mathbb{P}}\left(n^{-1/4}\right);$$
$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{a}^{[-k,-j]}(\boldsymbol{x})-a_0(\boldsymbol{x})\right|=o_{\mathbb{P}}(1);\quad \left\|\hat{a}^{[-k,-j]}(\cdot)-a_0(\cdot)\right\|_{P,2}=o_{\mathbb{P}}\left(n^{-1/4}\right);$$
$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{t}^{[-k]}(\boldsymbol{x})-t_{\dagger}^{[-k,-j]}(\boldsymbol{x})\right|=o_{\mathbb{P}}(1);\quad \left\|\hat{t}^{[-k]}(\cdot)-t_{\dagger}^{[-k,-j]}(\cdot)\right\|_{P,2}=o_{\mathbb{P}}\left(n^{-1/4}\right).$$

where $t_{\dagger}^{[-k,-j]}(\boldsymbol{x})=:\mathbb{E}\left[\operatorname{logit}\left\{\widehat{M}^{[-k,-j]}(a,\boldsymbol{x})\right\}\mid \boldsymbol{X}=\boldsymbol{x},\widehat{M}^{[-k,-j]}(\cdot)\right]$. We justify as follows that

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{r}^{[-k]}(\boldsymbol{x})-r_0(\boldsymbol{x})\right|=o_{\mathbb{P}}(1);\quad \left\|\hat{r}^{[-k]}(\cdot)-r_0(\cdot)\right\|_{P,2}=o_{\mathbb{P}}\left(n^{-1/4}\right).$$

First, given that logit is a smooth function, it is not hard to show that

$$\sup_{\boldsymbol{x}\in\mathcal{X},a\in\mathcal{A}}\left|\operatorname{logit}\left\{\widehat{M}^{[-k,-j]}(a,\boldsymbol{x})\right\}-\operatorname{logit}\{M_0(a,\boldsymbol{x})\}\right|=o_{\mathbb{P}}(1);$$
$$\left\|\operatorname{logit}\left\{\widehat{M}^{[-k,-j]}(\cdot)\right\}-\operatorname{logit}\{M_0(\cdot)\}\right\|_{P,2}=o_{\mathbb{P}}\left(n^{-1/4}\right),$$

under some mild regularity conditions. Then derive the error rate of $\breve{\beta}^{[-k]}$ as follows.

$$\begin{aligned}
&|\mathcal{I}_{-k}|^{-1}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_{-k,j}}\operatorname{logit}\left\{\widehat{M}^{[-k,-j]}(A_i,\boldsymbol{X}_i)\right\}\left\{A_i-\hat{a}^{[-k,-j]}(\boldsymbol{X}_i)\right\}\\
&=|\mathcal{I}_{-k}|^{-1}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_{-k,j}}\left(\operatorname{logit}\{M_0(A_i,\boldsymbol{X}_i)\}\{A_i-a_0(\boldsymbol{X}_i)\}\right.\\
&\quad+\left[\operatorname{logit}\left\{\widehat{M}^{[-k,-j]}(A_i,\boldsymbol{X}_i)\right\}-\operatorname{logit}\{M_0(A_i,\boldsymbol{X}_i)\}\right]\{A_i-a_0(\boldsymbol{X}_i)\}\\
&\quad+\operatorname{logit}\{M_0(A_i,\boldsymbol{X}_i)\}\left\{a_0(\boldsymbol{X}_i)-\hat{a}^{[-k,-j]}(\boldsymbol{X}_i)\right\}\\
&\quad\left.+\left[\operatorname{logit}\left\{\widehat{M}^{[-k,-j]}(A_i,\boldsymbol{X}_i)\right\}-\operatorname{logit}\{M_0(A_i,\boldsymbol{X}_i)\}\right]\left\{a_0(\boldsymbol{X}_i)-\hat{a}^{[-k,-j]}(\boldsymbol{X}_i)\right\}\right)\\
&=|\mathcal{I}_{-k}|^{-1}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_{-k,j}}\operatorname{logit}\{M_0(A_i,\boldsymbol{X}_i)\}\{A_i-a_0(\boldsymbol{X}_i)\}\\
&\quad+\left\|\hat{a}^{[-k,-j]}(\cdot)-a_0(\cdot)\right\|_{P,2}+\left\|\operatorname{logit}\left\{\widehat{M}^{[-k,-j]}(\cdot)\right\}-\operatorname{logit}\{M_0(\cdot)\}\right\|_{P,2}+O_{\mathbb{P}}\left(n^{-1/2}\right)\\
&=|\mathcal{I}_{-k}|^{-1}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_{-k,j}}\operatorname{logit}\{M_0(A_i,\boldsymbol{X}_i)\}\{A_i-a_0(\boldsymbol{X}_i)\}+o_{\mathbb{P}}\left(n^{-1/4}\right)+O_{\mathbb{P}}\left(n^{-1/2}\right),
\end{aligned}$$

under some mild regularity conditions. Similarly, we have

$$\begin{aligned}
&|\mathcal{I}_{-k}|^{-1}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_{-k,j}}\left\{A_i-\hat{a}^{[-k,-j]}(\boldsymbol{X}_i)\right\}^2\\
&=|\mathcal{I}_{-k}|^{-1}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_{-k,j}}\{A_i-a_0(\boldsymbol{X}_i)\}^2+o_{\mathbb{P}}\left(n^{-1/4}\right)+O_{\mathbb{P}}\left(n^{-1/2}\right).
\end{aligned}$$

And consequently, by Proposition 3.1 and

$$\check{\beta}^{[-k]} = \frac{\sum_{i \in \mathcal{I}_{-k}} \text{logit}\{M_0(A_i, \boldsymbol{X}_i)\}\{A_i - a_0(\boldsymbol{X}_i)\}}{\sum_{i \in \mathcal{I}_{-k}} \{A_i - a_0(\boldsymbol{X}_i)\}^2} + o_{\mathbb{P}}\left(n^{-1/4}\right) = \beta_0 + o_{\mathbb{P}}\left(n^{-1/4}\right).$$

Then by Assumption REG that β_0 and $|a_0(\boldsymbol{x})|$ are bounded and recall that

$$\hat{a}^{[-k]}(\boldsymbol{x}) = K^{-1} \sum_{j=1}^{K} \hat{a}^{[-k,-j]}(\boldsymbol{x}),$$

the estimator $\hat{r}^{[-k]}(\cdot)$ given by equation (3.5) satisfies that:

$$\begin{aligned}
&\sup_{\boldsymbol{x} \in \mathcal{X}} \left|\hat{r}^{[-k]}(\boldsymbol{x}) - r_0(\boldsymbol{x})\right| \\
&\leq \sup_{\boldsymbol{x} \in \mathcal{X}} \left|\hat{t}^{[-k]}(\boldsymbol{x}) - t_0(\boldsymbol{x})\right| + \left|\check{\beta}^{[-k]} - \beta_0\right| \left|a_0(\boldsymbol{x})\right| + \left|\check{\beta}^{[-k]}\right| \left|\hat{a}^{[-k]}(\boldsymbol{x}) - a_0(\boldsymbol{x})\right| \\
&\leq \sup_{\boldsymbol{x} \in \mathcal{X}, j} \left|\hat{t}^{[-k]}(\boldsymbol{x}) - t_\dagger^{[-k,-j]}(\boldsymbol{x})\right| + \left|t_\dagger^{[-k,-j]}(\boldsymbol{x}) - t_0(\boldsymbol{x})\right| + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1),
\end{aligned}$$

where $\sup_{\boldsymbol{x} \in \mathcal{X}, j} \left|t_\dagger^{[-k,-j]}(\boldsymbol{x}) - t_0(\boldsymbol{x})\right| = o_{\mathbb{P}}(1)$ is a consequence of

$$\sup_{\boldsymbol{x} \in \mathcal{X}, a \in \mathcal{A}} \left|\text{logit}\left\{\widehat{M}^{[-k,-j]}(a, \boldsymbol{x})\right\} - \text{logit}\{M_0(a, \boldsymbol{x})\}\right| = o_{\mathbb{P}}(1).$$

And then

$$\begin{aligned}
&\left\|\hat{r}^{[-k]}(\cdot) - r_0(\cdot)\right\|_{P,2} \\
&\leq \left\|\hat{t}^{[-k]}(\cdot) - t_0(\cdot)\right\|_{P,2} + \left|\check{\beta}^{[-k]} - \beta_0\right| \left\|a_0(\boldsymbol{x})\right\|_{P,2} + \left|\check{\beta}^{[-k]}\right| \left\|\hat{a}^{[-k]}(\boldsymbol{x}) - a_0(\boldsymbol{x})\right\|_{P,2} \\
&= \max_{j \in \{1, \ldots, K\}} \left\|t_\dagger^{[-k,-j]}(\cdot) - t_0(\cdot)\right\|_{P,2} + o_{\mathbb{P}}\left(n^{-1/4}\right) \\
&= \max_{j \in \{1, \ldots, K\}} \left\|\mathbb{E}\left[\text{logit}\left\{\widehat{M}^{[-k,-j]}(A, \boldsymbol{X})\right\} \mid \boldsymbol{X} = \boldsymbol{x}\right] - \mathbb{E}[\text{logit}\{M_0(A, \boldsymbol{X})\} \mid \boldsymbol{X} = \boldsymbol{x}]\right\|_{P,2} \\
&\quad + o_{\mathbb{P}}\left(n^{-1/4}\right) \\
&\leq \max_{j \in \{1, \ldots, K\}} \left\|\text{logit}\left\{\widehat{M}^{[-k,-j]}(\cdot)\right\} - \text{logit}\{M_0(\cdot)\}\right\|_{P,2} + o_{\mathbb{P}}\left(n^{-1/4}\right) = o_{\mathbb{P}}\left(n^{-1/4}\right).
\end{aligned}$$

Thus, $\hat{r}^{[-k]}(\boldsymbol{x})$ satisfies Assumption ML1.

APPENDIX A2.: PROOF OF THEOREM 4.1

Proof.

By (3.2), we have

$$n^{-1} \sum_{i=1}^{n} h\left(\boldsymbol{D}_i; \hat{\beta}_{\text{HD}}, \hat{\eta}\right) = n^{-1} \sum_{i=1}^{n} \widehat{\psi}(\boldsymbol{X}_i)\left\{Y_i e^{-\hat{\beta}_{\text{HD}} A_i} - (1 - Y_i) e^{\boldsymbol{X}_i^\top \hat{\boldsymbol{\gamma}}}\right\}\left\{A_i - g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right)\right\} = 0.$$

Our main involvement is to remove the approximation error $n^{-1}\sum_{i=1}^{n} h\big(\boldsymbol{D}_i;\widehat{\beta}_{\mathrm{HD}},\widehat{\eta}\big)-h\big(\boldsymbol{D}_i;\widehat{\beta}_{\mathrm{HD}},\bar{\eta}\big)$, asymptotically. Note that

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n} h\big(\boldsymbol{D}_i;\widehat{\beta}_{\mathrm{HD}},\widehat{\eta}\big)-h\big(\boldsymbol{D}_i;\widehat{\beta}_{\mathrm{HD}},\bar{\eta}\big)\\
&\quad=n^{-1}\sum_{i=1}^{n}\widehat{\psi}(\boldsymbol{X}_i)(1-Y_i)\Big\{e^{\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\gamma}}}-e^{\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\gamma}}}\Big\}\big\{A_i-g\big(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\alpha}}\big)\big\}\\
&\quad+n^{-1}\sum_{i=1}^{n}\widehat{\psi}(\boldsymbol{X}_i)\Big\{Y_i e^{-\widehat{\beta}_{\mathrm{HD}}A_i}-(1-Y_i)e^{\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\gamma}}}\Big\}\big\{g\big(\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\alpha}}\big)-g\big(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\alpha}}\big)\big\}\\
&\quad+n^{-1}\sum_{i=1}^{n}\big\{\mathrm{expit}\big(-\boldsymbol{X}_i^{\top}\tilde{\boldsymbol{\gamma}}\big)-\mathrm{expit}\big(-\boldsymbol{X}_i^{\top}\boldsymbol{\gamma}^{*}\big)\big\}\\
&\quad\times\Big\{Y_i e^{-\widehat{\beta}_{\mathrm{HD}}A_i}-(1-Y_i)e^{\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\gamma}}}\Big\}\big\{A_i-g\big(\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\alpha}}\big)\big\}\\
&\quad=:\Delta_1+\Delta_2+\Delta_3.
\end{aligned}$$

We handle the terms Δ_1, Δ_2 and Δ_3 separately as follows. First, we have

$$\begin{aligned}
\Delta_1&=n^{-1}\sum_{i=1}^{n}(1-Y_i)\{\widehat{\psi}(\boldsymbol{X}_i)-\bar{\psi}(\boldsymbol{X}_i)\}e^{\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\gamma}}}\Big\{1-e^{\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})}\Big\}\big\{A_i-g\big(\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\alpha}}\big)\big\}\\
&\quad+n^{-1}\sum_{i=1}^{n}(1-Y_i)\widehat{\psi}(\boldsymbol{X}_i)e^{\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\gamma}}}\Big\{1-e^{\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})}\Big\}\big\{g\big(\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\alpha}}\big)-g\big(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\alpha}}\big)\big\}\\
&\quad+n^{-1}\sum_{i=1}^{n}(1-Y_i)\bar{\psi}(\boldsymbol{X}_i)e^{\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\gamma}}}\Big\{1-e^{\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})}-\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\Big\}\big\{A_i-g\big(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\alpha}}\big)\big\}\\
&\quad+n^{-1}\sum_{i=1}^{n}(1-Y_i)\bar{\psi}(\boldsymbol{X}_i)e^{\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\gamma}}}\big\{g\big(\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\alpha}}\big)-g\big(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\alpha}}\big)\big\}\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\\
&\quad+n^{-1}\sum_{i=1}^{n}(1-Y_i)\bar{\psi}(\boldsymbol{X}_i)e^{\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\gamma}}}\big\{A_i-g\big(\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\alpha}}\big)\big\}\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\\
&=:\Delta_{11}+\Delta_{12}+\Delta_{13}+\Delta_{14}+\Delta_{15}
\end{aligned}$$

As $\sup_{i\in\{1,\ldots,n\}}\big|\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\big|=O_{\mathbb{P}}(1)$ by Assumption HD4, there exists $M_1=O(1)$ such that with probability approaching 1,

$$\Big|1-e^{\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})}\Big|\le M_1\big|\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\big|,\quad \Big|1-e^{\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})}-\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\Big|\le M_1\big\{\boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\big\}^2; \tag{A1}$$

$$|\widehat{\psi}(\boldsymbol{X}_i)-\bar{\psi}(\boldsymbol{X}_i)|=\frac{e^{\boldsymbol{X}_i^{\top}\boldsymbol{\gamma}^{*}}\Big|1-e^{\boldsymbol{X}_i^{\top}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{*})}\Big|}{\Big(1+e^{\boldsymbol{X}_i^{\top}\boldsymbol{\gamma}^{*}}\Big)\Big(1+e^{\boldsymbol{X}_i^{\top}\tilde{\boldsymbol{\gamma}}}\Big)}\le M_1\big|\boldsymbol{X}_i^{\top}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{*})\big|. \tag{A2}$$

And by Assumptions REG and HD4, there exists $M_2=\Theta(1)$ that $\sup_{i\in\{1,\ldots,n\}}\big|A_i-g\big(\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\alpha}}\big)\big|+\ |\ A_i-g\big(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\alpha}}\big)\ |\ \le M_2$. Consequently, by Assumptions HD4 and boundness of $\psi(\cdot)$, we have

$$
\begin{aligned}
|\Delta_{11}| &\le n^{-1}\sum_{i=1}^{n} M_1^2 M_2 e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}} \left|\boldsymbol{X}_i^\top(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\right| \left\|\boldsymbol{X}_i^\top(\hat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\right| \\
&\le M_1^2 M_2 \left[n^{-2} \sum_{i=1}^{n} e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}} \left\{\boldsymbol{X}_i^\top(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\right\}^2 \sum_{i=1}^{n} e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}} \left\{\boldsymbol{X}_i^\top(\hat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\right\}^2 \right]^{1/2} \\
&= O_{\mathbb{P}}\left(\frac{s\log p}{n}\right); \\
|\Delta_{12}| &\le n^{-1}\sum_{i=1}^{n} M_1 e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}} \left|\boldsymbol{X}_i^\top(\hat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\right| \left|g\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right) - g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right)\right| \\
&\le M_1 \left[n^{-2} \sum_{i=1}^{n} e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}} \left\{\boldsymbol{X}_i^\top(\hat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\right\}^2 \sum_{i=1}^{n} e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}} \left\{g\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right) - g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right)\right\}^2 \right]^{1/2} \\
&= O_{\mathbb{P}}\left(\frac{s\log p}{n}\right); \\
|\Delta_{13}| &\le n^{-1}\sum_{i=1}^{n} M_1 M_2 e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}} \left\{\boldsymbol{X}_i^\top(\hat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\right\}^2 = O_{\mathbb{P}}\left(\frac{s\log p}{n}\right); \\
|\Delta_{14}| &\le n^{-1}\sum_{i=1}^{n} e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}} \left|\boldsymbol{X}_i^\top(\hat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})\right| \left|g\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right) - g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right)\right| = O_{\mathbb{P}}\left(\frac{s\log p}{n}\right), \text{ similar to } |\Delta_{12}|.
\end{aligned}
$$

By Assumptions HD2 and HD4,

$$
|\Delta_{15}| \le \left\| n^{-1}\sum_{i=1}^{n} (1-Y_i)\bar{\psi}(\boldsymbol{X}_i) e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}} \left\{A_i - g\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right)\right\} \boldsymbol{X}_i \right\|_\infty \cdot \| \hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}} \|_1 = O_{\mathbb{P}}\left(\frac{s\log p}{n}\right).
$$

Thus, we have $|\Delta_1| = O_{\mathbb{P}}(s\log p/n)$, For Δ_2, we have

$$
\begin{aligned}
\Delta_2 = {} & n^{-1}\sum_{i=1}^{n} \hat{\psi}(\boldsymbol{X}_i) Y_i \left(e^{-\hat{\beta}_{\mathrm{HD}} A_i} - e^{-\beta_0 A_i}\right) \left\{g\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right) - g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right)\right\} \\
& + n^{-1}\sum_{i=1}^{n} \left\{\hat{\psi}(\boldsymbol{X}_i) - \bar{\psi}(\boldsymbol{X}_i)\right\} \left\{Y_i e^{-\beta_0 A_i} - (1-Y_i) e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}}\right\} \left\{g\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right) - g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right)\right\} \\
& + n^{-1}\sum_{i=1}^{n} \bar{\psi}(\boldsymbol{X}_i) \left\{Y_i e^{-\beta_0 A_i} - (1-Y_i) e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}}\right\} \\
& \quad \times \left\{g\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right) - g\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}\right) - g'\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right) \boldsymbol{X}_i^\top (\hat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}})\right\} \\
& + n^{-1}\sum_{i=1}^{n} \bar{\psi}(\boldsymbol{X}_i) \left\{Y_i e^{-\beta_0 A_i} - (1-Y_i) e^{\boldsymbol{X}_i^\top \bar{\boldsymbol{\gamma}}}\right\} g'\left(\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}\right) \boldsymbol{X}_i^\top (\hat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}) \\
=: {} & \Delta_{21} + \Delta_{22} + \Delta_{23} + \Delta_{24}.
\end{aligned}
$$

Again using Assumptions REG and HD4, there exists $M_3 = \Theta(1)$ such that

$$
\left| e^{-\hat{\beta}_{\mathrm{HD}} A_i} - e^{-\beta_0 A_i} \right| \le M_3 \left| \hat{\beta}_{\mathrm{HD}} - \beta_0 \right|. \tag{A3}
$$

And by Assumption HD3 and the mean value theorem, for each i, there exits t_i lying between $\boldsymbol{X}_i^\top \hat{\boldsymbol{\alpha}}$ and $\boldsymbol{X}_i^\top \bar{\boldsymbol{\alpha}}$ such that

$$\begin{aligned}&\left|g\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)-g\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\alpha}}\right)-g'\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)\boldsymbol{X}_i^\top(\hat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}})\right|\\&\le\left|g'\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)-g'(t_i)\right|\left|\boldsymbol{X}_i^\top(\hat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}})\right|\le L\left\{\boldsymbol{X}_i^\top(\hat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}})\right\}^2.\end{aligned}\tag{A4}$$

These combined with (A1), (A2) and Assumptions REG and HD4 lead to that

$$\begin{aligned}|\Delta_{21}|&=O\left(\left|\hat{\beta}_{\mathrm{HD}}-\beta_0\right|\left[n^{-1}\sum_{i=1}^n\left\{g\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)-g\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\alpha}}\right)\right\}^2\right]^{1/2}\right)=O_{\mathbb{P}}\left(\frac{s\log p}{n}\right);\\|\Delta_{22}|&=O\left(\left[n^{-1}\sum_{i=1}^n\left\{1+e^{\boldsymbol{X}_i\bar{\boldsymbol{\gamma}}}\right\}\{\hat{\psi}(\boldsymbol{X}_i)-\bar{\psi}(\boldsymbol{X}_i)\}^2\right.\right.\\&\left.\left.\times n^{-1}\sum_{i=1}^n\left\{1+e^{\boldsymbol{X}_i\bar{\boldsymbol{\gamma}}}\right\}\left\{g\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)-g\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\alpha}}\right)\right\}^2\right]^{1/2}\right)=O_{\mathbb{P}}\left(\frac{s\log p}{n}\right);\\|\Delta_{23}|&=O\left(n^{-1}\sum_{i=1}^n\left\{1+e^{\boldsymbol{X}_i\bar{\boldsymbol{\gamma}}}\right\}\left\{\boldsymbol{X}_i^\top(\hat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}})\right\}^2\right)=O_{\mathbb{P}}\left(\frac{s\log p}{n}\right).\end{aligned}$$

And by Assumptions HD2 and HD4,

$$\begin{aligned}|\Delta_{24}|&\le\left\|n^{-1}\sum_{i=1}^n\bar{\psi}(\boldsymbol{X}_i)\left\{Y_ie^{-\beta_0A_i}-(1-Y_i)e^{\boldsymbol{X}_i^\top\bar{\boldsymbol{\gamma}}}\right\}g'\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)\boldsymbol{X}_i\right\|_\infty\cdot\|\hat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}}\|_1\\&=O_{\mathbb{P}}\left(\frac{s\log p}{n}\right).\end{aligned}$$

So we also have $|\Delta_2|=O_{\mathbb{P}}(s\log p/n)$. For Δ_3, we have

$$\begin{aligned}\Delta_3&=n^{-1}\sum_{i=1}^n\left\{\mathrm{expit}\left(-\boldsymbol{X}_i^\top\tilde{\boldsymbol{\gamma}}\right)-\mathrm{expit}\left(-\boldsymbol{X}_i^\top\boldsymbol{\gamma}^*\right)\right\}Y_i\left(e^{-\hat{\beta}_{\mathrm{HD}}A_i}-e^{-\beta_0A_i}\right)\left\{A_i-g\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)\right\}\\&\quad+n^{-1}\sum_{i=1}^n\left[\left\{\mathrm{expit}\left(-\boldsymbol{X}_i^\top\tilde{\boldsymbol{\gamma}}\right)-\mathrm{expit}\left(-\boldsymbol{X}_i^\top\boldsymbol{\gamma}^*\right)-\mathrm{expit}'\left(-\boldsymbol{X}_i^\top\boldsymbol{\gamma}^*\right)\boldsymbol{X}_i^\top(\boldsymbol{\gamma}^*-\tilde{\boldsymbol{\gamma}})\right\}\right.\\&\quad\left.\times\left\{Y_ie^{-\beta_0A_i}-(1-Y_i)e^{\boldsymbol{X}_i^\top\bar{\boldsymbol{\gamma}}}\right\}\left\{A_i-g\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)\right\}\right]\\&\quad+n^{-1}\sum_{i=1}^n\mathrm{expit}'\left(-\boldsymbol{X}_i^\top\boldsymbol{\gamma}^*\right)\boldsymbol{X}_i^\top(\boldsymbol{\gamma}^*-\tilde{\boldsymbol{\gamma}})\left\{Y_ie^{-\beta_0A_i}-(1-Y_i)e^{\boldsymbol{X}_i^\top\bar{\boldsymbol{\gamma}}}\right\}\left\{A_i-g\left(\boldsymbol{X}_i^\top\bar{\boldsymbol{\alpha}}\right)\right\}\\&=:\Delta_{31}+\Delta_{32}+\Delta_{33}.\end{aligned}$$

Again using the mean value theorem and the fact that $|\mathrm{expit}'(u)-\mathrm{expit}'(v)|\le|u-v|$ for any $u,v\in\mathbb{R}$ (by $|\mathrm{expit}''(\cdot)|\le1$), we have

$$\left|\mathrm{expit}\left(-\boldsymbol{X}_i^\top\tilde{\boldsymbol{\gamma}}\right)-\mathrm{expit}\left(-\boldsymbol{X}_i^\top\boldsymbol{\gamma}^*\right)-\mathrm{expit}'\left(-\boldsymbol{X}_i^\top\boldsymbol{\gamma}^*\right)\boldsymbol{X}_i^\top(\boldsymbol{\gamma}^*-\tilde{\boldsymbol{\gamma}})\right|\le\left\{\boldsymbol{X}_i^\top(\boldsymbol{\gamma}^*-\tilde{\boldsymbol{\gamma}})\right\}^2.$$

Then by (A2), (A3), and Assumptions REG and HD4, we have

$$|\Delta_{31}| = O\left(\left|\hat{\beta}_{\mathrm{HD}} - \beta_0\right|\left[n^{-1}\sum_{i=1}^{n}\left\{\boldsymbol{X}_i^{\top}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\right\}^2\right]^{1/2}\right) = O_{\mathbb{P}}\left(\frac{s\log p}{n}\right);$$
$$|\Delta_{32}| = O\left(n^{-1}\sum_{i=1}^{n}\left\{1 + e^{\boldsymbol{X}_i\bar{\boldsymbol{\gamma}}}\right\}\left\{\boldsymbol{X}_i^{\top}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\right\}^2\right) = O_{\mathbb{P}}\left(\frac{s\log p}{n}\right).$$

And by Assumptions HD2 and HD4,

$$|\Delta_{33}| \le \left\|n^{-1}\sum_{i=1}^{n}\mathrm{expit}'\left(-\boldsymbol{X}_i^{\top}\boldsymbol{\gamma}^*\right)\left\{Y_i e^{-\beta_0 A_i} - (1 - Y_i)e^{\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\gamma}}}\right\}\left\{A_i - g\left(\boldsymbol{X}_i^{\top}\bar{\boldsymbol{\alpha}}\right)\right\}\boldsymbol{X}_i\right\|_{\infty}$$
$$\times \left\|\boldsymbol{\gamma}^* - \tilde{\boldsymbol{\gamma}}\right\|_1 = O_{\mathbb{P}}\left(\frac{s\log p}{n}\right).$$

Thus, we have $|\Delta_3| = O_{\mathbb{P}}(s\log p/n)$, and by Assumption HD5,

$$n^{-1}\sum_{i=1}^{n} h\left(\boldsymbol{D}_i; \hat{\beta}_{\mathrm{HD}}, \hat{\eta}\right) - h\left(\boldsymbol{D}_i; \hat{\beta}_{\mathrm{HD}}, \bar{\eta}\right) = O_{\mathbb{P}}\left(\frac{s\log p}{n}\right) = o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right).$$

which leads to that

$$n^{-1}\sum_{i=1}^{n} h\left(\boldsymbol{D}_i; \hat{\beta}_{\mathrm{HD}}, \bar{\eta}\right) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) = 0.$$

This combined with Remark 2.1 that $\mathbb{E}h(\boldsymbol{D}; \beta_0, \bar{\eta}) = 0$ under Assumption HD1, the regularity Assumption REG, and Theorem 5.21 of Van der Vaart (2000), comes to the conclusion of Theorem 4.1. □

APPENDIX A3.: PROOF OF THEOREM 4.2

Following the general results of the DML estimator with nonlinear Neyman orthogonal score presented in Section 3.3 and Theorem 3.3 of Chernozhukov, Chetverikov, et al. (2018), we only need to verify their Assumptions 3.3 and 3.4 on our score function $h(\boldsymbol{D}; \beta, \eta)$, specifically, Assumptions A1 and A2 presented as follows.

Assumption A1 (Moment condition with neyman orthogonality).

It holds: (a) $\mathbb{E}h(\boldsymbol{D}; \beta_0, \eta_0) = 0$ *and* $\mathscr{B}$ *contains an interval of length* $\Theta\left(n^{-1/2}\log n\right)$ *centred at* β_0; *(b) the map* $(\beta, \eta) \to \mathbb{E}h(\boldsymbol{D}; \beta_0, \eta_0)$ *is twice continuously Gateaux-differentiable; (c)* $|\mathbb{E}h(\boldsymbol{D}; \beta, \eta_0)| \ge \min\{\,|\,J_0(\beta - \beta_0)\,|\,, c_0\}$ *where the parameters* $\eta_0 = \{r_0(\cdot), m_0(\cdot), \bar{\psi}\}$, $c_0 = \Theta(1)$ *and* $J_0 = \partial_\beta \mathbb{E}h(\boldsymbol{D}; \beta, \eta_0)|_{\beta=\beta_0} = \Theta(1)$; *(d)* $h(\boldsymbol{D}; \beta_0, \eta_0)$ *obeys Neyman orthogonality, i.e.,* $\partial_\eta \mathbb{E}h(\boldsymbol{D}; \beta_0, \eta_0)[\eta - \eta_0] = 0$ *for all* $\eta \in \mathscr{E}$ *where the parameter space of* η: $\mathscr{E} \subseteq \{\eta : \mathbb{E}|h(\boldsymbol{D}; \beta_0, \eta_0)[\eta - \eta_0]| < \infty\}$.

Assumption A2 (Quality of the nuisance estimators).

It holds: (a) $\hat{\eta}^{[-k]}$ *belongs to the realisation set* $\mathcal{T}_n$ *for each* $k \in \{1, 2, \ldots, K\}$, *with probability approaching 1 where* $\mathcal{T}_n$ *satisfies* $\eta_0 \in \mathcal{T}_n$ *and conditions given as follows; (b) The space of* β, $\mathcal{B}$ *is bounded and for each* $\eta \in \mathcal{T}_n$, *the functional space* $\mathcal{F}_\eta = \{h(\cdot\,; \beta, \eta) : \beta \in \mathcal{B}\}$ *is measurable and its uniform covering number satisfies: there exists positive constant* $R = \Theta(1)$ *and* $v = \Theta(1)$ *such that*

$$\sup_Q \log \mathcal{N}(\epsilon \|F_\eta(\cdot)\|_{Q,2}, \mathcal{F}_\eta, \|\cdot\|_{Q,2}) \leq v \log(R/\epsilon), \quad \forall \epsilon \in (0, 1],$$

where $F_\eta(\cdot)$ *is a measurable envelope function for* $\mathcal{F}_\eta$: $F_\eta(\boldsymbol{D}) \geq |h(\boldsymbol{D}; \beta, \eta)|$ *for all* $\boldsymbol{D}$ *and* $\beta \in \mathcal{B}$, *and there exists* $q > 2$ *such that* $\|F_\eta(\cdot)\|_{P,q} = O(1)$; *(c) there exists sequence* τ_n:

$$\begin{aligned}
&\sup_{\eta = \{r, m, \psi\} \in \mathcal{T}_n, \beta \in \mathcal{B}} |\mathbb{E}h(\boldsymbol{D}; \beta, \eta) - \mathbb{E}h(\boldsymbol{D}; \beta, \{r_0, m_0, \psi\})| = o(\tau_n), \\
&\sup_{\eta \in \mathcal{T}_n, |\beta - \beta_0| \leq \tau_n} \mathbb{E}[h(\boldsymbol{D}; \beta, \eta) - h(\boldsymbol{D}; \beta_0, \{r_0, m_0, \psi\})]^2 \\
&\qquad + \mathbb{E}[h(\boldsymbol{D}; \beta_0, \{r_0, m_0, \psi\}) - h(\boldsymbol{D}; \beta_0, \eta_0)]^2 = o(1), \\
&\sup_{r \in (0,1), \eta \in \mathcal{T}_n, |\beta - \beta_0| \leq \tau_n} \partial_r^2 \mathbb{E}h\{\boldsymbol{D}; \beta_0 + r(\beta - \beta_0), \eta_0 + r(\eta - \{r_0(\cdot), m_0(\cdot), \psi\})\} = o\left(n^{-1/2}\right);
\end{aligned}$$

and (d) $\mathbb{E}h^2(\boldsymbol{D}; \beta_0, \eta_0) = \Theta(1)$.

We simplified and adapted the original assumptions in Chernozhukov, Chetverikov, et al. (2018) to form Assumptions A1–A2, according to our own setting. The only nontrivial change made here is that in Assumption A2 (c), we require

$$\begin{aligned}
&\sup_{\eta = \{r, m, \psi\} \in \mathcal{T}_n, \beta \in \mathcal{B}} |\mathbb{E}h(\boldsymbol{D}; \beta, \eta) - \mathbb{E}h(\boldsymbol{D}; \beta, \{r_0, m_0, \psi\})| = o(\tau_n), \\
&\sup_{r \in (0,1), \eta \in \mathcal{T}_n, |\beta - \beta_0| \leq \tau_n} \partial_r^2 \mathbb{E}h\{\boldsymbol{D}; \beta_0 + r(\beta - \beta_0), \eta_0 + r(\eta - \{r_0(\cdot), m_0(\cdot), \psi\})\} = o\left(n^{-1/2}\right);
\end{aligned} \tag{A5}$$

instead of:

$$\begin{aligned}
&\sup_{\eta \in \mathcal{T}_n, \beta \in \mathcal{B}} |\mathbb{E}h(\boldsymbol{D}; \beta, \eta) - \mathbb{E}h(\boldsymbol{D}; \beta, \eta_0)| = o(\tau_n), \\
&\sup_{r \in (0,1), \eta \in \mathcal{T}_n, |\beta - \beta_0| \leq \tau_n} |\partial_r^2 \mathbb{E}h\{\boldsymbol{D}; \beta_0 + r(\beta - \beta_0), \eta_0 + r(\eta - \eta_0)\}| = o\left(n^{-1/2}\right),
\end{aligned}$$

as used in Assumption 3.4 (c) of Chernozhukov, Chetverikov, et al. (2018). The first inequality of (A5) is used by Chernozhukov, Chetverikov, et al. (2018) to derive a preliminary rate for the DML estimator: $|\hat{\beta}_{\mathrm{ML}} - \beta_0| = o_{\mathbb{P}}(\tau_n)$(see their Step 1 of the proof of Lemma 6.3), and the second inequality of (A5) is used in their Step 3 the proof of Lemma 6.3 to process the second order error of

$$n^{-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} h(\boldsymbol{D}_i;\beta_0,\eta_0) - h\left(\boldsymbol{D}_i;\beta,\hat{\eta}^{[-k]}\right)$$

uniformly for β satisfying $|\beta-\beta_0|\le\tau_n$.

Note that our modified two assumptions are still sufficient for deriving these results. Given that $\mathbb{E}h(\boldsymbol{D};\beta_0,\{r_0,m_0,\psi\})=0$ holds for all ψ (see Remark 2.1), there is actually no need to consider $\mathbb{E}h(\boldsymbol{D};\beta,\{r_0,m_0,\psi\})-\mathbb{E}h(\boldsymbol{D};\beta,\{r_0,m_0,\psi_0\})$ when deriving $\left|\hat{\beta}_{\mathrm{ML}}-\beta_0\right|=o_{\mathbb{P}}(\tau_n)$. While for the Step 3 of Chernozhukov, Chetverikov, et al. (2018), one can instead handle the second order error of

$$n^{-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} h\left(\boldsymbol{D}_i;\beta_0,\left\{r_0,m_0,\hat{\psi}^{[-k]}\right\}\right) - h\left(\boldsymbol{D}_i;\beta,\hat{\eta}^{[-k]}\right),$$

again using that $\mathbb{E}h(\boldsymbol{D};\beta_0,\{r_0,m_0,\psi\})=0$ holds for all ψ, and then remove the remaining error:

$$n^{-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} h(\boldsymbol{D}_i;\beta_0,\{r_0,m_0,\psi_0\}) - h\left(\boldsymbol{D}_i;\beta_0,\left\{r_0,m_0,\hat{\psi}^{[-k]}\right\}\right),$$

through concentration based on $\partial_\psi\mathbb{E}h(\boldsymbol{D};\beta_0,\eta_0)[\psi-\psi_0]=0$ and the second inequality of Assumption A2 (c): $\sup_{\eta\in\mathcal{T}_n,|\beta-\beta_0|\le\tau_n}\mathbb{E}[h(\boldsymbol{D};\beta_0,\{r_0,m_0,\psi\})-h(\boldsymbol{D};\beta_0,\eta_0)]^2=o(1)$.

Alternatively, this modification essentially reduces our requirement on the quality of $\hat{\psi}^{[-k]}(\cdot)$, as mentioned in Remark 4.6. As will be seen from the proof below, to fulfil the modified Assumption A2 (c), we do not require $\left\|\hat{\psi}^{[-k]}(\cdot)-\overline{\psi}(\cdot)\right\|_{P,2}=o_{\mathbb{P}}(\tau_n)$ as will happen when following the original version of Chernozhukov, Chetverikov, et al. (2018), but only needs $\hat{\psi}^{[-k]}(\cdot)$ to be uniformly consistent (though the former one may still be justifiable in our case because we take $\overline{\psi}(\boldsymbol{x})=\mathrm{expit}\{-\bar{r}(\boldsymbol{x})\}$). We now verify Assumptions A1 and A2 based on our Assumptions REG and ML1.

Proof.

Assumption A1 (a) is directly given by our logistic partial model assumption 1.1. Assumption A1 (b) is naturally satisfied as $h(\boldsymbol{D};\beta,\eta_0+r(\eta-\eta_0))$ and is a twice continuously differentiable in (β,r). Assumption A1 (c) is directly given by Assumption REG. And Assumption A1 (d) holds by equation (2.1), combined with the model assumptions (1.1) and $\mathbb{E}[A\mid\boldsymbol{X}=\boldsymbol{x},Y=0]=m_0(\boldsymbol{x})$.

By Assumption ML1 and $\overline{\psi}(\boldsymbol{x})=\mathrm{expit}(-\bar{r}(\boldsymbol{x}))$, there exists $\zeta_{1,n}=o(1)$ and $\zeta_{2,n}=o\left(n^{-1/4}\right)$ such that $\hat{\eta}^{[-k]}$ belongs to

$$\mathscr{T}_n =: \Big\{\eta = (r, m, \psi): \sup_{x \in \mathscr{X}} \Big|\psi(x) - \overline{\psi}(x)\Big| + |r(x) - r_0(x)| + |m(x) - m_0(x)| \le \zeta_{1,n}, \text{ and } \|r(\cdot) - r_0(\cdot)\|_{P,2} + \|m(\cdot) - m_0(\cdot)\|_{P,2} \le \zeta_{2,n}\},$$

with probability approaching 1, for each $k \in \{1, \ldots, K\}$. We define $\mathscr{T}_n$ of Assumption A2 in this way such that A2 (a) is satisfied. Now we validate Assumption A2 (b). By Assumption REG that A and β belong to compact sets, and $|m(x)| \le m_0(x) + o(1)$ is uniformly bounded for $\eta = \{r, m, \psi\} \in \mathscr{T}_n$, there exists positive $C_1 = \Theta(1)$ (1) such that for $\eta = \{r, m, \psi\} \in \mathscr{T}_n$,

$$\begin{aligned} h(\boldsymbol{D}; \beta, \eta) &= \psi(\boldsymbol{X})\{Ye^{-\beta A} - (1 - Y)e^{r(\boldsymbol{X})}\}\{A - m(\boldsymbol{X})\} \\ &\le \left|\psi(\boldsymbol{X})Ye^{-\beta A}\{A - m(\boldsymbol{X})\}\right| + \left|\psi(\boldsymbol{X})e^{r(\boldsymbol{X})}(1 - Y)\{A - m(\boldsymbol{X})\}\right| \\ &\le C_1\{\psi(\boldsymbol{X}) + \psi(\boldsymbol{X})e^{r(\boldsymbol{X})}\} = C_1\{\mathrm{expit}(-\psi(\boldsymbol{X})) + \mathrm{expit}(\psi(\boldsymbol{X}))\} \\ &\le C_1 + 1; \\ \partial_\beta h(\boldsymbol{D}; \beta, \eta)^2_{P,2} &= \mathbb{E}\left[\psi(\boldsymbol{X})Ye^{-\beta A}A\{A - m(\boldsymbol{X})\}\right]^2 \le C_1. \end{aligned}$$

Then by Example 19.7 of Van der Vaart (2000), Assumption A2 (b) holds with $v = 1$ and R being the diameter of $\mathscr{B}$. Note that Assumption A2 (d) is again directly given by Assumption REG. It remains to verify Assumption A2 (c). For each $\eta = \{r, m, \psi\} \in \mathscr{T}_n$ and $\beta \in \mathscr{B}$, using the boundness of β, A, $m_0(x)\psi(x)$ and $\psi(x)e^{r(x)}$, there exists $C_2 = \Theta(1)$ such that

$$\begin{aligned} &|\mathbb{E}h(\boldsymbol{D}; \beta, \eta) - \mathbb{E}h(\boldsymbol{D}; \beta, \{r_0, m_0, \psi\})| \\ &\le \left|\mathbb{E}\psi(\boldsymbol{X})Ye^{-\beta A}\{m_0(\boldsymbol{X}) - m(\boldsymbol{X})\}\right| + \left|\mathbb{E}\psi(\boldsymbol{X})(1 - Y)e^{r(\boldsymbol{X})}\{m_0(\boldsymbol{X}) - m(\boldsymbol{X})\}\right| \\ &\quad + \left|\mathbb{E}\psi(\boldsymbol{X})e^{r(\boldsymbol{X})}(1 - Y)\{1 - e^{r_0(\boldsymbol{X}) - r(\boldsymbol{X})}\}\{A - m(\boldsymbol{X}_0)\}\right| \\ &\le C_2\|(m_0(\boldsymbol{X}) - m(\boldsymbol{X})\|_{P,2} + \|r_0(\boldsymbol{X}) - r(\boldsymbol{X})\|_{P,2} + \|r_0(\boldsymbol{X}) - r(\boldsymbol{X})\|^2_{P,2}) \le 3C_2\zeta_{2,n}. \end{aligned}$$

So we take $\tau_n = n^{-1/4}$ and by $\zeta_{2,n} = o(n^{-1/4})$, the first inequality of Assumption A2 (c) is satisfied. Again by the boundness of β, A, $m_0(x)\psi(x)$ and $\psi(x)e^{r(x)}$; and $\zeta_{1,n}$, $\tau_n = o(1)$, there exists $C_3 = \Theta(1)$ such that

$$\begin{aligned} &\mathbb{E}[h(\boldsymbol{D}; \beta, \eta) - h(\boldsymbol{D}; \beta_0, \{r_0, m_0, \psi\})]^2 + \mathbb{E}[h(\boldsymbol{D}; \beta_0, \{r_0, m_0, \psi\}) - h(\boldsymbol{D}; \beta_0, \eta_0)]^2 \\ &\le \mathbb{E}[h(\boldsymbol{D}; \beta, \eta) - h(\boldsymbol{D}; \beta_0, \eta)]^2 + \mathbb{E}[h(\boldsymbol{D}; \beta_0, \eta) - h(\boldsymbol{D}; \beta_0, \{r_0, m_0, \psi\})]^2 \\ &\quad + \mathbb{E}[h(\boldsymbol{D}; \beta_0, \{r_0, m_0, \psi\}) - h(\boldsymbol{D}; \beta_0, \eta_0)]^2 \\ &\le \mathbb{E}\left[\psi(\boldsymbol{X})e^{-\beta_0 A}\{e^{(\beta_0 - \beta)A} - 1\}\{A - m(\boldsymbol{X})\}\right]^2 \\ &\quad + \mathbb{E}\left[\psi(\boldsymbol{X})e^{-\beta_0 A}\{m_0(\boldsymbol{X}) - m(\boldsymbol{X})\}\right]^2 + \mathbb{E}\left[\psi(\boldsymbol{X})e^{r(\boldsymbol{X})}\{m_0(\boldsymbol{X}) - m(\boldsymbol{X})\}\right]^2 \\ &\quad + \mathbb{E}\left[\psi(\boldsymbol{X})e^{r(\boldsymbol{X})}(1 - Y)\{1 - e^{r_0(\boldsymbol{X}) - r(\boldsymbol{X})}\}\{A - m_0(\boldsymbol{X})\}\right]^2 \\ &\quad + \mathbb{E}\left[\left|\overline{\psi}(\boldsymbol{X}) - \psi(\boldsymbol{X})\right|\{e^{-\beta_0 A} + e^{r_0(\boldsymbol{X})}\}|A - m_0(\boldsymbol{X})|\right]^2 \\ &\le C_3 \sup_{a \in \mathscr{A}} \left|e^{(\beta_0 - \beta)a} - 1\right| + C_3 \sup_{x \in \mathscr{X}} \Big\{\left|\left(1 + e^{r(x)}\right)[\overline{\psi}(x) - \psi(x)]\right| \\ &\qquad + |m(x) - m_0(x)| + \left|e^{r(x) - r_0(x)} - 1\right|\Big\} \\ &\le C_3 \sup_{a \in \mathscr{A}} \left|e^{(\beta_0 - \beta)a} - 1\right| + C_3 \sup_{x \in \mathscr{X}} \Big\{\left|\overline{\psi}(x) - \psi(x)\right| \\ &\qquad + |m(x) - m_0(x)| + \{\mathrm{expit}(r_0(x)) + 1\}\left|e^{r(x) - r_0(x)} - 1\right|\Big\} \\ &\le C_3\left\{\left(e^{\tau_n C_3} - 1\right) + 2\zeta_{1,n} + 2\left|e^{\zeta_{1,n}} - 1\right|\right\} = o(1), \end{aligned}$$

which validates the second inequality of Assumption A2 (c). At last, for each $r \in (0, 1)$, denote by $\beta^* = \beta_0 + r(\beta - \beta_0)$, $\eta^* = \{r^*, m^*, \psi\} = \{r_0(\cdot), m_0(\cdot), \psi\} + r(\eta - \{r_0(\cdot), m_0(\cdot), \psi\})$. Similar as the above deduction, we have that there exists $C_4 = \Theta(1)$,

$$\begin{aligned}
&\partial_r^2 \mathbb{E} h\{\boldsymbol{D}; \beta_0 + r(\beta - \beta_0), \eta_0 + r(\eta - \{r_0(\cdot), m_0(\cdot), \psi\})\} \\
&= \mathbb{E}\psi(\boldsymbol{X}) Y e^{-\beta^* A} A(\beta - \beta_0)\{m_0(\boldsymbol{X}) - m(\boldsymbol{X})\} \\
&\quad + \mathbb{E}\psi(\boldsymbol{X})(1 - Y) e^{r^*(\boldsymbol{X})}\{r_0(\boldsymbol{X}) - r(\boldsymbol{X})\}\{m_0(\boldsymbol{X}) - m(\boldsymbol{X})\} \\
&\le C_4 |\beta - \beta_0| \cdot \mathbb{E}|m_0(\boldsymbol{X}) - m(\boldsymbol{X})| + C_4 \mathbb{E}|r_0(\boldsymbol{X}) - r(\boldsymbol{X})| \cdot \mathbb{E}|m_0(\boldsymbol{X}) - m(\boldsymbol{X})| \\
&= O\left(\|m(\cdot) - m_0(\cdot)\|_{P,2}^2\right) + O\left\{(\beta - \beta_0)^2\right\} + O\left(\|r(\cdot) - r_0(\cdot)\|_{P,2}^2\right) \\
&= O\left(\zeta_{2,n}^2\right) + o\left(\tau_n^2\right) = o\left(n^{-1/4}\right).
\end{aligned}$$

□

Using the verified Assumptions A1 and A2, one can follow nearly the same proof procedures as those of Theorem 3.3 and Lemma 6.3 in Chernozhukov, Chetverikov, et al. (2018) to prove our Theorem 4.2. The only minor difference concerning the processing of $\psi^{[-k]}$ has been presented as above. As we point out, one can handle this smoothly by first considering $n^{-1}\sum_{k=1}^{K}\sum_{i \in \mathcal{J}_k} h\left(\boldsymbol{D}_i; \beta_0, \left\{r_0, m_0, \widehat{\psi}^{[-k]}\right\}\right)$ when deriving the as initial rate and asymptotic expansion of $\widehat{\beta}_{\mathrm{ML}}$ as $\mathbb{E}\left[h\left(\boldsymbol{D}_i; \beta_0, \left\{r_0, m_0, \widehat{\psi}^{[-k]}\right\}\right) \mid \widehat{\psi}^{[-k]}\right] = 0$, and finally concentrate $n^{-1}\sum_{k=1}^{K}\sum_{i \in \mathcal{J}_k} h(\boldsymbol{D}_i; \beta_0, \{r_0, m_0, \psi_0\}) - h\left(\boldsymbol{D}_i; \beta_0, \left\{r_0, m_0, \widehat{\psi}^{[-k]}\right\}\right)$ using that $\partial_\psi \mathbb{E} h(\boldsymbol{D}; \beta_0, \eta_0)[\psi - \psi_0] = 0$ and $\widehat{\psi}^{[-k]}(\cdot)$ is uniformly consistent.

APPENDIX A4.: NUMERICAL IMPLEMENTATION OF THE HD APPROACH

We present and demonstrate the implementation procedure of our HD approach mentioned in Remark 3.3 that uses LASSO instead of the Dantzig equation and modifies the construction procedures to make it solvable using the **R** packages *glmnet* and *RCAL*. Let $G(u) = \int g(u)du$, and recall that $\tilde{\boldsymbol{\gamma}}$ is some initial estimator obtained through ℓ_1-regularised logistic regression for Y versus $\{A, \boldsymbol{X}\}$ and $\widehat{\psi}(\boldsymbol{x}) = \mathrm{expit}\left(-\boldsymbol{x}^\top \tilde{\boldsymbol{\gamma}}\right)$. Then we fit

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \; n^{-1} \sum_{i=1}^{n} (1 - Y_i)\widehat{\psi}(\boldsymbol{X}_i) e^{\boldsymbol{X}_i^\top \tilde{\boldsymbol{\gamma}}}\left\{-A_i \boldsymbol{X}_i^\top \boldsymbol{\alpha} + G\left(\boldsymbol{X}_i^\top \boldsymbol{\alpha}\right)\right\} + \lambda_\alpha \| \boldsymbol{\alpha} \|_1, \tag{A6}$$

to obtain $\widehat{\boldsymbol{\alpha}}$. It is not hard to see that the KKT (or subgradient) condition of (A6) is equivalent to the ℓ_∞-constraint in (3.1). When the link function of $g(\cdot)$ is identity (liner model) or $\mathrm{expit}(\cdot)$ (logistic model), can be solved using the **R** package *glmnet* with proper specification of the sample weights, i.e., $(1 - Y_i)\widehat{\psi}(\boldsymbol{X}_i) e^{\boldsymbol{X}_i^\top \tilde{\boldsymbol{\gamma}}}$. Then we solve

$$n^{-1} \sum_{i=1}^{n} \widehat{\psi}(\boldsymbol{X}_i)\left\{Y_i e^{-\beta A_i} - (1 - Y_i) e^{\boldsymbol{X}_i^\top \tilde{\boldsymbol{\gamma}}}\right\}\left\{A_i - g\left(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\alpha}}\right)\right\} = 0,$$

to obtain a preliminary estimator $\tilde{\beta}$. It can be shown that when either $r(\boldsymbol{x})$ or $m(\boldsymbol{x})$ is correctly specified, the estimator $\tilde{\beta}$ should approach β_0 at the rate $O_{\mathbb{P}}\left\{(s\log p/n)^{1/2}\right\}$, i.e., the ℓ_2 errors of $\tilde{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$. So $\tilde{\beta}$ provides a good enough approximation of β_0 that can be used for the ℓ_1-regularised (weighted) calibration regression (Tan, 2020b) to estimate $\boldsymbol{\gamma}$:

$$\min_{\beta\in\mathbb{R},\eta\in\mathbb{R}^p} n^{-1}\sum_{i=1}^{n}\hat{\psi}(\boldsymbol{X}_i)e^{\boldsymbol{X}_i^\top\tilde{\boldsymbol{\gamma}}}g'\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\alpha}}\right)\left\{Y_i e^{-\tilde{\beta}A_i-\boldsymbol{X}_i^\top\boldsymbol{\gamma}}+(1-Y_i)\left(\tilde{\beta}A_i+\boldsymbol{X}_i^\top\boldsymbol{\gamma}\right)\right\}+\lambda_\gamma\|\boldsymbol{\gamma}\|_1. \tag{A7}$$

Similarly, the KKT condition of (A7) corresponds to the ℓ_∞-constraints in (3.2), though they are not always imposing the same moment conditions: when the nuisance model $r(\boldsymbol{x})$ is misspecified, $\tilde{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\gamma}}$ typically have different limits. We use **R** package *RCAL* to solve (A7) with the response taken as Y_i, regressors as $\boldsymbol{X}_i$, sample weight $\hat{\psi}(\boldsymbol{X}_i)e^{\boldsymbol{X}_i^\top\tilde{\boldsymbol{\gamma}}}g'\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\alpha}}\right)$ and offset $\tilde{\beta}A_i$ for each i. Denoting the solution of (A7) as $\hat{\boldsymbol{\gamma}}$, we finally obtain the estimator $\hat{\beta}_{\mathrm{HD}}$ by solving

$$n^{-1}\sum_{i=1}^{n}\hat{\psi}(\boldsymbol{X}_i)e^{\boldsymbol{X}_i^\top(\tilde{\boldsymbol{\gamma}}-\hat{\boldsymbol{\gamma}})}\left\{Y_i e^{-\beta A_i}-(1-Y_i)e^{\boldsymbol{X}_i^\top\hat{\boldsymbol{\gamma}}}\right\}\left\{A_i-g\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\alpha}}\right)\right\}=0. \tag{A8}$$

Here the final estimating equation is asymptotically equivalent to the second row of (3.2) only when $r(\boldsymbol{x})$ is correctly specified $\tilde{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\gamma}}$ have the same limiting values). When $r(\boldsymbol{x})$ is misspecified, the orthogonal score function used in (A8), denoted by $h'(\boldsymbol{D},\beta_0,\eta)$, is not the same as the $h(\boldsymbol{D},\beta_0,\eta)$ used in the main text. We will point out that this does not hurt the Neyman orthogonality of $h'(\boldsymbol{D},\beta_0,\eta)$. It is because when $r(\boldsymbol{x})$ is misspecified but $m(\boldsymbol{x})$ is correct (by our model Assumption HD1, at least one nuisance model need to be correct), $\partial_r h'(\boldsymbol{D};\beta_0,\bar{\eta})[r-\bar{r}]$ is naturally satisfied due to the correctness of $m(\boldsymbol{x})$, and $\partial_m h'(\boldsymbol{D};\beta_0,\bar{\eta})[m-\bar{m}]$ is satisfied according to the KKT (moment) condition of (A7). When $m(\boldsymbol{x})$ is misspecified but $r(\boldsymbol{x})$ is correct, $\partial_m h'(\boldsymbol{D};\beta_0,\bar{\eta})[m-\bar{m}]$ is naturally satisfied and (A8) is asymptotically equivalent with

$$n^{-1}\sum_{i=1}^{n}\hat{\psi}(\boldsymbol{X}_i)\left\{Y_i e^{-\beta A_i}-(1-Y_i)e^{\boldsymbol{X}_i^\top\hat{\boldsymbol{\gamma}}}\right\}\left\{A_i-g\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\alpha}}\right)\right\}=0,$$

as $\tilde{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\gamma}}$ approach the true $\boldsymbol{\gamma}_0$ and the second order errors is asymptotically negligible. So $\partial_r h'(\boldsymbol{D};\beta_0,\bar{\eta})[r-\bar{r}]$ is satisfied by (A7). Thus, our modified construction procedure does not break the theoretical guarantee of $\hat{\beta}_{\mathrm{HD}}$.

APPENDIX A5.: ADDITIONAL DETAILS OF NUMERICAL EXPERIMENTS

First, we present the mean vector and covariance matrix used to generate A and $\boldsymbol{X}$ in Section 5.1:

i. Take $\boldsymbol{\mu}_1 = (0.4, -0.25, -0.25, 0, \ldots, 0)$, $\boldsymbol{\mu}_0 = \mathbf{0}$, and

$$\left(\boldsymbol{\Sigma}^{-1}\right)_{ij} = \begin{cases} 1.5 & i = j = 1 \\ 1.2 & i = j \geq 2 \\ 0.2 & i = 1, 2 \leq j \leq 5 \text{ or } 2 \leq i \leq 5, j = 1 \\ 0 & \text{else} \end{cases}$$

ii. Take $\boldsymbol{\mu}_1 = (0.4, -0.25, -0.25, 0, \ldots, 0)$, $\boldsymbol{\mu}_0 = \mathbf{0}$,

$$\left(\Sigma^{-1}\right)_{ij} = \begin{cases} 1.5 & i = j = 1 \\ 1.2 & i = j \geq 2 \\ 0.2 & i = 1, 2 \leq j \leq 5 \text{ or } 2 \leq i \leq 5, j = 1 \\ 0 & \text{else} \end{cases}$$

and

$$\left(\boldsymbol{\Sigma}_0^{-1}\right)_{ij} = \begin{cases} 1.5 & i = j = 1 \\ 1.2 & i = j \geq 2 \\ 0.2 & i = 1, 2 \leq j \leq 5 \text{ or } 2 \leq i \leq 5, j = 1 \\ 0.075 & i = 3, 4, j = 2 \text{ or } i = 2, j = 3, 4 \text{ or } i = 3, j = 4 \text{ or } i = 4, j = 3 \\ 0 & \text{else} \end{cases}$$

iii. Take the covariance of $\boldsymbol{X}$ as

$$\boldsymbol{\Sigma}_{ij} = \begin{cases} 0.5 & i = j \\ 0.15 & i \leq 4, j \leq 4, i \neq j \\ 0 & \text{else} \end{cases}.$$

Then we present the specific choice on the basis functions for data generation in Section 5.2. In specific, we take

$$f_a(\boldsymbol{x}) = \left\{ \frac{1}{1+e^{x_1}}, \frac{1}{1+e^{x_2}}, \sin(x_3), \cos(x_4), I(x_5 > 0), I(x_6 > 0), x_7 x_8, x_9 x_{10} \right\}^\top;$$

$$\boldsymbol{\zeta}_a = (1, -1, 0.5, 0.5, 0.25, -0.25, 0.1, 0.1)^\top,$$

for $a_0(\boldsymbol{x})$, the conditional mean of $\boldsymbol{A}$ given $\boldsymbol{X} = \boldsymbol{x}$. And that

$$f_r(\boldsymbol{x}) = \left\{ x_1 x_2 x_3, x_4 x_5, x_6^3, \sin^2(x_7), \cos(x_8), \frac{1}{1+x_9^2}, \frac{1}{1+e^{x_{10}}}, I(x_{11} > 0), I(x_{12} > 0) \right\}^\top;$$

$$\boldsymbol{\zeta}_r = (0.1, 0.1, 0.1, -0.5, 0.5, 1, -1, 0.25, -0.25)^\top,$$

to specify $r_0(\boldsymbol{x})$.

REFERENCES

Athey S. and Imbens GW (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives 31(2), 3–32. [PubMed: 29465214]

Belloni A, Chernozhukov V, Chetverikov D, Hansen C. and Kato K. (2018). High-dimensional econometrics and regularized GMM. Paper 1806.01888, arXiv.org.

Bentancor A. and Clarke D. (2017). Assessing plan B: The effect of the morning after pill on children and women. Economic Journal 127(607), 2525–52.

Bickel PJ., Ritov Y and Tsybakov AB (2009). Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics 37(4), 1705–32.

Bradic J, Wager S. and Zhu Y. (2019). Sparsity double robust inference of average treatment effects. Paper 1905.00744, arXiv.org.

Bühlmann P. and van de Geer S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications, Berlin: Springer Science & Business Media.

Candes E. and Tao T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. Annals of Statistics 35(6), 2313–51.

Chen HY (2007).A semiparametric odds ratio model for measuring association. Biometrics 63(2),413–21. [PubMed: 17688494]

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. and Robins J. (2018). Double/debiased machine learning for treatment and structural parameters. Econometrics Journal 21(1), C1–C68.

Chernozhukov V, Newey W. and Robins J. (2018). Double/debiased machine learning using regularized Riesz representers. CEMMAP Working paper 15/18, Institute of Fiscal Studies, University College London, UK.

Chernozhukov V, Newey W, Robins J. and Singh R. (2018). Double/debiased machine learning of global and local parameters using regularized Riesz representers. Paper 1802.08667, arXiv.org.

Colangelo K. and Lee Y-Y (2020). Double debiased machine learning nonparametric inference with continuous treatments. Paper 2004.03036, arXiv.org.

Cui Y. and Tchetgen Tchetgen E. (2019). Bias-aware model selection for machine learning of doubly robust functionals. Paper 1911.02029, arXiv.org.

Dimitriadou E, Hornik K, Leisch F, Meyer D. and Weingessel A. (2004). R package e1071: Misc functions of the Department of Statistics. Software package, Technical University of Vienna, Austria.

Dukes O. and Vansteelandt S. (2020). Inference for treatment effect parameters in potentially misspecified high-dimensional models. Biometrika, 108, 321–34.

Farrell MH, Liang T. and Misra S. (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. Paper 1809.09953, arXiv.org.

Friedman J, Hastie T. and Tibshirani R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1), 1–22. [PubMed: 20808728]

Ghosh Sand Tan Z(2020).Doublyrobustsemiparametricinferenceusingregularizedcalibratedestimation with high-dimensional data. Paper 2009.12033, arXiv.org.

Giné E. and Nickl R. (2016).´ Mathematical Foundations of Infinite-Dimensional Statistical Models, vol. 40. New York: Cambridge University Press.

Greenwell B, Boehmke B. and Cunningham J, GBM Developers(2020). Package ‘gbm’, R package version, vol. 2.

Janková J. and van de Geer S. (2016). Confidence regions for high-dimensional generalized linear models under sparsity. Paper 1610.01353, arXiv.org.

Knaus MC (2018). A double machine learning approach to estimate the effects of musical practice on student’s skills. Paper 1805.10300, arXiv.org.

Knaus MC (2020). Double machine learning based program evaluation under unconfoundedness. Paper 2003.03191, arXiv.org.

Kuchibhotla AK and Chakrabortty A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. Paper 1804.02605, arXiv.org.

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z. and Kenkel B, R Core Team, et al. (2020). Package caret, CRAN.

Liaw A. and Wiener M. (2002). Classification and regression by RandomForest. R News 2(3), 18–22.

Lin X. and Carroll RJ (2006). Semiparametric estimation in general repeated measures problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 69–88.

Liu M, Xia Y, Cho K. and Cai T. (2020). Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. Paper 2004.00816, arXiv.org.

Ma R, Cai TT and Li H. (2020). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. Journal of the American Statistical Association, 116(534), 984–98. [PubMed: 34421157]

Negahban SN, Ravikumar P, Wainwright MJ and Yu B. (2012). A unified framework for high-dimensional analysis of *M*-estimators with decomposable regularizers. Statistical Science 27(4), 538–57.

Nekipelov D,Semenova V. and Syrgkanis V(2018).Regularizedorthogonalmachinelearningfornonlinear semiparametric models. Paper 1806.04823, arXiv.org.

Ning Y, Sida P. and Imai K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. Biometrika 107(3), 533–54.

Oprescu M, Syrgkanis V. and Wu ZS (2019). Orthogonal random forest for causal inference. Proceedings of Machine Learning Research 97, 4932–41.

Ripley B. and Venables W. (2016). Package nnet. R package version 7.3–12, CRAN.

Semenova V. and Chernozhukov V. Debiased machine learning of conditional average treatment effects and other causal functions. Econometrics Journal, Published ahead of print, 29 August 2020. 10.1093/ectj/utaa027.

Severini TA and Staniswalis JG (1994). Quasi-likelihood estimation in semiparametric models. Journal of the American statistical Association 89(426), 501–11.

Smucler E, Rotnitzky A. and Robins JM (2019). A unifying approach for doubly-robust ℓ_1-regularized estimation of causal contrasts. Paper 1904.03737, arXiv.org.

Tan Z. (2019a). On doubly robust estimation for logistic partially linear models. Statistics & Probability Letters 155, 108577.

Tan Z. (2019b). RCAL: Regularized calibrated estimation. R package version 1.0, CRAN.

Tan Z. (2020a). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. Annals of Statistics 48(2), 811–37.

Tan Z. (2020b). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. Biometrika 107(1), 137–58.

Tchetgen Tchetgen EJ, Robins JM and Rotnitzky A. (2010). On doubly robust estimation in a semiparametric odds ratio model. Biometrika 97(1), 171–80. [PubMed: 23049119]

Van de Geer S, Bühlmann P, Ritov Y. and Dezeure R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Annals of Statistics 42(3), 1166–202.

Van der Vaart AW (2000). Asymptotic statistics, vol. 3. Cambridge: Cambridge University Press.

Wager S. and Athey S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113(523), 1228–42.

Xia L, Li B. Nanand Y (2020). A revisit to debiased Lasso for generalized linear models. Paper 2006.12778, arXiv.org.

Yang J-C., Chuang H-C and Kuan C-M (2020). Double machine learning with gradient boosting and its application to the big *n* audit quality effect. Annals issue in honor of George Tiao: Statistical learning for dependent data. Journal of Econometrics 216(1), 268–83.

Zhu Y. and Bradic J. (2018). Significance testing in non-sparse high-dimensional linear models. Electronic Journal of Statistics 12(2), 3312–64.

Zimmert M. and Lechner M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. Paper 1908.08779, arXiv.org.

Table 1.

Average mean square error (MSE), average absolute bias (Bias), and average coverage probability (CP) of 95% CI of our HD estimator with the sample size set as 1,000, 1,500 and 2,000, under configurations (i)–(iii) described in Section 5.1. Number of repetition for each setting is 300.

	Configuration (i)			**Configuration (ii)**			**Configuration (iii)**		
n	1,000	1,500	2,000	1,000	1,500	2,000	1,000	1,500	2,000
MSE	0.008	0.006	0.004	0.007	0.005	0.004	0.007	0.004	0.003
Bias	0.021	0.021	0.018	0.020	0.022	0.018	0.030	0.016	0.013
CP	0.91	0.92	0.96	0.94	0.92	0.94	0.94	0.93	0.96

Table 2.

Average mean square error (MSE), average absolute bias (Bias), and average coverage probability (CP) of 95% CI of our ML estimator with sample sizes set as 1,000 and 2,000, which the nuisance models estimated using the four ML algorithms as well as the 'Best' approach described in Section 5.2. The number of repetitions for each setting is 300.

	***n* = 1,000**					***n* = 2,000**				
	GBM	**RF**	**SVM**	**NN**	**Best**	**GBM**	**RF**	**SVM**	**NN**	**Best**
MSE	0.013	0.014	0.013	0.012	0.013	0.007	0.008	0.008	0.007	0.007
Bias	0.037	0.046	0.048	0.015	0.037	0.039	0.051	0.049	0.042	0.039
CP	0.92	0.95	0.93	0.94	0.93	0.90	0.92	0.90	0.91	0.91

Table 3.

Point estimations, 95% CI lower/upper bounds (LB/UB) and two-side p-values for $\beta_0^{(1)}$ (log odds ratio of early gestation fetal death to the treatment of the EC pill) of our HD and ML (with the five different realisations described in Section 5.2) approaches.

Method	HD	GBM	SVM	RF	NN	Best
β_0	−0.171	−0.220	−0.153	−0.215	−0.177	−0.190
CI LB	−0.310	−0.355	−0.279	−0.378	−0.391	−0.324
CI UB	−0.028	−0.085	−0.027	−0.052	0.038	−0.057
p-value	0.018	0.000	0.016	0.007	0.088	0.003

Table 4.

Point estimations, 95% CI lower/upper bounds (LB/UB) and two-side p-values for $\beta_0^{(2)}$ (log odds ratio of new birth to the treatment of the EC pill) of our HD and ML (with the five different realisation described in Section 5.2) approaches.

Method	HD	GBM	SVM	RF	NN	Best
β_0	−0.186	−0.135	−0.124	−0.112	−0.148	−0.125
CI LB	−0.287	−0.235	−0.217	−0.224	−0.259	−0.222
CI UB	−0.089	−0.035	−0.030	0.000	−0.036	−0.029
p-value	0.000	0.007	0.009	0.033	0.010	0.009