Dimension-Free Average Treatment Effect Inference with Deep Neural Networks*

Xinze Du, Yingying Fan, Jinchi Lv, Tianshu Sun and Patrick Vossler
University of Southern California

November 30, 2021

Abstract

This paper investigates the estimation and inference of the average treatment effect (ATE) using deep neural networks (DNNs) in the potential outcomes framework. Under some regularity conditions, the observed response can be formulated as the response of a mean regression problem with both the confounding variables and the treatment indicator as the independent variables. Using such formulation, we investigate two methods for ATE estimation and inference based on the estimated mean regression function via DNN regression using a specific network architecture. We show that both DNN estimates of ATE are consistent with dimension-free consistency rates under some assumptions on the underlying true mean regression model. Our model assumptions accommodate the potentially complicated dependence structure of the observed response on the covariates, including latent factors and nonlinear interactions between the treatment indicator and confounding variables. We also establish the asymptotic normality of our estimators based on the idea of sample splitting, ensuring precise inference and uncertainty quantification. Simulation studies and real data application justify our theoretical findings and support our DNN estimation and inference methods.

Running title: ATED

Key words: Nonparametric inference; Average treatment effect; Dimension-free; Consistency and rate of convergence; Asymptotic distribution; Deep neural network

*Xinze Du is Ph.D. candidate, Department of Mathematics, University of Southern California, Los Angeles, CA 90089 (E-mail: xinzedu@usc.edu). Yingying Fan is Centennial Chair in Business Administration and Professor, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: fanyingy@marshall.usc.edu). Jinchi Lv is Kenneth King Stonier Chair in Business Administration and Professor, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: jinchilv@marshall.usc.edu). Tianshu Sun is Robert R. Dockson Associate Professor in Business Administration, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: tianshus@marshall.usc.edu). Patrick Vossler is Ph.D. candidate, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: pvossler@marshall.usc.edu). This work was supported by NSF Grants DMS-1953356 and EF-2125142.
1 Introduction

The estimation and inference of the average treatment effect (ATE) are foundational research topics in causal inference. Under the potential outcomes framework, the observed outcome $Y$ of a unit corresponds to one of two potential outcomes; one value for when the unit receives treatment and the other value for when the unit does not. The average treatment effect is defined as the population mean of the difference between these two potential outcomes. Since only one of the two potential outcomes can be observed for each unit, the estimation of ATE faces the common challenges in missing data problems. There is a large literature on ATE estimation. To name a few, see, for example, [3, 4, 11, 16, 21, 23]. See also the recent review papers [2, 15] on the existing methods and some new developments.

Under some regularity conditions, the observed outcome $Y$ can be formulated as the response of a mean regression problem with covariates $(X^\top, T)^\top$, where $X$ is the vector of covariates measuring the characteristics of the unit and $T$ is the treatment indicator taking values 0 and 1. Here, $T = 1$ means that the unit receives the treatment and $T = 0$ otherwise. Under such a model assumption, the average treatment effect is the expected difference of the mean regression functions corresponding to the treated and untreated groups. This motivates the estimation of ATE based on the estimated mean regression function, giving rise to the projection and imputation estimate [2].

In the era of big data, we have the luxury of collecting many covariates for each unit. Since it is generally challenging to test for confounding, a conservative approach is to include most, if not all, covariates with the aim of making the unconfoundedness assumption approximately correct. However, the large number of covariates, together with the potentially complicated interactions between covariates $X$ and the treatment indicator $T$, increases the challenge of ATE estimation and inference. On the one hand, while parametric regression models are relatively robust to the increased dimensionality of covariates, they impose stringent model structure assumptions which are unlikely to hold in practice, causing the issue of model misspecification. On the other hand, nonparametric models are much more flexible with mild model structure assumptions, but they can suffer from the curse of dimensionality, resulting in slower convergence rates. As a result, statistical inference, such as confidence interval construction, is more challenging in the nonparametric setting.

This paper explores two methods for ATE estimation and inference based on the nonparametric method of deep neural networks (DNNs) with theoretical underpinning. In recent years, DNNs have been popularly used to model the potentially complicated dependence structure of the response on covariates, thanks to their attractive approximation power. We first propose directly applying DNN for estimating the underlying mean regression function and then constructing an ATE estimate based on the estimated mean regression function. To overcome the curse of dimensionality, we adopt the specific deep neural network structure introduced and theoretically investigated in [5]. Such a network is recursively defined using some specifically designed two-layer neural networks as building blocks. As a result, some layers of the DNN are only sparsely connected. The specific structure of the DNN ensures dimension-free convergence rate of the resulting nonparametric mean regression estimate, as
formally revealed in [5].

Although elegant, the results in [5] are not directly applicable to our current model setting, mainly because of the discrete treatment indicator $T$, resulting in the nonsmoothness of the mean regression function with respect to its covariates. Similar to most other non-parametric regression methods, the theoretical study of the DNN estimate in [5] requires that the mean regression function has enough smoothness with respect to all covariates. To adapt the theory to our setting, we define a new function that linearly interpolates the values of the true mean regression function when $T = 0$ and $T = 1$. This new function has enough smoothness with respect to all its covariates, and thus the theory developed in [5] is applicable. We emphasize that this technical treatment is only for theoretical derivation and does not affect the practical implementation. In fact, the intermediate values of the newly constructed mean regression function when $T \in (0,1)$ are not used in our applications.

An ATE estimate based on the empirical mean over the same data for fitting the DNN can be obtained with the estimated mean regression function. We show that such an estimate is asymptotically consistent in estimating the true ATE, and the consistency rate is dimension-free, depending only on the smoothness parameter and another parameter controlling the number of hidden neurons. This result is consistent with that in [5]. However, despite the nice property of dimension-free consistency rate, such ATE estimate does not enjoy the asymptotic normality because of the bias. Therefore, we exploit the idea of sample splitting, where the ATE estimate is constructed as the empirical mean of the estimated DNN regression function evaluated on an independent inference data set. The similar sample splitting idea has been popularly used in the literature; see, for example, [9]. We show that if the sample used for DNN training is much larger than the sample used for inference, then the resulting ATE estimate enjoys the asymptotic normality, ensuring valid statistical inference.

We then incorporate the idea of DNN modeling into the doubly robust ATE estimation [13, 12]. We show that with the DNN estimate of the mean regression function discussed above, only very mild conditions on the propensity score estimation are needed for the doubly robust estimator to be consistent. For the asymptotic normality, we resort to the same sample splitting idea. We show that equally split samples, together with some additional mild assumptions on the propensity score estimation, can be sufficient for the doubly robust estimator to obtain asymptotic normality. In particular, we prove that the propensity score estimate based on the same DNN architecture gives us one such estimate.

The remaining of the paper is organized as follows. In Section 2, we introduce our model setting and two DNN-based ATE estimation methods. In Section 3, we study the sampling properties of these two estimators including their asymptotic normality. Sections 4 and 5 present numerical results using simulated examples and a real data example, respectively. Section 6 contains some conclusions and directions for future study. All technical proofs are deferred to the Appendix and the Supplementary Material.
For a real-valued multivariate function \( f \) stand for the collections of real numbers and positive integers, respectively, and \( \mathbb{N}_0 = \mathbb{N} \cup \{0\} \). We use \( \| \cdot \| \) to denote the Euclidean norm of vectors. To facilitate the technical presentation, we first introduce some necessary notation that will be used throughout the paper.

### 1.1 Notation

Let \( \mathcal{N}(\epsilon, \mathcal{F}, \| \cdot \|) \) denote the covering number of \( \mathcal{F} \) with metric \( \| \cdot \| \) at scale \( \epsilon > 0 \); see, e.g., [14]. That is, for a metric space \( (\mathcal{G}, \| \cdot \|) \) with \( \mathcal{F} \subset \mathcal{G} \), we define

\[
\mathcal{N}(\epsilon, \mathcal{F}, \| \cdot \|) = \min\{|M| : M \subset \mathcal{F} \subset \bigcup_{f \in M} B(f, \epsilon)\},
\]

where \( B(f, \epsilon) = \{g \in \mathcal{G} : \|f - g\| < \epsilon\} \) represents a ball centered at \( f \) with radius \( \epsilon \) in the metric space, and \( |\cdot| \) denotes the cardinality of a set.

Let \( \|f\|_\infty \) be the supremum norm of \( f : \mathbb{R}^p \to \mathbb{R} \), that is, \( \|f\|_\infty = \sup_{x \in \mathbb{R}^p} |f(x)| \); for any set \( A \subset \mathbb{R}^p \), define \( \|f(x)\|_\infty, A = \sup_{x \in A} |f(x)| \).

### 2 ATE inference using deep neural networks

#### 2.1 Model setting

Consider the potential outcomes framework of causal inference (see, e.g., [20]), where a set of independent and identically distributed (i.i.d.) observations \( \mathcal{D} \) are obtained. Here, for \( i = 1, \cdots, n_\mathcal{D} \) with \( n_\mathcal{D} := |\mathcal{D}| \), the \( i \)th observation in \( \mathcal{D} \) is denoted as \( (X_i, T_i, Y_i) \), where \( X_i = (X_{i1}, \cdots, X_{ip})^\top \) represents the vector of \( p \) covariates, \( T_i \) is the treatment indicator (1 for treated and 0 for untreated), and \( Y_i \in \mathbb{R} \) is the scalar response. The observed response takes the form \( Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) \), with the two potential outcomes \( Y_i(1) \) and \( Y_i(0) \) representing the outcomes with and without treatment, respectively. A common estimate of interest is the average treatment effect (ATE) defined as

\[
\tau = \mathbb{E}[Y_i(1) - Y_i(0)].
\]

Note that the potential outcome \( Y_i(t), t \in \{0, 1\} \), is latent when the individual \( i \) receives the opposite treatment \( T_i = 1 - t \), making the ATE estimation and inference challenging.

We assume the following nonparametric regression model for the observed response \( Y_i \)

\[
Y_i = m(X_i, T_i) + \varepsilon_i,
\]

where \( m(x, t) = \mathbb{E}[Y_i|X_i = x, T_i = t] \) is the underlying regression function, and \( \varepsilon_i \) is the model error with zero mean and finite variance and is independent of both \( X_i \) and \( T_i \). Throughout we make the commonly used assumptions that 1) \( T_i \perp \perp \{Y_i(0), Y_i(1)\}|X_i \) and
2) $0 < \mathbb{P}[T_i = 1|X_i] < 1$ almost surely, where the former is commonly referred to as the unconfoundedness assumption and the latter is called overlap assumption. The main focus of our paper is to develop statistical inference method for ATE using the nonparametric tool of DNN regression with theoretical underpinning.

We start by discussing the estimation of ATE, which is usually the first step of statistical inference. Under the above two assumptions of unconfoundedness and overlap, the right-hand side of (2) can be further written as

$$\tau = \mathbb{E}[m(X_i, 1) - m(X_i, 0)].$$

(4)

This suggests that we estimate the ATE using the empirical counterpart

$$\hat{\tau} = \hat{\mathbb{E}}_{X}[\hat{m}(X, 1) - \hat{m}(X, 0)],$$

(5)

where $\hat{m}(x, t)$ is an empirical estimate of $m(x, t)$ for $t = 0, 1$, and $\hat{\mathbb{E}}_{X}$ stands for the empirical mean with respect to $X$.

For the intuitive estimate in (5) to work well, we need to construct accurate estimates $\hat{m}(x, t)$ for $t = 0, 1$. With its appealing approximation property, DNN regression is a natural method to use for achieving this goal. For ATE estimation, the empirical mean $\hat{\mathbb{E}}_{X}$ in (5) can be constructed using the same data as those for learning $\hat{m}(x, t)$. However, if the goal is statistical inference, we will need to resort to data splitting and use an independent set to calculate the empirical mean to make the estimation bias under control in establishing the asymptotic normality of our estimator. A similar idea has been advocated in the literature; see, for example, [9]. We will formalize the above statements in subsequent sections.

In what follows, we will discuss two estimators: one is constructed using the exact intuition in (5), and the other one is the doubly robust estimate that also exploits information from the propensity score.

### 2.2 ATE inference based on DNN estimate

In the multivariate regression, it is well-known that classical nonparametric method can suffer from the curse of dimensionality when dimensionality $p$ is not very small. Fortunately, under certain network architectures of the DNN, one can learn a broad class of smooth functions accurately with the aid of modern optimization circumventing the curse of dimensionality; see, e.g., the recent work in [5]. In this paper, we will consider the same DNN network architecture described by the following functional space $\mathcal{H}^{(l)}_{M,p^*,p,\alpha}$ for the construction of ATE estimator.

**Definition 1.** Given positive integers $p^*, p, M, K$ and positive constant $\alpha$, for each $l \in \mathbb{N}$, the function space $\mathcal{H}^{(l)}_{M,p^*,p,\alpha}$ is defined recursively as

$$\mathcal{H}^{(l)}_{M,p^*,p,\alpha} = \left\{ h : \mathbb{R}^{p+1} \to \mathbb{R} \mid h(x) = \sum_{k=1}^{K} g_k(f_{1,k}(x), f_{2,k}(x), \ldots, f_{p^*,k}(x)) \right\},$$

for some $g_k \in \mathcal{H}^{(l)}_{M,p^*,p,\alpha}$ and $f_{j,k} \in \mathcal{H}^{(l-1)}_{M,p^*,p,\alpha}$. 5
where
\[
\mathcal{H}_{M,p^*,p,\alpha}^{(0)} = \left\{ f : \mathbb{R}^{p^*+1} \to \mathbb{R} \mid f(x) = \sum_{i=1}^{M} \mu_i \cdot \sigma(\sum_{j=1}^{4p^*} \lambda_{i,j} \cdot \sigma(\sum_{v=1}^{p^*+1} \theta_{i,j,v} \cdot x^{(v)} + \theta_{i,j,0}) + \lambda_{i,0}) + \mu_0 \text{ with } |\mu_i| \leq \alpha, |\lambda_{i,j}| \leq \alpha, \text{ and } |\theta_{i,j,v}| \leq \alpha \right\},
\]
and \(\sigma(\cdot)\), specified as the sigmoid function in our theoretical study, is the activation function.
Here, \(\mu_i, \lambda_{i,j}, \theta_{i,j,v} \in \mathbb{R}\) are weight coefficients, \(x^{(v)}\) denotes the \(v\)th component of vector \(x\), and \(\cdot\) means the regular scalar multiplication which is explicitly spelled out for the presentation clarity.

The architecture of the DNN described in Definition 1 has been investigated in [5], with an illustrative diagram given in Figure 1 therein. As can be seen from the definition, the DNN is a feedforward network defined recursively using the two-layer network in \(\mathcal{H}_{M,p^*,p,\alpha}^{(0)}\). As a result, many of the hidden layers are sparsely connected. The parameters \(p^*, K, \text{ and } M\) are all tuning parameters that need to be selected by the practitioner.

For a function \(f(x)\) and some positive value \(y\), define the truncation function
\[
\text{trunc}(f(x), y) = \begin{cases} f(x), & \text{if } |f(x)| \leq y, \\ y \cdot \text{sign}(f(x)), & \text{if } |f(x)| > y, \end{cases}
\]
where \(\text{sign}(t)\) denotes the sign function that takes value 1 if \(t > 0\), value \(-1\) if \(t < 0\), and value 0 if \(t = 0\). Given an i.i.d. sample \(D\), define
\[
\hat{m}_D(x, t) = \arg \min_{h \in \mathcal{H}^{(l)}} \left| D \right| \sum_{i \in D} (Y_i - h(X_i, T_i))^2
\]
as the optimal neural network in \(\mathcal{H}^{(l)} := \mathcal{H}_{M,p^*,p,\alpha}^{(l)}\) that minimizes the squared loss. Hereafter, with an abuse of notation, we use \(i \in D\) to represent the corresponding data \((Y_i, X_i, T_i) \in D\).

To increase the robustness of the DNN estimate, we truncate \(\hat{m}_D(x, t)\) as
\[
m_D(x, t) = \text{trunc}(\hat{m}_D(x, t), C \log n_D),
\]
where \(C\) is some large enough universal positive constant.

With the estimate \(m_D(x, t)\), we are halfway done with constructing the DNN estimate of \(\tau\) based on the intuition in (5). It remains to specify the empirical mean \(\hat{E}_X\) in (5). A natural estimate is to average over covariates \(X_i\) from the same learning data \(D\), that is,
\[
\hat{\tau}_D = \frac{1}{|D|} \sum_{i \in D} [m_D(X_i, 1) - m_D(X_i, 0)].
\]
We will show that such an estimate is consistent in estimating \(\tau\). However, the consistency rate is not fast enough for \(\hat{\tau}_D\) to achieve the asymptotic normality, hindering its ability for valid statistical inference.

To overcome such difficulty, we resort to the method of unbalanced sample splitting, which
allows us to separate the randomness in the approximation step from the randomness in the inference step. Specifically, we assume that the available data set \( D \) can be randomly split into two independent data sets, the learning set \( D_1 \) and the inference set \( D_2 \), with \( |D_1| = n^\gamma \) and \( \gamma > 1 \) some constant, and \( |D_2| = n \). Here, without loss of generality, we assume that \( n^\gamma \) is an integer. The learning set \( D_1 \) is used to compute the estimated nonparametric mean regression function \( m_{D_1}(x, t) \) as defined in (8). Then the inference set is also included to calculate the final ATE estimate, i.e.,

\[
\hat{\tau}(D_1, D_2) = \frac{1}{|D_2|} \sum_{i \in D_2} \hat{\tau}_i(D_1),
\]

with \( \hat{\tau}_i(D_1) = m_{D_1}(X_i, 1) - m_{D_1}(X_i, 0) \). We will show in Section 3 that the estimator defined in (10) achieves the asymptotic normality. From our technical analysis, we will also see that the unbalanced sample splitting plays a pivotal role in establishing the asymptotic normality.

### 2.3 Doubly robust estimate

The DNN estimate of \( \tau \) discussed in the previous section does not require the estimation of the propensity score function. This section explores a different type of estimator, the doubly robust estimator, for its robustness to the misspecification of either the mean regression function or the propensity score function. In addition, we will make it clear that the asymptotic normality of the doubly robust estimator can be achieved with equally split samples, making its practical implementation attractive.

We consider the same model as in (3). Given a data set \( D \) of i.i.d. observations, one can use the same DNN method as discussed in Section 2.2 to estimate the regression function, yielding an estimate \( m_D(x, t) \). We denote by \( \hat{m}_D(\cdot) = (\hat{m}_{D,1}(\cdot), \hat{m}_{D,0}(\cdot)) \) with \( \hat{m}_{D,t}(x) = m_D(x, t) \) for \( t = 0, 1 \) for notational simplicity. We denote the propensity score estimate as \( \hat{e}_D(x) \) that may be estimated by existing methods such as matching and stratification. Note that so far, we have not imposed any specific assumptions on the estimation accuracy of the propensity score function.

For a given data point \((Y_i, X_i, T_i)\), let us define

\[
\phi_i(\hat{e}_D, \hat{m}_D) = \frac{T_i}{\hat{e}_D(X_i)}(Y_i - \hat{m}_{D,1}(X_i)) + \hat{m}_{D,1}(X_i)
\]

and

\[
\psi_i(\hat{e}_D, \hat{m}_D) = \frac{1 - T_i}{1 - \hat{e}_D(X_i)}(Y_i - \hat{m}_{D,0}(X_i)) + \hat{m}_{D,0}(X_i).
\]

Then the doubly robust estimator based on data in \( D \) can be constructed as

\[
\hat{\tau}_{DR,D}(\hat{e}_D, \hat{m}_D) = \frac{1}{|D|} \sum_{i \in D} (\phi_i(\hat{e}_D, \hat{m}_D) - \psi_i(\hat{e}_D, \hat{m}_D)).
\]
We further define the population counterpart of $\hat{\tau}_{DR,D}(\hat{c}_D, \hat{m}_D)$ as

$$\tau(\hat{c}_D, \hat{m}_D) = \mathbb{E}(X,T,Y)(\phi(\hat{c}_D, \hat{m}_D) - \psi(\hat{c}_D, \hat{m}_D)),$$

where $\phi(\hat{c}_D, \hat{m}_D) = \mathbb{E}(X,Y|c_D, T, Y)(Y - \hat{m}_{D,1}(X)) + \hat{m}_{D,1}(X)$, $(X, T, Y)$ represents an independent new observation from the same distribution as $(X_1, T_1, Y_1) \in D$, $\psi(\hat{c}_D, \hat{m}_D)$ is defined analogously, and the expectation in (14) is taken with respect to $(X, T, Y)$.

As discussed in the previous section, the above estimate (13) is consistent in estimating the ATE under some regularity conditions. However, the estimation bias renders the asymptotic normality invalid. Next, we discuss the doubly robust estimator based on the idea of data splitting. Suppose we randomly split the set of available observations into two equal sized sets $D_1$ and $D_2$. Using data in $D_1$, we calculate the estimates $\hat{m}_{D_1,t}$, $t = 0, 1$, and $\hat{c}_D$ the same way as specified at the beginning of this section. Then the doubly robust estimator is constructed similar to (13) except that (11) and (12) are evaluated on the data in $D_2$; that is,

$$\hat{\tau}_{DR,D_2}(\hat{c}_{D_1}, \hat{m}_{D_1}) = \frac{1}{|D_2|} \sum_{i \in D_2} \left( \phi_i(\hat{c}_{D_1}, \hat{m}_{D_1}) - \psi_i(\hat{c}_{D_1}, \hat{m}_{D_1}) \right).$$ (15)

3 Asymptotic distributions of regular and doubly robust DNN estimators for ATE

Note that $m(x,t)$ takes a discrete covariate $T$ as an input, which greatly increases the theoretical challenges and makes the existing tools for studying the sampling properties of DNN inapplicable. For the purpose of motivating our technical analysis, let us temporarily assume that the propensity score $e(x) = P(T = 1|X = x)$ is known. Note that

$$\mathbb{E}[Y|X = x, e(X) = t] = \mathbb{E}[m(X,T)|X = x, e(X) = t]$$

$$= m(x,1)P(T = 1|X = x, e(X) = t) + m(x,0)P(T = 0|X = x, e(X) = t)$$

$$= [m(x,1)t + m(x,0)(1-t)] \mathbb{I}\{e(x) = t\},$$ (16)

where $\mathbb{I}\{\cdot\}$ stands for the indicator function. We extend the domain of the underlying regression function $m(x,t)$ to $\mathbb{R}^p \times [0, 1]$ and define the intermediate values as

$$m(x,t) = m(x,1)t + m(x,0)(1-t)$$ (17)

for $t \in (0, 1)$. It is seen that the extended function is infinitely differentiable with respect to $t$ in $(0, 1)$, and still satisfies our regression model assumption (3) on the boundary when $t \in \{0, 1\}$. Observe that $m(x,t)$ in (17) is a function defined on $\mathbb{R}^p \times [0, 1]$, and can be roughly understood as the underlying nonparametric regression function with $Y$ the response, and $(X^T, e(X))^T$ the new covariate vector\(^*$). The advantage of having $m(x, t)$ in (17) is that it is

\(^*$Rigorously speaking, this mean regression function is only defined on \{(x, t) : e(x) = t\}. Also, the overlap assumption prevents $e(X)$ from taking values 0 and 1. We temporarily ignore these constraints for the sake of motivating our technical analysis.
smooth with respect to \( t \), which will greatly facilitate us in developing new machine learning theory.

For observational studies, the propensity score function information is typically unknown. As a consequence, \( m(\mathbf{x}, t) \) in (17) is not directly estimable in the whole range of \( t \in [0, 1] \). Nevertheless, we still use the formulation in (17) keeping in mind that we only have observations on the boundary of the domain for \( t \in [0, 1] \) (i.e., the observed \( T_i \)'s). Since our theory does not rely on the values of \( m(\mathbf{x}, t) \) when \( t \in (0, 1) \), such treatment should not cause any problems in our technical analyses.

To set up the technical preparation, we briefly review the major definitions and notation from [5] below.

**Definition 2.** Given \( s > 0 \) and \( C > 0 \), the \((s,C)\)-smooth function class for functions of \( p \) real variables with \( s = q + r \), \( q \in \mathbb{N}_0 \), and \( 0 < r \leq 1 \) is defined as

\[
\mathcal{S}_{s,C,p} = \left\{ m : \mathbb{R}^{p+1} \to \mathbb{R} \mid \frac{\partial^q m(y)}{\partial x_1^{\alpha_1} \cdots \partial x_{p+1}^{\alpha_{p+1}}} - \frac{\partial^q m(z)}{\partial x_1^{\alpha_1} \cdots \partial x_{p+1}^{\alpha_{p+1}}} \leq C \|y - z\|^r \right\}
\]

for any \( y, z \in \mathbb{R}^{p+1} \) and \( \sum_{i=1}^{p+1} \alpha_i = q \) with \( \alpha_i \in \mathbb{N}_0 \), \( i = 1, \ldots, p+1 \).

In what follows, for the ease of presentation, we refer to \( \mathcal{S}_{s,C} \) as the function class that includes all the \( \mathcal{S}_{s,C,p} \) functions for all positive integers \( p \). The smoothness restrictions on the function class are commonly exploited for deriving nontrivial results on the rates of convergence for nonparametric estimators. In particular, the \((s,C)\)-smoothness condition in Definition 2 has been used to derive the distribution-free rates of convergence for nonparametric regression estimators; see, e.g., Section 3.2 of [14].

Now we are ready to introduce a generalized function class with some additional specific structures. These specific structures are well suited for our study and will assist us in the theoretical derivations. Recall that to facilitate our theory, the domain of the regression function \( m(\cdot, \cdot) \) is extended to \( \mathbb{R}^{p+1} \), while the values outside of the original domain do not convey any practical meaning. As will be seen in Condition 1 below, we assume that \( m(\cdot, \cdot) \) belongs to the class of \((s,C)\)-smooth generalized hierarchical interactive functions, which is formally defined as follows.

**Definition 3.** The \((s,C)\)-smooth generalized hierarchical interactive function class of order \( p^* \in \mathbb{N} \) and level \( l \in \mathbb{N} \) is defined recursively as

\[
\mathcal{M}_{p^*,l}(\mathcal{S}_{s,C}) = \left\{ m : \mathbb{R}^{p+1} \to \mathbb{R} \mid m(\mathbf{x}) = \sum_{k=1}^{K} g_k(f_{i,k}(\mathbf{x}), f_{2,k}(\mathbf{x}), \ldots, f_{p^*,k}(\mathbf{x})) \text{ with } g_k \in \mathcal{S}_{s,C}, \right. \\
\left. f_{i,k} \in \mathcal{M}_{p^*,l-1}(\mathcal{S}_{s,C}) \text{ for } i = 1, 2, \ldots, p^* \text{ and } k = 1, 2, \ldots, K \right\},
\]

where \( K \) is some positive integer and \( p + 1 \) is the dimensionality of the augmented covariate vector. When \( l = 0 \), \( \mathcal{M}_{p^*,0}(\mathcal{S}_{s,C}) \) is defined as

\[
\left\{ m : \mathbb{R}^{p+1} \to \mathbb{R} \mid m(\mathbf{x}) = f(\mathbf{a}_1^\top \mathbf{x}, \mathbf{a}_2^\top \mathbf{x}, \ldots, \mathbf{a}_{p^*}^\top \mathbf{x}) \text{ with } f \in \mathcal{S}_{s,C}, \mathbf{a}_i \in \mathbb{R}^{p+1} \text{ for } i = 1, 2, \ldots, p^* \right\}.
\]

The class of functions in Definition 3 above is rich enough to contain numerous commonly
used function classes such as the additive models, interaction models, and projection pursuit models. As a result, the assumption that the underlying mean regression function $m(\cdot, \cdot) \in \mathcal{M}_{p^*, l}(S_s, C)$ allows for rich model structures including interactions between the treatment indicator and the covariates, and also the latent factor structure in covariates. We also note that such a hierarchical structure resembles that of DNNs, which entails the approximation capabilities of DNN estimates defined in (8). See also [5] for some related discussions.

We are now ready to introduce the regularity conditions that are needed to facilitate our technical analysis.

**Condition 1.**

(i) The covariate vector $X$ has bounded support and response $Y$ has subGaussian distribution with $\mathbb{E}\exp(cY^2) < \infty$, where $c > 0$ is some constant.

(ii) The regression function $m(\cdot, t) \in \mathcal{M}_{p^*, l}(S_s, C)$ for some $s > 0$ and $C > 0$. By the definition of $\mathcal{M}_{p^*, l}(S_s, C)$, all partial derivatives of order no larger than $q$ of functions $g_k$ and $f_{i,k}$ are bounded by some universal positive constant in magnitude, and all the functions $g_k$ are Lipschitz continuous with Lipschitz constant $L > 0$.

(iii) For $\mathcal{H}^{(l)}_{M, p^*, p, \alpha}$, the parameters are taken as $M = \left\lceil c_1 n^{\frac{p^*}{2(p + r)}} \right\rceil$ and $\alpha = n^{c_2}$ for sufficiently large positive constants $c_1$ and $c_2$; the parameters $K$ and $p^*$ in defining $\mathcal{H}^{(l)}_{M, p^*, p, \alpha}$ are taken the same as in defining $\mathcal{M}_{p^*, l}(S_s, C)$ in part (ii) above.

(iv) There exists some constant $\delta > 0$ such that $e(X) \in [\delta, 1 - \delta]$ almost surely.

The boundedness of the support of the covariate distribution is commonly assumed in nonparametric regression and helps bound the complexity of the DNN function class. The slightly stronger assumption on overlap in Condition 1(iv) helps simplify the technical analysis. We also note that the parameters $K$ and $p^*$ in constructing the network $\mathcal{H}^{(l)}_{M, p^*, p, \alpha}$ should be correctly specified and thus equal to the ones in $\mathcal{M}_{p^*, l}(S_s, C)$, and this assumption is inherited from [5]. Establishing the theory when $K$ or $p^*$ is misspecified in constructing the DNN is highly challenging and left for future investigation.

### 3.1 Asymptotic normality of the regular DNN estimator

We start with presenting the consistency of the DNN estimator without data splitting defined in (9).

**Proposition 1.** Assume that Condition 1 with the sigmoid activation function $\sigma(x) = \frac{e^x}{e^x + 1}$ in $\mathcal{H}^{(l)}$ holds. Then the estimator $\hat{\tau}_D$ defined in (9) satisfies that $|\hat{\tau}_D - \tau| = o_P\{(\log n_D)^2 n_D^{-\frac{p^*}{2(p + r)}}\}$ as $n_D = |D| \to \infty$.

The proof of Proposition 1 uses some key results established in [14]. Thanks to the specific DNN network architecture in Definition 1, the rate of convergence in Proposition 1 above is free of dimensionality $p$. The intuition is that the underlying regression function
$m(x, t)$ has the sparsity structure specified in $\mathcal{M}_{p^*, l}(S_{s, C})$, whose complexity is controlled by $p^*$. Thus, the dimension-free convergence rate is attainable.

We now present the asymptotic normality of the data splitting estimator (10). Recall that after splitting, we have data sets of sizes $|D_1| = n^\gamma$ and $|D_2| = n$. To gain some high-level understanding, consider the decomposition

$$\sqrt{n}(\hat{m}(D_1, D_2) - \tau) = \frac{1}{\sqrt{n}} \sum_{i \in D_2} (\tau_i - \tau) + \frac{1}{\sqrt{n}} \sum_{i \in D_2} (\tilde{\tau}_i(D_1) - \tau_i),$$

(18)

where $\tau_i = m(X_i, 1) - m(X_i, 0)$ and $\tilde{\tau}_i(D_1)$ is defined in Section 2.2. Note that the first term on the right-hand side of (18) is the scaled summation of i.i.d. mean zero random variables and thus is asymptotically normal. For the second term on the right-hand side of (18), since the proof of Proposition 1 shows that $m_{D_1}(x, t)$ is consistent in estimating $m(x, t)$, it follows that the second term is negligible when the sample size of $D_1$ is much larger than that of $D_2$. These results are formally presented in Theorem 1 below.

**Theorem 1.** Assume that the conditions of Proposition 1 hold and $\gamma > 1 + \frac{p^*}{2\gamma}$. Then we have

$$\sqrt{n}(\hat{m}(D_1, D_2) - \tau) \xrightarrow{p} N(0, \sigma^2)$$

(19)

as $n \to \infty$, where $\sigma^2 = \text{Var}(m(X, 1) - m(X, 0))$.

The requirement of $\gamma > 1 + \frac{p^*}{2\gamma}$ in Theorem 1 above can be relaxed if the regression function $m(x, t)$ takes a more specific form, as formally presented in the condition of the corollary below.

**Condition 2.**

(i) The regression function $m(x, t) \in \mathcal{M}_{p^*, l}(S_{s, C})$, where all functions $g_k$ and $f_{i,k}$ with $k = 1, \ldots, K$ and $i = 1, \ldots, p^*$ appearing in the definition of $\mathcal{M}_{p^*, l}(S_{s, C})$ are polynomials taking the following generic form

$$f(x) = \sum_{|\alpha| \leq q} r_{\alpha} x^\alpha$$

with some $q \in \mathbb{N}_0$, $r_{\alpha} \in \mathbb{R}$ the regression coefficient, $x = (x_1, \ldots, x_{p+1})^\top$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{p+1})$, and $x^\alpha = x_1^{\alpha_1} \cdots x_{p+1}^{\alpha_{p+1}}$. Here, assume that $\alpha_i \in \mathbb{N}_0$ and $|\alpha| = \sum_{i=1}^{p+1} \alpha_i$.

(ii) Denote by $q_0$ the highest order of all the polynomials in part (i). Take the parameters in $H^{(l)}_{M, p^*, l, \alpha}$ as $M = \left[ c_1 n^{2q_0 + p^*} \right]$ and $\alpha = n^{c_2}$ for sufficiently large positive constants $c_1$ and $c_2$, where the non-decreasing sequence $\{\lambda_n\}$ is defined as

$$\lambda_n = \inf \{ s \in \mathbb{N} : n(\lambda) \geq n + 1 \}$$

with

$$n(\lambda) = \inf \{ n \in \mathbb{N} : \frac{\log(n)}{2s + p^*} \geq \log\left( \frac{3}{2(2q_0 + 3)} \right) + \log(\log n) \}.$$
In addition, parameters \( K \) and \( p^* \) in defining \( H_{M,p^*,p,\alpha} \) and \( M_{p^*,l}(S_s,C) \) in part (i) are the same.

The parameter \( s \) in Condition 2(i) above can take some arbitrary positive value in \( \mathbb{R} \), in view of the specific form of functions involved in the definition. The constants \( c_1 \) and \( c_2 \) in Condition 2(ii) are generally different from the corresponding constants in Condition 1, because the former ones depend generally on \( p^*, p, \) and \( q_0 \), while the latter ones can depend on parameter \( p^*, p, \) and \( s \) in Condition 1.

**Corollary 1.** Assume that (i) and (iv) of Condition 1 and Condition 2 hold with the sigmoid activation function \( \sigma(x) = \frac{e^x}{1 + e^x} \) in \( H(l) \). Let \( |D_1| = n(\log n)^k \) with \( |D_2| = n \) for some \( k > 4 + p^* \). Then for \( \hat{\tau}(D_1,D_2) \) defined in (10), we have
\[
\sqrt{n}(\hat{\tau}(D_1,D_2) - \tau) \xrightarrow{\mathbb{P}} N(0,\sigma^2)
\] as \( n \to \infty \), where \( \sigma^2 \) is as defined in Theorem 1.

Rigorously speaking, Corollary 1 cannot be proved by directly applying Proposition 1 or Theorem 1. The main difficulty is that although a regression function \( m(x) \) satisfying Condition 2 belongs to \( M_{p^*,l}(S_s,C) \) over all \( s \in \mathbb{N} \), the probabilistic statements in proving Proposition 1 and Theorem 1 do not hold uniformly for all \( s \in \mathbb{N} \). Thus, we cannot simply set \( s \) to infinity to prove Corollary 1. Instead, we must first establish results similar to those in [5] in order to prove Corollary 1. Nevertheless, since the function class in Corollary 1 is much smaller, we downgrade the importance of the result and name it a corollary. Whether a similar result holds for a broader class of analytic functions that are infinitely differentiable is an interesting question for future study.

Compared to Theorem 1, the weaker assumption in Corollary 1 on \( \gamma_n \) indicates that the asymptotic normality is possible with nearly balanced sample splitting. The fundamental reason is that, by modifying the proof of Theorem 1 in [5] to require a stronger structural assumption on the mean regression function \( m(x,t) \), we can show that the DNN regression function achieves a near \( n^{-1/2} \) convergence rate (up to some logarithmic factor).

For the asymptotic normality in Theorem 1 and Corollary 1 to be practically applicable, we need an accurate variance estimate. Let us consider the following natural choice
\[
\hat{\sigma}^2(D_1,D_2) = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i \in D_2} \hat{\tau}^2_i(D_1) - \left( \frac{1}{n} \sum_{i \in D_2} \hat{\tau}_i(D_1) \right)^2 \right).
\] (21)

The independence between the data in \( D_1 \) and \( D_2 \) and the consistency of \( m_{D_1}(x,t) \) (cf. the proof of Proposition 1) ensure that \( \hat{\sigma}^2(D_1,D_2) \) introduced in (21) is a consistent estimator of \( \sigma^2 \), yielding the following asymptotic normality with the estimated variance.

**Theorem 2.** Under the conditions of Theorem 1, we have the asymptotic normality using the variance estimator defined in (21)
\[
\frac{\sqrt{n}(\hat{\tau}(D_1,D_2) - \tau)}{\hat{\sigma}(D_1,D_2)} \xrightarrow{\mathbb{P}} N(0,1)
\] (22)
as $n \to \infty$. Moreover, it holds that

$$|\hat{\sigma}^2(D_1, D_2) - \sigma^2| = o_P((\log n)^4 n^{-1/2})$$

for large enough $n$.

Theorem 2 above makes the practical construction of confidence intervals (CIs) possible when sample size $n$ is large. In particular, a level 100(1 − $\alpha$)% CI for $\tau$ is given by

$$(\hat{\tau}(D_1, D_2) - n^{-1/2}\hat{\sigma}(D_1, D_2)z_{\alpha/2}, \hat{\tau}(D_1, D_2) + n^{-1/2}\hat{\sigma}(D_1, D_2)z_{\alpha/2}),$$

where $z_{\alpha/2}$ is the 100(1 − $\alpha/2$)th percentile of the standard normal distribution. Corollary 2 below summarizes the results that are parallel to those in Corollary 1.

**Corollary 2.** Under the conditions of Corollary 1, the asymptotic normality in (22) holds. In addition, we have $|\hat{\sigma}^2(D_1, D_2) - \sigma^2| = O_P((\log(n))^4 n^{-1/2})$.

### 3.2 Asymptotic normality of the doubly robust DNN estimator

Recall that we use the balanced sample splitting in constructing the doubly robust estimator. We slightly abuse the notation and use $n$ to denote the common sample size for both $D_1$ and $D_2$ in this section. We require the following condition on the propensity score estimation for investigating the sampling properties of the doubly robust estimator.

**Condition 3.**

(i) There exists some constant $C_2 > 0$ such that for any $n$, the propensity score estimate $\hat{e}_{D_1}(x)$ constructed from sample $D_1$ satisfies that

$$\frac{1}{C_2 \log n} \leq \hat{e}_{D_1}(X) \leq 1 - \frac{1}{C_2 \log n},$$

for $X$ almost surely, where $X$ is an independent observation from the same distribution as $X_1$.

(ii) It holds that

$$\mathbb{E}\left\{\frac{1}{n} \sum_{i \in D_1} |\hat{e}_{D_1}(X_i) - e(X_i)|^2\right\} = o\left(\frac{1}{\log n}\right).$$

(iii) Assume that

$$\mathbb{E}_{D_1}\mathbb{E}_X|\hat{e}_{D_1}(X) - e(X)|^2 = o(n^{-1/2}),$$

where $X$ is an independent observation from the same distribution as $X_1$.

Condition 3(i) can be easily satisfied if we define a truncated propensity score estimator; see (33) below for an example. Condition 3(ii) is a mild consistency assumption on $\hat{e}_{D_1}$. Condition 3(iii) plays a crucial role in establishing the asymptotic normality of the doubly robust estimator based on sample splitting. We will suggest a propensity score estimator that satisfies all these conditions toward the end of this section.
**Proposition 2.** Assume that the conditions of Proposition 1 hold and the propensity score estimator \( \hat{e}_{D_1} \) satisfies (i) and (ii) of Condition 3. Then the doubly robust estimator defined in (13) satisfies that
\[
|\hat{\tau}_{DR,D_1}(\hat{e}_{D_1},\hat{m}_{D_1}) - \tau| = o_P(1).
\]

Proposition 2 does not give us an explicit convergence rate because of the very weak assumptions on the propensity score estimator \( \hat{e}_{D_1} \). The explicit rate can be derived at the cost of assuming the faster convergence rate for assumptions on the propensity score estimator \( \hat{e}_{D_1} \) as \( n \to \infty \).

**Theorem 3.** Assume that the conditions of Proposition 1 hold and the propensity score estimator \( \hat{e}_{D_1} \) satisfying (i) and (iii) of Condition 3, the doubly robust ATE estimator based on the sample splitting defined in (15) has the asymptotic normality
\[
\sqrt{n}(\hat{\tau}_{DR,D_2}(\hat{e}_{D_1},\hat{m}_{D_1}) - \tau) \overset{d}{\to} N(0,\sigma^2_{DR})
\]
as \( n \to \infty \), where \( \sigma^2_{DR} = \text{Var}(m_1(X) - m_0(X)) + \text{Var}(\varepsilon)E_{\varepsilon|X}\frac{1}{(1-e(X))} \).

Comparing Theorem 3 with Theorem 1, the doubly robust estimator has larger asymptotic variance than the regular DNN estimator \( \hat{\tau}(D_1,D_2) \). This is reflected in the results of our simulation studies. Similar to the DNN estimate presented in the previous section, the asymptotic variance \( \sigma^2_{DR} \) can be estimated using a plug-in estimator
\[
\hat{\sigma}^2_{DR,D_2}(\hat{e}_{D_1},\hat{m}_{D_1}) = \frac{n}{n-1}\left(\frac{1}{n} \sum_{i \in D_2} \hat{\tau}_i^2(\hat{e}_{D_1},\hat{m}_{D_1}) - \left(\frac{1}{n} \sum_{i \in D_2} \hat{\tau}_i(\hat{e}_{D_1},\hat{m}_{D_1})\right)^2\right),
\]
where \( \hat{\tau}_i(\hat{e}_{D_1},\hat{m}_{D_1}) = \phi_i(\hat{e}_{D_1},\hat{m}_{D_1}) - \psi_i(\hat{e}_{D_1},\hat{m}_{D_1}) \) and all the notation is the same as in Section 2.3.

**Theorem 4.** Under the conditions of Theorem 3, it holds that
\[
\frac{\sqrt{n}(\hat{\tau}_{DR,D_2}(\hat{e}_{D_1},\hat{m}_{D_1}) - \tau)}{\hat{\sigma}_{DR,D_2}(\hat{e}_{D_1},\hat{m}_{D_1})} \overset{d}{\to} N(0,1)
\]
as \( n \to \infty \). In addition, we have
\[
|\hat{\sigma}^2_{DR,D_2}(\hat{e}_{D_1},\hat{m}_{D_1}) - \sigma^2_{DR}| = o_P((\log n)^2n^{-1/4}).
\]

Next we consider a specific propensity score estimator that satisfies the conditions of Theorem 3. We start with introducing the condition below which restricts the structure of the true propensity score.

**Condition 4.** The propensity score \( e(x) \in \mathcal{M}_{p^*}d(S_{s_e,C_e}) \) for some constants \( s_e = q_e + r_e > 0 \) with \( q_e \in \mathbb{N}_0 \) and \( 0 < r_e \leq 1 \), and \( C_e > 0 \). Moreover, all partial derivatives of order no larger than \( q_e \) of functions \( g_k \) and \( f_{i,k} \) involved in the definition of \( \mathcal{M}_{p^*}d(S_{s_e,C_e}) \) are bounded by some universal positive constant in magnitude, and all functions \( g_k \) are Lipschitz continuous with Lipschitz constant \( L_e > 0 \).
Observe that the above condition on the propensity score resembles Condition 1(ii) for \( m(x, t) \) except that the ambient dimensionality is \( p \) instead here. The smoothness parameters in these two conditions can be different and one may use \( \min\{ s, s_e \} \) to unify them. Condition 4 above accommodates commonly used propensity score functions such as the logistic function of form \( e(x) = \frac{\exp(a^T x)}{1 + \exp(a^T x)} \) with \( a \in \mathbb{R}^p \) the regression coefficient vector. It is seen that \( e(x) = f(a^T x) \) for \( f(x) = \frac{\exp(x)}{1 + \exp(x)} \). Thus, the propensity score function belongs to the function class \( M_{1,0}(S_{s_e, C_e}) \) for any positive \( s \geq 1 \) and some \( C_e \) depending on \( s \). For example, by letting \( s = 1 \), we see that \( f(x) \in S_{1,1/4,1} \) (see Definition 2) and the condition of \( 1 = p^* < 2s = 2 \) in Theorem 3 holds.

We next introduce the DNN estimate for the propensity score. Let us define

\[
\tilde{e}_{D_1}(x) = \frac{1}{2} + \operatorname{trunc} \left( \tilde{e}_{D_1}(x, t) - \frac{1}{2} \cdot \frac{1}{C_2 \log(n)} \right),
\]

(33)

where \( C_2 > 0 \) is some constant and

\[
\tilde{e}_{D_1}(x) = \arg \min_{h \in \mathcal{H}_{M, p^*, p-1, \alpha}^{(l)}} \frac{1}{n} \sum_{i \in D_1} |T_i - h(X_i)|^2.
\]

(34)

We make the same assumption that parameters \( K \) and \( p^* \) in \( \mathcal{H}_{M, p^*, p-1, \alpha} \) above are set at their true values in Condition 4 for defining \( \mathcal{M}_{p^*, f}(S_{s_e, C_e}) \).

**Corollary 3.** Assume that the conditions of Proposition 1 hold with \( p^* < 2s_e \) and Condition 4 holds. Then the propensity score estimator \( \tilde{e}_{D_1}(\cdot) \) defined in (33) satisfies (i) and (iii) of Condition 3. Consequently, the resulting doubly robust estimator enjoys the same asymptotic normality as in Theorems 3 and 4.

### 4 Simulation studies

In this section, we consider simulation examples mimicking observational data to verify the theoretical results obtained in Section 3 for the two ATE estimates and illustrate their finite-sample performance.

#### 4.1 Simulation results of the mean difference DNN estimator for ATE

Consider the following main effect for the control group \( T = 0 \)

\[
m_0(x) = \mathbb{E}(Y_i(0)|x) = x_1^2 + x_2 + x_3^2,
\]

(35)

where we choose \( x = (X_1, \cdots, X_p)^\top \sim \text{Uniform}(\mathbb{R}^p) \). The treatment propensity score \( P(T = 1|x) \) is defined as

\[
e(x) = \frac{1}{4}(1 + \beta_{2,4}(x_3)),
\]

(36)

†This can be verified by Faà di Bruno’s formula for high order derivatives of the composite function \( f(x) = f_1 \circ f_2 \), where \( f_1(x) = \frac{x}{1+x} \) and \( f_2(x) = e^x \).
where $\beta_{2,4}$ denotes the beta distribution with shape parameters 2 and 4. Finally, the treatment effect is kept fixed at $\tau(x) = 1$ and we assume an additive model error of $\varepsilon \sim N(0, 1)$.

A similar simulation setting was first proposed in [22] with a linear main effect function. We use a slightly more complicated main effect function, but our goal is the same as in [22]. Specifically, we intend to test the ability of our estimator to correct for bias due to an interaction between the propensity score and the main effect. This simulation setting mirrors the challenge in observational studies in which the treatment assignment is correlated with the potential outcomes. Thus, the statistical method must accurately adjust for the observed covariates to avoid a biased estimate.

We generate a data set of size $n_D$ from the above observational data model in (35)–(36) and we set $p = 50$. Then we randomly split the data into two parts: a training sample $\mathcal{D}_1$ of size $n_1 = cn$ and an inference sample $\mathcal{D}_2$ of size $n$, where $n = 1000$ and we consider the choices of $c = 1, \ldots, 5$. For each generated data set, we apply a deep neural network (DNN) model with the feedforward network structure to the training sample. More specifically, we employ a DNN with three hidden layers, where the number of neurons in each hidden layer is set as $p + 1$ since we include the treatment assignment as an input into our network. Furthermore, we set the learning rate and batch size as 0.001 and 128, respectively, and allow the number of epochs to vary from 100 to 800. We optimize the network parameters using the Adam optimizer. Finally, we consider the two popular choices of the sigmoid activation and the ReLU activation for the activation function.

We begin with the imbalanced samples version of the ATE estimate with DNN defined in Section 2.2. A joint nonparametric regression function $\hat{m}_{\mathcal{D}_1}(x, t)$ can be constructed based on the training sample $\mathcal{D}_1$. Then we can construct the regular DNN ATE estimator using the inference sample $\mathcal{D}_2$. The simulation example is repeated 200 times to generate the distribution of the resulting regular DNN ATE estimator.

Figure 1 and Table 1 present the results of the imbalanced samples version of the ATE estimate with DNN as a function of the choice of activation (i.e., sigmoid vs. ReLU), the training-to-inference ratio $c$, and the number of epochs varying from 100 to 800.

From Figure 1 and Table 1, we see that sigmoid activation generally outperforms ReLU activation in terms of the bias and variance. Indeed, our technical assumptions exclude ReLU because of its nonsmoothness. Developing theory for ReLU is an interesting research topic for future study. The empirical distribution of the ATE estimator is rather close to the normal distribution that is nearly centered around the true value of the ATE $\tau$. Furthermore, we observe that the results improve as the training-to-inference ratio $c$ increases, which is consistent with our theory. We also observe that the performance of the ATE estimator becomes better as the number of epochs grows. However, since the risk of overfitting also increases when the number of epochs is too large, we recommend to cap it to prevent overfitting of the DNN model. We present additional simulation results with different numbers of epochs in Section C of the Supplementary Material.
Figure 1: The scaled density of the ATE estimate over 200 replications for different training sample sizes and different activation functions. Here we use a fixed inference sample size of $n = 1000$ and train each network for 800 epochs. From top to bottom, the training sample size $n_1$ increases from 1000 to 5000. The true treatment effect of $\tau = 1$ is shown as a red vertical line. Results for different training lengths can be found in Section C of the Supplementary Material.

4.2 Simulation results of the doubly robust DNN estimator for ATE

We now turn to the doubly robust version of the ATE estimate with DNN as defined in Section 2.3. The simulation setting is the same as in Section 4.1. A key difference is that in addition to constructing an estimated regression function $\hat{m}_{D_1}(x,t)$ based on the training sample $D_1$, we will also construct the estimated propensity score $\hat{e}_{D_1}(x)$ based on the same training sample $D_1$. Then using the inference sample $D_2$, we can construct the doubly robust DNN ATE estimator as given in (15). For the construction of the estimated propensity score with DNN, we can always fix a relatively small number of epochs for the training of the network (e.g., at 100 across all the settings) and at the same time, constrain the estimated propensity score within $[0.01, 1 - 0.01]$. The main purpose of these modifications is to prevent the over- or perfect fitting of the propensity score. Moreover, we will vary the number of epochs for the construction of $\hat{m}_{D_1}(x,t)$ with DNN as in Section 4.1.

Figure 2 and Table 2 present the results of the doubly robust version of the ATE estimate with DNN as a function of the choice of activation (i.e., sigmoid vs. ReLU), the training-to-inference ratio $c$, and the number of epochs varying from 100 to 500 (for the construction of the estimated joint regression function $\hat{m}_{D_1}(x,t)$ as mentioned above).

From Figure 2 and Table 2, we see that the sigmoid doubly robust estimator has comparable performance to that of the difference of means estimate (i.e., our first method), but with slightly larger variance. This is consistent with our theoretical results in Theorems 1 and 3. It is also interesting to observe that for balanced samples (i.e., the case of $c = 1$), the performance of the sigmoid doubly robust estimator was rather close to that of the sigmoid
Table 1: Results of the same simulation setting as in Figure 1 aggregated over 200 replications. In each replication, the networks are trained for 800 epochs. Results for different training lengths can be found in Section C of the Supplementary Material.

| n | Activation | Mean   | Median | SD    | MSE    |
|---|------------|--------|--------|-------|--------|
| 1000 | ReLU | 0.9567 | 0.9592 | 0.09532 | 0.01091 |
|     | Sigmoid | 1.0196 | 1.0175 | 0.07522 | 0.00601 |
| 2000 | ReLU | 0.9760 | 0.9703 | 0.07188 | 0.00572 |
|     | Sigmoid | 0.9927 | 0.9864 | 0.05797 | 0.00340 |
| 3000 | ReLU | 0.9837 | 0.9776 | 0.07215 | 0.00544 |
|     | Sigmoid | 0.9911 | 0.9899 | 0.04983 | 0.00255 |
| 4000 | ReLU | 0.9769 | 0.9740 | 0.06647 | 0.00493 |
|     | Sigmoid | 0.9881 | 0.9926 | 0.04098 | 0.00181 |
| 5000 | ReLU | 0.9821 | 0.9866 | 0.06029 | 0.00394 |
|     | Sigmoid | 0.9941 | 0.9931 | 0.04124 | 0.00173 |

Table 1: Results of the same simulation setting as in Figure 1 aggregated over 200 replications. In each replication, the networks are trained for 800 epochs. Results for different training lengths can be found in Section C of the Supplementary Material.

mean difference estimator. When $c$ grows, the latter one has much improved performance while the former stays more or less the same. The fact that the training-to-inference sample ratio has more impact on difference of means estimate is also consistent with our theory. On the contrary, the ReLU doubly robust estimator had excessively large variance. Also, the training and network tuning for the purpose of ATE inference with ReLU can be more challenging according to our empirical experience. These suggest against the use of ReLU for our application. Results corresponding to different numbers of epochs are presented in Section C of the Supplementary Material.

5 Real data application

As a supplement to our theoretical results and our simulation studies, we demonstrate the practical usage of our proposed methods by studying the effect of 401(k) eligibility on accumulated assets as in [7, 10, 1].

There has been a considerable line of research focused on understanding the effect of a 401(k) plan on the accumulated assets of a household. The challenge here is that there is heterogeneity amongst savers and the decision to enroll in a 401(k) plan is non-random. To address the endogeneity of 401(k) participation, [17, 18] and [8] used data from the 1991 Survey of Income and Program Participation (SIPP) and argued that eligibility for enrolling in a 401(k) plan can be taken as exogenous after controlling for observables, particularly income. The crux of their argument is that, around the time this data was collected, 401(k) plans were still relatively new and most people based their employment decisions on income, not on whether their employer offered a 401(k) plan. Thus, eligibility for a 401(k) plan could be taken as exogenous conditional on income, and the causal effect of 401(k) eligibility could

\[ \text{This is because though a 401(k) plan is a tax-deferred retirement plan that is provided through an employer. Therefore only workers in firms that offer 401(k) plans are eligible.} \]
| n_1  | Estimate Type          | Activation | Mean     | Median   | SD       | MSE       |
|------|------------------------|------------|----------|----------|----------|-----------|
| 1000 | Difference of Means Estimate | ReLU       | 0.9665   | 0.9751   | 0.10528  | 0.01215   |
|      |                        | Sigmoid    | 1.0177   | 1.0193   | 0.07743  | 0.00628   |
|      | Doubly Robust Estimate | ReLU       | 0.9757   | 0.9716   | 0.19882  | 0.03992   |
|      |                        | Sigmoid    | 0.9620   | 0.9631   | 0.08996  | 0.00949   |
| 2000 | Difference of Means Estimate | ReLU       | 0.9792   | 0.9743   | 0.17514  | 0.03095   |
|      |                        | Sigmoid    | 0.9750   | 0.9771   | 0.07266  | 0.00588   |
|      | Doubly Robust Estimate | ReLU       | 0.9840   | 0.9843   | 0.05472  | 0.00324   |
|      |                        | Sigmoid    | 0.9754   | 0.9749   | 0.08674  | 0.00809   |
| 3000 | Difference of Means Estimate | ReLU       | 0.9896   | 0.9882   | 0.04791  | 0.00239   |
|      |                        | Sigmoid    | 0.9772   | 0.9881   | 0.13725  | 0.01926   |
|      | Doubly Robust Estimate | ReLU       | 0.9754   | 0.9749   | 0.08674  | 0.00809   |
|      |                        | Sigmoid    | 0.9751   | 0.9715   | 0.08919  | 0.00854   |
| 4000 | Difference of Means Estimate | ReLU       | 0.9785   | 0.9808   | 0.07092  | 0.00547   |
|      |                        | Sigmoid    | 0.9806   | 0.9882   | 0.04791  | 0.00239   |
|      | Doubly Robust Estimate | ReLU       | 0.9874   | 0.9886   | 0.06328  | 0.00450   |
|      |                        | Sigmoid    | 0.9837   | 0.9852   | 0.04295  | 0.00210   |
| 5000 | Difference of Means Estimate | ReLU       | 1.0051   | 0.9869   | 0.16570  | 0.02735   |
|      |                        | Sigmoid    | 0.9785   | 0.9753   | 0.08152  | 0.00707   |
|      | Doubly Robust Estimate | ReLU       | 0.9874   | 0.9886   | 0.06703  | 0.00463   |
|      |                        | Sigmoid    | 0.9941   | 0.9953   | 0.03458  | 0.00122   |
|      | Doubly Robust Estimate | ReLU       | 0.9843   | 0.9710   | 0.13930  | 0.01955   |
|      |                        | Sigmoid    | 0.9857   | 0.9902   | 0.07306  | 0.00552   |

Table 2: The simulation results corresponding to Figure 2 for 800 training epochs. Results for different training lengths can be found in Section C of the Supplementary Material.
Figure 2: The scaled density of the ATE estimate over 200 replications for different training sample sizes and different activation functions. The red curves correspond to the DNN estimate defined in (10) and the blue curves correspond to the doubly robust estimate defined in (15). The true treatment effect of \( \tau = 1 \) is shown as a red vertical line. Here we use a fixed inference sample size of \( n = 1000 \) and train each network for 800 epochs. From top to bottom, the training sample size \( n_1 \) increases from 1000 to 5000. Results for different training lengths can be found in Section C of the Supplementary Material.

We use the same data as in [7], which consists of 9915 observations at the household level from the 1991 SIPP. Specifically, we use net financial assets as our outcome variable and the covariates are age, income, family size, years of education, and indicators for marital status, two-earner status, defined benefit pension status, IRA participation, and home ownership. Since 401(k) eligibility is used as our treatment variable, it is important to note that our estimate of interest is now the average intention to treat.

We randomly sample (without replacement) with sample size varying from 20% to 50% of the data for the inference set and use the remaining data as our training set. With the randomly sampled training and inference sets, we calculate our mean difference estimate and doubly robust estimate for the average intention to treatment. Finally, we repeat this process 100 times to generate a distribution of the estimates.

The results are summarized in Figure 3 and Table 3. It is seen that the distributions of both estimates are uni-modal and close to symmetric, which is similar to what we have observed in the simulation studies. Compared to the results in [7], both of our estimators have distributions concentrating around the ATE estimate obtained in [7] for their quadratic spline specification without variable selection of 8093. However, our estimates have larger robust standard deviations. This is expected because our methods rely on sample splitting, and as revealed in Theorems 2 and 4, the convergence rates are determined by the inference set size, which we vary from 20% to 50% of the total data in our application, whereas [7]
used the entire sample and bootstrap to estimate the robust standard deviation. Comparing our mean difference estimate with our doubly robust estimate, we see that the latter has larger standard deviations which is consistent with our theory and our simulation studies. In addition, we observe that the estimates from the ReLU network have longer-tailed distributions. Our empirical results also suggest that the intention to treat effect is indeed significantly different from zero. Results corresponding to different numbers of epochs are included in Section C of the Supplementary Material.

| Inference Proportion | Estimate Type          | Activation | Median | Robust SD |
|----------------------|------------------------|------------|--------|-----------|
|                      | Difference of Means Estimate | ReLU       | 7780   | 2362      |
|                      |                        | Sigmoid    | 6911   | 2442      |
| 0.2                  | Doubly Robust Estimate | ReLU       | 7440   | 4488      |
|                      |                        | Sigmoid    | 8025   | 3384      |
|                      | Difference of Means Estimate | ReLU       | 7400   | 2669      |
|                      |                        | Sigmoid    | 6659   | 2036      |
| 0.3                  | Doubly Robust Estimate | ReLU       | 8127   | 3460      |
|                      |                        | Sigmoid    | 7723   | 2289      |
|                      | Difference of Means Estimate | ReLU       | 6764   | 2429      |
|                      |                        | Sigmoid    | 6764   | 2429      |
| 0.4                  | Doubly Robust Estimate | ReLU       | 7614   | 3110      |
|                      |                        | Sigmoid    | 7614   | 2035      |
|                      | Difference of Means Estimate | ReLU       | 7473   | 3743      |
|                      |                        | Sigmoid    | 7473   | 3743      |
| 0.5                  | Doubly Robust Estimate | ReLU       | 7934   | 4051      |
|                      |                        | Sigmoid    | 7603   | 1872      |

Table 3: The real data results corresponding to Figure 3 for 800 training epochs. Results for different training lengths can be found in Section C of the Supplementary Material.

### 6 Discussions

In this paper, we have considered the estimation and inference of ATE using deep neural networks. Under the potential outcomes framework, the observed response follows a non-parametric mean regression model, and ATE can be written as the expected difference of the mean regression function corresponding to the treatment and control groups. We have proposed to use DNN to learn the mean regression function, and construct the ATE estimate based on the DNN estimate. We have also derived the asymptotic normality of the ATE estimate using the idea of sample splitting. These ideas and results are further extended to the doubly robust estimator based on the inverse propensity score weighting. Simulation
Figure 3: The scaled density of the ATE estimate over 100 replications for different training sample size proportions and different activation functions. The red curves correspond to the DNN estimate defined in (10) and the blue curves correspond to the doubly robust estimate defined in (15). The red vertical line is the ATE estimate reported in [7] from the quadratic spline specification without variable selection of 8093. The rows in the figure correspond to different sizes of the inference set varying from 20% to 50% of the data. In this figure, both estimates come from networks trained for 800 epochs. Results for different training lengths can be found in Section C of the Supplementary Material.

studies and a real data application demonstrate the practical utilities of our methods. The current theory excludes the ReLU activation because of the smoothness assumption required in establishing the main results. Developing theory for more general activation functions is an interesting topic for future study. In addition, our current consistency rates are derived for functions with finite smoothness parameter $s$. We conjecture that the rates in Propositions 1 and 2 can be improved to nearly parametric rate when $s = \infty$. We leave such study for future investigation.

A Proofs of main results

We provide the proofs of Theorems 1–4, Propositions 1–2, and Corollary 3 in this Appendix. The remaining proofs and additional technical details are contained in the Supplementary Material. Throughout the paper, we use $C$ to denote a generic positive constant whose value may change from line to line.

A.1 Proof of Proposition 1

Observe that the difference between our ATE estimator $\hat{\tau}_D$ and the true value of the ATE $\tau$ consists of two major parts: the approximation error and the estimation error. The first part comes from the fact that $m_D(x, t)$ can be generally biased for nonparametric function
approximation, while the second part is because we estimate the population expectation
based on a given sample of size $n_D = |D|$. There is also an interplay between these two
parts. The theoretical results in [5] have tackled the approximation side, and our goal
here is to bound the estimation error given the regression function $m_D(x, t)$. Since the
regression function varies as $n_D \to \infty$, we focus on bounding the error for all possible
learned regression functions in the function class $H^{(l)}$ to accommodate the approximation
process. This is possible thanks to the relatively limited complexity of function class $H^{(l)}$, or
more precisely, the bound on the covering number of $H^{(l)}$ according to the learning theory
literature. Meanwhile, the correlation between the treatment group and the control group
makes no difference. Due to the symmetry of the two groups in estimation, the bounds for
one part can be naturally applied to the other part. Thus, for simplicity, we focus only on
one part, e.g., the treatment group, in our technical analysis. Throughout the proof, we will
use the notation $D = \{(X_i, T_i, Y_i)\}_{i=1}^{n_D}$ to denote the available data set. We will also drop
the subscript and write $n := n_D$.

Specifically, to bound $|\tau - \hat{\tau}_D|$, the treatment part and the control part can be separated as

$$|\tau - \hat{\tau}_D| \leq |E_X m(X, 1) - \frac{1}{n} \sum_{i=1}^{n} m_D(X_i, 1)|$$

$$+ |E_X m(X, 0) - \frac{1}{n} \sum_{i=1}^{n} m_D(X_i, 0)|,$$

(37)

where $E_X$ represents the expectation over an independent data point $X$ from the same
distribution as $X_1$. For the treatment part of (37), we have

$$|E_X m(X, 1) - \frac{1}{n} \sum_{i=1}^{n} m_D(X_i, 1)| \leq |E_X [m(X, 1) - m_D(X, 1)]|$$

$$+ |E_X m_D(X, 1) - \frac{1}{n} \sum_{i=1}^{n} m_D(X_i, 1)|.$$

An application of Theorem 1 in [5] leads to

$$E_D E_{X,T} |m(X, T) - m_D(X, T)|^2 \leq C_2 (\log n)^3 n^{-\frac{2\gamma}{2\gamma + p}}$$

(38)

with $C_2$ some positive constant for $n$ sufficiently large, where $E_D$ stands for the expectation
over data in $D$. This immediately entails that

$$E_{X,T} |m(X, T) - m_D(X, T)|^2 = o_P \left((\log n)^4 n^{-\frac{2\gamma}{2\gamma + p}}\right)$$

(39)
by Chebyshev’s inequality. This together with Condition 1(iv) ensures that

\[
\|E_X[m(X, 1) - m_D(X, 1)]\|
\leq \sqrt{E_X[m(X, 1) - m_D(X, 1)]^2}
\leq \frac{1}{\sqrt{\delta}} \sqrt{E_X\left\{|m(X, 1) - m_D(X, 1)|^2 \varepsilon(X) + |m(X, 0) - m_D(X, 0)|^2(1 - \varepsilon(X))\right\}}
\leq \frac{1}{\sqrt{\delta}} \sqrt{E_X, T|m(X, T) - m_D(X, T)|^2}
= o_P((\log n)^2/n^{1+\varepsilon}). \tag{40}
\]

On the other hand, from Theorem 9.1 in [14], one can bound the difference between the empirical average and its expectation as

\[
P\left(\|E_X m_D(X, 1) - \frac{1}{n} \sum_{i=1}^n m_D(X_i, 1)\| > \varepsilon_n\right)
\leq P\left(\sup_{\hat{m} \in H(t)} \|E_X \hat{m}(X, 1) - \frac{1}{n} \sum_{i=1}^n \hat{m}(X_i, 1)\| > \varepsilon_n\right)
\leq 8E_{\mu_n}[N(\varepsilon_n, H(t), L_1(\mu_n))] \exp\left(-\frac{n\varepsilon_n^2}{128}\right), \tag{41}
\]

where \(N(\varepsilon_n, H(t), L_1(\mu_n))\) stands for the covering number of the function class \(H(t)\) with metric \(\|f\|_{L_1(\mu_n)} = \mu_n(|f|) = \frac{1}{n} \sum_{i=1}^n |f(X_i)|\) at scale \(\varepsilon_n > 0\) and \(\mu_n\) the empirical measure. It follows from the fundamental theory of covering numbers that

\[
E_{\mu_n}[N(\varepsilon_n, H(t), L_1(\mu_n))] \leq N(\varepsilon_n, H(t), \|\cdot\|_\infty) \tag{42}
\]

and

\[
N(\varepsilon_n, H(t), \|\cdot\|_\infty) \leq \exp(C_3(\log n)M) \tag{43}
\]

with some positive constant \(C_3\), given that \(\varepsilon \geq \frac{1}{\sqrt{n}}\) for some positive constant \(C_4\); see, e.g., Lemma 2 in [5].

With the choice of \(\varepsilon_n = \sqrt{\frac{128 \log(n)N(\frac{1}{n}, H(t), \|\cdot\|_\infty))}{\log n + C_4(\log n)M}}\), one can deduce that

\[
P\left(\|E_X m_D(X, 1) - \frac{1}{n} \sum_{i=1}^n m_D(X_i, 1)\| > C_5\sqrt{\frac{\log n + C_4(\log n)M}{n}}\right)
\leq P\left(\|E_X m_D(X, 1) - \frac{1}{n} \sum_{i=1}^n m_D(X_i, 1)\| > \varepsilon_n\right)
\leq 8E_{\mu_n}[N(\varepsilon_n, H(t), L_1(\mu_n))] \exp\left(-\frac{n\varepsilon_n^2}{128}\right)
\leq \frac{8N(\varepsilon_n, H(t), \|\cdot\|_\infty)}{n \cdot N(\frac{1}{\sqrt{n}}, H(t), \|\cdot\|_\infty)}
\leq \frac{8}{n}, \tag{44}
\]

24
where $C_5$ is some positive constant. Here, the first inequality in (44) results from the fact that $\epsilon_n \leq C_5 \sqrt{\frac{\log n + C_3 (\log n) M}{n}}$ holds for some positive constant $C_5$. The second and third inequalities are implied by inequalities (41), (42), and (43). Finally, the last inequality is due to the monotone decreasing property of the covering number with respect to the scale. Hence, we can obtain by Condition 1(iii) that

\[
\left| \mathbb{E} m_D (X, 1) - \frac{1}{n} \sum_{i=1}^{n} m_D (X_i, 1) \right| = o_P \left( \sqrt{\log(n) n^{-\frac{2s}{2s+p^*}}} \right). \tag{45}
\]

Combining the above bounds in (40) and (45) yields

\[
\left| \mathbb{E} m(X, 0) - \frac{1}{n} \sum_{i=1}^{n} m_D (X_i, 0) \right| = o_P \left( \sqrt{\frac{2s}{2s+p^*}} \right). \tag{46}
\]

Similarly, it can be derived for the control part that

\[
\left| \mathbb{E} m(X, 0) - \frac{1}{n} \sum_{i=1}^{n} m_D (X_i, 0) \right| = o_P \left( \sqrt{\frac{2s}{2s+p^*}} \right). \tag{47}
\]

Therefore, in view of (37), (46), and (47), we have

\[
\left| \tau - \hat{\tau}_D \right| = o_P \left( \sqrt{\frac{2s}{2s+p^*}} \right), \tag{48}
\]

which completes the proof of Proposition 1.

A.2 Proof of Theorem 1

The high-level idea of the proof has been summarized in the main text just before Theorem 1. For the ease of presentation, we write $\mathcal{D}_2 = \{(X_i, T_i, Y_i)\}_{i=1}^{n}$. Let us consider the decomposition

\[
\sqrt{n} (\hat{\tau}(\mathcal{D}_1, \mathcal{D}_2) - \tau) = \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(X_i, 1) - \mathbb{E} X m(X, 1) \right)
- \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(X_i, 0) - \mathbb{E} X m(X, 0) \right)
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( m_D (X_i, 1) - m(X_i, 1) \right) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( m_D (X_i, 0) - m(X_i, 0) \right)
:= A_1 - A_0 + B_1 - B_0. \tag{49}
\]

The first two terms together $A_1 - A_0$ can be written as the sum of i.i.d random variables with bounded variance. Thus, an application of the classical central limit theorem (CLT) leads to

\[
A_1 - A_0 \overset{p}{\to} N(0, \sigma^2). \tag{50}
\]

We next prove that $B_1 = o_P(1)$ and $B_0 = o_P(1)$. Then these results together with (50)
can complete the proof of this theorem. Since the proofs for terms $B_1$ and $B_0$ are almost identical, we only show the former. First, since $\mathcal{D}_1$ and $\mathcal{D}_2$ are independent, each containing i.i.d. observations, an application of Chebyshev’s inequality entails that for any $x > 0$, it holds that

$$
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} \left( m_{\mathcal{D}_1}(X_i, 1) - m(X_i, 1) - \mathbb{E}_X[m_{\mathcal{D}_1}(X, 1) - m(X, 1)] \right) \right| > n^{-1/2}x \middle| \mathcal{D}_1 \right) 
\leq \sum_{i=1}^{n} \mathbb{E}_X \left| m_{\mathcal{D}_1}(X_i, 1) - m(X_i, 1) \right|^2 - n(\mathbb{E}_X[m_{\mathcal{D}_1}(X, 1) - m(X, 1)])^2 
\leq \frac{\mathbb{E}_X \left| m_{\mathcal{D}_1}(X_i, 1) - m(X_i, 1) \right|^2}{x^2}.
$$

Noting that $|\mathcal{D}_1| = n^\gamma$, by (38) we have

$$
\mathbb{E}_{\mathcal{D}_1} \mathbb{E}_X \left| m_{\mathcal{D}_1}(X_i, 1) - m(X_i, 1) \right|^2 \leq C(\gamma \log n)^3 n^{-\frac{2\gamma s}{2r+s}}.
$$

Taking $x = (\log n)^2 n^{-\frac{\gamma}{2r+s}}$ and by the properties of the conditional expectation, we can deduce that

$$
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} \left( m_{\mathcal{D}_1}(X_i, 1) - m(X_i, 1) - \mathbb{E}_X m_{\mathcal{D}_1}(X, 1) + \mathbb{E}[m(X, 1)] \right) \right| > n^{-1/2}x \right) 
\leq \frac{\mathbb{E}_{\mathcal{D}_1} \mathbb{E}_X \left| m_{\mathcal{D}_1}(X_i, 1) - m(X_i, 1) \right|^2}{x^2} 
\to 0.
$$

This result along with (40) entails that

$$
|B_1| = \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( m_{\mathcal{D}_1}(X_i, 1) - m(X_i, 1) \right) \right| = o_p((\log n)^2 n^{-\frac{\gamma}{2r+s}}) = o_P(1),
$$

which concludes the proof of Theorem 1.

### A.3 Proof of Theorem 2

Denote by $\sigma^2(\mathcal{D}_1)$ the population variance for $\hat{\tau}(\mathcal{D}_1)$ conditional on $\mathcal{D}_1$; that is,

$$
\sigma^2(\mathcal{D}_1) = \text{Var}(\hat{\tau}(\mathcal{D}_1)|\mathcal{D}_1),
$$

where $\hat{\tau}(\mathcal{D}_1) = m_{\mathcal{D}_1}(X, 1) - m_{\mathcal{D}_1}(X, 0)$. Hence, we have

$$
|\hat{\sigma}^2(\mathcal{D}_1, \mathcal{D}_2) - \sigma^2| \leq |\hat{\sigma}^2(\mathcal{D}_1, \mathcal{D}_2) - \sigma^2(\mathcal{D}_1)| + |\sigma^2(\mathcal{D}_1) - \sigma^2|.
$$

We can obtain that $m(x, t)$ is bounded on its domain by the bounded support assumption in Condition 1(i) and the smoothness assumption on the mean regression function $m(x, t)$.
For the second term on the right-hand side of (54), it holds that
\[
|\sigma^2(D_1) - \sigma^2| = |\mathbb{E}_n[m_{D_1}(X, 1) - \mathbb{E}_n m_{D_1}(X, 1) - m_{D_1}(X, 0) + \mathbb{E}_n m_{D_1}(X, 0)]^2|D_1]
- \mathbb{E}_n[m_{D_1}(X, 1) - m_{D_1}(X, 0) + \mathbb{E}_n m_{D_1}(X, 0)]^2|D_1]
\leq C \log(n^\gamma) |\mathbb{E}_n[m_{D_1}(X, 1) - m_{D_1}(X, 1)] - \mathbb{E}_n[m_{D_1}(X, 1) - m_{D_1}(X, 1)]|
+ \mathbb{E}_n|m_{D_1}(X, 0) - m(X, 0)| - \mathbb{E}_n[m_{D_1}(X, 0) - \mathbb{E}_n m_{D_1}(X, 0)]
\leq 2C \log(n^\gamma) |\mathbb{E}_n[m_{D_1}(X, 1) - m(X, 1)] + \mathbb{E}_n[m_{D_1}(X, 0) - m(X, 0)]|
\]
where $C \log(n^\gamma)$ comes from the truncation step involved in the definition of $m_{D_1}$. Then an application of inequality (40) yields
\[
\mathbb{E}_n|m_{D_1}(X, 1) - m(X, 1)| \leq \left\{ \mathbb{E}_n|m_{D_1}(X, 1) - m(X, 1)|^2 \right\}^{1/2} = o_P((\log n^\gamma) \frac{n - \gamma}{2 + \gamma}).
\]
The same result can be obtained for $\mathbb{E}_n|m_{D_1}(X, 0) - m(X, 0)|$ using similar arguments. Thus, we can obtain that
\[
|\sigma^2(D_1) - \sigma^2| = o_P((\log n^\gamma) \frac{n - \gamma}{2 + \gamma}).
\]  

(55)

The first term on the right-hand side of (54) can be tackled with an application of the weak law of large numbers for a triangular array. In particular, we set $Z_{n,i} = \hat{\tau}_i(D_1)$ for $i \in \mathcal{D}_2$ with $\hat{\tau}_i(D_1)$ defined below (10). Observe that $|Z_{n,i}|$ is upper bounded by $C \log(n^\gamma)$ with some positive constant $C$. Then, by Chebyshev’s inequality conditional on $D_1$ for arbitrary $\epsilon > 0$, it holds that
\[
\mathbb{P}\left(\left|\sum_{i \in \mathcal{D}_2}(Z_{n,i} - \mathbb{E}[Z_{n,i}|D_1])\right| > \epsilon \right)
= \mathbb{E}\left[\mathbb{P}\left(\left|\sum_{i \in \mathcal{D}_2}(Z_{n,i} - \mathbb{E}[Z_{n,i}|D_1])\right| > \epsilon \right) |D_1]\right]
\leq \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}[Z_{n,i}^2 | D_1]\right] \leq \frac{C^2 \gamma^2 \log^2(n)}{\epsilon^2 n}
\]
and similarly,
\[
\mathbb{P}\left(\left|\sum_{i \in \mathcal{D}_2}(Z_{n,i}^2 - \mathbb{E}[Z_{n,i}^2 | D_1])\right| > \epsilon \right) \leq \frac{C^4 \gamma^4 \log^4(n)}{\epsilon^2 n}.
\]
By choosing $\epsilon = \frac{\log^3(n)}{\sqrt{n}}$, we have with probability at least $1 - \frac{C^4 \gamma^4}{\log^4(n)} - \frac{C^3 \gamma^3}{\log^3(n)}$ that
\[
\left|\sum_{i \in \mathcal{D}_2}(Z_{n,i} - \mathbb{E}[Z_{n,i}|D_1])\right| \leq \frac{\log^3(n)}{\sqrt{n}} \quad \text{and} \quad \left|\sum_{i \in \mathcal{D}_2}(Z_{n,i}^2 - \mathbb{E}[Z_{n,i}^2 | D_1])\right| \leq \frac{\log^3(n)}{\sqrt{n}}.
\]
Thus, we can deduce that

\[
|\hat{\sigma}^2(D_1, D_2) - \sigma^2(D_1)| = \left| \frac{n}{n-1} \left( \frac{1}{n} \sum_{i \in D_2} (Z_{n,i}^2 - \mathbb{E}[Z_{n,i}^2|D_1]) - \left( \frac{1}{n} \sum_{i \in D_2} Z_{n,i} \right)^2 + \mathbb{E}[\hat{Z}_{n,i}^2|D_1] \right) + \frac{1}{n-1} \sigma^2(D_1) \right|
\]

\[
\leq 2 \frac{1}{n} \sum_{i \in D_2} (Z_{n,i}^2 - \mathbb{E}[Z_{n,i}^2|D_1]) + 4C \log(n) \left| \frac{1}{n} \sum_{i \in D_2} (Z_{n,i} - \mathbb{E}[Z_{n,i}|D_1]|) + \frac{2\sigma^2(D_1)}{n} \right|
\]

\[
\leq 2 \frac{\log^3(n)}{\sqrt{n}} + \frac{4C\gamma \log^4(n)}{\sqrt{n}} + \frac{2\sigma^2}{n} + \frac{2}{n} |\sigma^2(D_1) - \sigma^2|
\]

for \( n \) large enough with probability at least \( 1 - \frac{C^4}{\log^2(n)} = \frac{C^4}{\log^4(n)} \).

Therefore, combining the bounds in (55) and (56) yields that

\[
|\hat{\sigma}^2(D_1, D_2) - \sigma^2| = o_P(\frac{\log^3(n)}{\sqrt{n}} + \frac{\log^4(n)}{\sqrt{n}} + O(\frac{1}{n}) + o_P((\log n)^3 n^{-\frac{4}{4+\gamma}})).
\]

Since we assume that \( \gamma > 1 + \frac{C^4}{25} \), it follows that

\[
|\hat{\sigma}^2(D_1, D_2) - \sigma^2| = o_P(n^{-1/2}(\log n)^4).
\]

The above consistency result together with Theorem 1 and Slutsky’s lemma completes the proof of Theorem 2.

### A.4 Proof of Proposition 2

Recall that we assume that \( |D_1| = n \) and for the \( i \)th observation in \( D_1 \), we denote it as \((X_i, T_i, Y_i)\). We start with the decomposition\(^3\)

\[
|\hat{\tau}_{DR,D_1}(\hat{\epsilon}, \hat{m}) - \tau| = \frac{1}{n} \sum_{i \in D_1} (\phi_i(\hat{\epsilon}, \hat{m}) - \phi_i(\hat{\epsilon}, m)) - \frac{1}{n} \sum_{i \in D_1} (\psi_i(\hat{\epsilon}, \hat{m}) - \psi_i(\hat{\epsilon}, m))
\]

\[
+ \frac{1}{n} \sum_{i \in D_1} (\phi_i(\hat{\epsilon}, m) - \phi_i(e, m)) - \frac{1}{n} \sum_{i \in D_1} (\psi_i(\hat{\epsilon}, m) - \psi_i(e, m))
\]

\[
+ \frac{1}{n} \sum_{i \in D_1} [(\phi_i(e, m) - \mathbb{E}\phi(e, m)) - (\psi_i(e, m) - \mathbb{E}\psi(e, m))]
\]

\[
:= E_1 - E_0 + F_1 - F_0 + G.
\]

We will show that each term on the right-hand side of (58) is an \( o_P(1) \) term with some rate of convergence. Since the treatments for terms \( E_1 \) and \( E_0 \) are similar, we will only deal with the former one. The explicit form of term \( E_1 \) can be written as

\[
E_1 = \frac{1}{n} \sum_{i \in D_1} (\frac{T_i}{\hat{\epsilon}(X_i)} - 1)(m_1(X_i) - \hat{m}_1(X_i)).
\]

\(^3\)The subscripts \( D_i \) for \( \hat{\epsilon}_{D_i} \) and \( \hat{m}_{D_i} \) are omitted in this proof and the proofs of Theorem 3 and Theorem 4 for notational simplicity.
Based on (i) of Condition 3, we can deduce that

\[
|E_1| \leq \frac{1}{n} \sum_{i \in D_1} 2C_2 \log(n)|m_1(X_i) - \hat{m}_1(X_i)|
\]

\[
\leq 2C_2 \log(n) \left( \frac{1}{n} \sum_{i \in D_1} |m_1(X_i) - \hat{m}_1(X_i)|^2 \right)^{1/2}
\]

\[
\leq 2C_2 \log(n) \left( \frac{1}{n} \sum_{i \in D_1} |m_1(X_i) - \hat{m}_1(X_i)|^2 - \mathbb{E}_X|m_1(X) - \hat{m}_1(X)|^2 \right)
\]

\[
+ \mathbb{E}_X|m_1(X) - \hat{m}_1(X)|^2 \right)^{1/2}.
\]

The first term inside the square root on the right-hand side above can be bounded by applying Theorem 9.1 in [14], using arguments similar to those used for obtaining inequality (41). Specifically, let us define a new function class

\[
\tilde{H}^{(l)} = \{g : g(x, t) = (\text{trunc}(\hat{m}(x, t), C \log n) - m(x, t))^2 \text{ with } \hat{m} \in H^{(l)} \}.
\]

Then for \( n \) sufficiently large, it holds that

\[
\mathbb{E}_{\mu_n}[\mathbb{N}(\epsilon_n, \tilde{H}^{(l)}, L_1(\mu_n))] \leq \mathbb{N}(\frac{\epsilon_n}{2C(\log n)}, H^{(l)}, ||\cdot||_\infty).
\]

Thus, an application of similar arguments as in the proof of (41) leads to

\[
\left| \frac{1}{n} \sum_{i \in D_1} |\hat{m}_1(X_i) - m_1(X_i)|^2 - \mathbb{E}_X|\hat{m}_1(X) - m_1(X)|^2 \right| = o_P(\sqrt{\log(n)n^{-\frac{s_2^2}{2+sp^*}}})
\]

The expectation term above can be bounded similar to (39). Hence, we can obtain that

\[
|E_1| = o_P(\log^3(n)n^{-\frac{s_2^2}{2+sp^*}}) + o_P(\log^{5/4}(n)n^{-\frac{s_2^2}{2+sp^*}}) = o_P(1).
\]

(60)

As for term \( F_1 \), we can write it as

\[
F_1 = \frac{1}{n} \sum_{i \in D_1} \left( \frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)} \right)T_i(Y_i - m_1(X_i)) = \frac{1}{n} \sum_{i \in D_1} \left( \frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)} \right)T_i \epsilon_i,
\]

which entails that \( EF_1 = 0 \). Due to (i) of Condition 3, it follows that

\[
\mathbb{E}[F_1^2 | X_1, X_2, \ldots, X_n] \leq \mathbb{V}ar[\varepsilon] \frac{1}{n} \sum_{i \in D_1} \left( \frac{e(X_i) - \hat{e}(X_i)}{\hat{e}(X_i)e(X_i)} \right)^2
\]

\[
\leq \frac{\mathbb{V}ar[\varepsilon] C_2^2 \log^2(n)}{\delta^2} \frac{1}{n} \sum_{i \in D_1} (e(X_i) - \hat{e}(X_i))^2.
\]

Then an application of (ii) of Condition 3 implies that

\[
\mathbb{E}[F_1^2] = o_P(1),
\]

(63)
which shows that term $F_1$ vanishes in probability asymptotically thanks to Chebyshev’s inequality.

Applying similar arguments to terms $E_0$ and $F_0$, we can obtain that

$$|E_0|, |F_0| = o_P(1).$$

(64)

On the other hand, due to the boundedness of both $\phi_i(e, m)$ and $\psi_i(e, m)$, an application of the law of large numbers entails that

$$G = o_P(1).$$

(65)

Therefore, it follows that the doubly robust estimator $\hat{\tau}_{DR,D_1}(\hat{e}, \hat{m})$ is a consistent estimator of $\tau$, which concludes the proof of Proposition 2.

A.5 Proof of Theorem 3

The proof idea is similar to that of Theorem 1, which begins with the decomposition

$$\sqrt{n}|\hat{\tau}_{DR,D_2}(\hat{e}, \hat{m}) - \tau| = \frac{1}{\sqrt{n}} \sum_{i \in D_2} [\phi_i(\hat{e}, \hat{m}) - \phi_i(\hat{e}, m)] - \frac{1}{\sqrt{n}} \sum_{i \in D_2} [\psi_i(\hat{e}, \hat{m}) - \psi_i(\hat{e}, m)]$$

$$+ \frac{1}{\sqrt{n}} \sum_{i \in D_2} [\phi_i(e, m) - \phi_i(e, m)] - \frac{1}{\sqrt{n}} \sum_{i \in D_2} [\psi_i(e, m) - \psi_i(e, m)]$$

$$+ \frac{1}{\sqrt{n}} \sum_{i \in D_2} [\phi_i(e, m) - \mathbb{E}(Y, X, T) \phi(e, m)]$$

$$- \frac{1}{\sqrt{n}} \sum_{i \in D_2} [\psi_i(e, m) - \mathbb{E}(Y, X, T) \psi(e, m)]$$

$$:= H_1 - H_0 + I_1 - I_0 + J_1 - J_0.$$  

(66)

We first consider term $H_1$ above. Since $D_1$ and $D_2 = \{(X_i, T_i, Y_i)\}_{i=1}^n$ are independent, it follows from Chebyshev’s inequality that for any $x > 0$,

$$\mathbb{P}(H_1 - \mathbb{E}[H_1|D_1] > x|D_1) \leq \frac{\sum_{i \in D_2} \mathbb{E}[(\phi_i(\hat{e}, \hat{m}) - \phi_i(\hat{e}, m))^2|D_1]}{nx^2}$$

$$\leq \frac{\mathbb{E}[(\mathbb{E}[(Y, X, T)] - 1)(\hat{m}_1(X_1) - m_1(X_1))]^2|D_1]}{x^2}$$

$$\leq C^2_2 (\log n)^2 \mathbb{E}[(\hat{m}_1(X_1) - m_1(X_1))^2|D_1]/x^2,$$

where in the last step we have used (i) of Condition 3. By the definition of the conditional probability and (38), we can obtain that

$$\mathbb{P}(H_1 - \mathbb{E}[H_1|D_1] \geq x) \leq C^2_2 (\log n)^2 \mathbb{E}_{D_1} \mathbb{E}[(\hat{m}_1(X_1) - m_1(X_1))^2|D_1]/x^2$$

$$\leq C^2_2 (\log n)^2 o((\log n)^3 n^{-\frac{2\epsilon}{3\epsilon + \epsilon'}})/x^2.$$

30
By letting \( x = (\log n)^3 n^{-\frac{1}{2+\tau_p^*}} \), we have

\[
H_1 - \mathbb{E}[H_1|\mathcal{D}_1] = o_P((\log n)^3 n^{-\frac{1}{2+\tau_p^*}}).
\]

Further, we can deduce that

\[
|\mathbb{E}[H_1|\mathcal{D}_1]| = \sqrt{n|\mathbb{E}_{X_1,T_1}[\left(\frac{T_1}{\hat{e}(X_1)} - 1\right)(m_1(X_1) - \hat{m}_1(X_1))]|}
\leq C_2 \sqrt{n \log(n)}|\mathbb{E}_{X_1}[\left(e(X_1) - \hat{e}(X_1)\right)(m_1(X_1) - \hat{m}_1(X_1))]|,
\leq C_2 \sqrt{n \log(n)} \sqrt{\mathbb{E}_{X_1}[e(X_1) - \hat{e}(X_1)]^2} \sqrt{\mathbb{E}_{X_1}[m_1(X_1) - \hat{m}_1(X_1)]^2}.
\]

Combining (iii) of Condition 3, the assumption of \( p^* < 2s \), and inequality (38) results in

\[
\mathbb{E}[H_1|\mathcal{D}_1] = o_P(\sqrt{n \log(n)n^{-1/4}(\log n)^3 n^{-\frac{1}{2+\tau_p^*}}}) = o_P(1).
\]

Thus, the above results together entail that

\[
H_1 = (H_1 - \mathbb{E}[H_1|\mathcal{D}_1]) + \mathbb{E}[H_1|\mathcal{D}_1] = o_P((\log n)^3 n^{-\frac{1}{2+\tau_p^*}} + o_P(1)) = o_P(1).
\]

Similar arguments can be applied to term \( I_1 \). In particular, note that \( \mathbb{E}[I_1|\mathcal{D}_1] = 0 \). Also, it holds that

\[
\mathbb{P}(\{I_1 \geq x\}) = \mathbb{E}_{\mathcal{D}_1}\mathbb{P}(\{I_1 - \mathbb{E}[I_1|\mathcal{D}_1] \geq x|\mathcal{D}_1\})
\leq \mathbb{E}_{\mathcal{D}_1} \frac{\mathbb{E}[I_1^2|\mathcal{D}_1]}{x^2}
\leq \mathbb{E}_{\mathcal{D}_1} \mathbb{E}[|T_1(Y_1 - m_1(X_1))\hat{e}(X_1) - e(X_1)|^2|\mathcal{D}_1]/x^2
\leq \mathbb{E}_{\mathcal{D}_1} \mathbb{E}[\hat{e}(X_1) - e(X_1)|^2|\mathcal{D}_1]/x^2
\leq \mathbb{E}_{\mathcal{D}_1} \mathbb{E}[\hat{e}(X_1) - e(X_1)|^2|\mathcal{D}_1]/x^2
= o \left( (\log n)^2 n^{-1/2} \right)/x^2,
\]

where \((X_1, T_1, Y_1) \in \mathcal{D}_2\) is independent of \( \mathcal{D}_1 \). Taking \( x = (\log n)^2 n^{-1/4} \), we can obtain that

\[
I_1 = o_P\left( (\log n)^2 n^{-1/4} \right).
\]

Similarly, we can show that

\[
H_0 = o_P(1) \quad \text{and} \quad I_0 = o_P(1).
\]

Moreover, it holds that \( J_1 - J_0 \) converges in distribution to \( \mathcal{N}(0, \sigma_{DR}^2) \). Therefore, combining all these results yields that

\[
\begin{align*}
\hat{\tau}_{DR,n}(\hat{e}_{D_1}, \hat{m}_{D_1}) - \tau & \xrightarrow{d} \mathcal{N}(0, \sigma_{DR}^2), \quad (67)
\end{align*}
\]

where \( \sigma_{DR}^2 = \text{Var}_{X_1,T_1}[\phi(e, m) - \psi(e, m)] = \text{Var}(m_1(X) - m_0(X)) + \text{Var}(\epsilon) \mathbb{E}_{\epsilon}(\frac{1}{1 - \epsilon(X)}) \).

31
This completes the proof of Theorem 3.

A.6 Proof of Corollary 3

We only need to verify (i) and (iii) of Condition 3. Indeed, (i) of Condition 3 holds due to the truncation on \( \hat{e}_{D_1}(x) \). As for (iii) of Condition 3, we can show by Proposition 1, in which bound (38) can be applied to \( \hat{e}(X) \) as well, that

\[
\mathbb{E}_{D_1, \mathbb{E}_X} [\hat{e}_{D_1}(X) - e(X)]^2 \leq C \log^3(n)n^{-\frac{2n}{2n+p}}
\]

for some constant \( C \) and all sufficiently large \( n \). Since \( p^* < 2s_\epsilon \), the right-hand side of (68) is indeed an \( o(n^{-1/2}) \) term. Therefore, given (i) and (iii) of Condition 3, the desired conclusions of Corollary 3 follow from Theorem 3.

A.7 Proof of Theorem 4

Recall that \( D_2 = \{(X_i, T_i, Y_i)\}_{i=1}^n \). Let us define

\[
\sigma^2_{DR}(\hat{e}_{D_1}, \hat{m}_{D_1}) = \text{Var}[\phi_1(\hat{e}_{D_1}, \hat{m}_{D_1}) - \psi_1(\hat{e}_{D_1}, \hat{m}_{D_1})|D_1].
\]

Observe that

\[
\sigma^2_{DR} = \text{Var}[\phi_1(e, m) - \psi_1(e, m)]
\]

and \( \phi_1 \) and \( \psi_1 \) are defined on the observation \((X_1, T_1, Y_1) \in D_2\). Then an application of similar arguments as in the proof of Theorem 2 shows that \( \sigma^2_{DR}(\hat{e}_{D_1}, \hat{m}_{D_1}) \) is a consistent estimator of \( \sigma^2_{DR} \). It holds that

\[
\phi_1(\hat{e}_{D_1}, \hat{m}_{D_1}) - \psi_1(\hat{e}_{D_1}, \hat{m}_{D_1})
= \frac{T_1}{\hat{e}_{D_1}(X_1)} (m_1(X_1) - \hat{m}_{D_1,1}(X_1)) + \hat{m}_{D_1,1}(X_1)
- \frac{1 - T_1}{1 - \hat{e}_{D_1}(X_1)} (m_0(X_1) - \hat{m}_{D_1,0}(X_1)) - \hat{m}_{D_1,0}(X_1)
+ \left( \frac{T_1}{\hat{e}_{D_1}(X_1)} - \frac{1 - T_1}{1 - \hat{e}_{D_1}(X_1)} \right) \varepsilon_1.
\]

Since \( \varepsilon_1 \) is independent of \((X_1, T_1)\) and \( D_1 \) and has mean zero, it follows that

\[
\sigma^2_{DR}(\hat{e}_{D_1}, \hat{m}_{D_1}) = \text{Var}\left( \frac{T_1}{\hat{e}_{D_1}(X_1)} (m_1(X_1) - \hat{m}_{D_1,1}(X_1)) \right)
+ \hat{m}_{D_1,1}(X_1) - \frac{1 - T_1}{1 - \hat{e}(X_1)} (m_0(X_1) - \hat{m}_{D_1,0}(X_1)) - \hat{m}_{D_1,0}(X_1)
+ \text{Var}\left( \left( \frac{T_1}{\hat{e}_{D_1}(X_1)} - \frac{1 - T_1}{1 - \hat{e}_{D_1}(X_1)} \right) \varepsilon_1 \bigg| D_1 \right)
:= I_1 + I_2.
\]
Similarly, we can show that
\[
\sigma_{DR}^2 = \text{Var}\left( m_1(X_1) - m_0(X_1) \right) + \text{Var}\left( \left( \frac{T_1}{e(X_1)} - \frac{1 - T_1}{1 - e(X_1)} \right) \varepsilon_1 \right)
\]
\[= I_1 + II_2.\]

First, let us consider term $I_2 - II_2$. By the independence of $\varepsilon_1$ with $(X_1, T_1)$ and $D_1$, we have
\[
|I_2 - II_2| = \text{Var}(\varepsilon_1) \left| \mathbb{E}\left[ \left( \frac{T_1}{e_{D_1}(X_1)} - \frac{1 - T_1}{1 - e_{D_1}(X_1)} \right)^2 \right] \right| - \mathbb{E}\left[ \left( \frac{T_1}{e(X_1)} - \frac{1 - T_1}{1 - e(X_1)} \right)^2 \right] - \mathbb{E}\left[ e(X_1) \right] - \mathbb{E}\left[ \frac{1}{e(X_1)} - \frac{1}{1 - e(X_1)} \right]
\]
\[\leq \text{Var}[\varepsilon_1] C_2^2 \log^2(n) \mathbb{E}[|\hat{e}_{D_1}(X_1) - e(X_1)||D_1],\]
where in the last step we have used the boundedness assumption of $\hat{e}_{D_1}$ stated in Condition 3(ii). Furthermore, by Condition 3(iii) and the fact that $Y_1$ is a sub-Gaussian random variable, it follows from Chebyshev’s inequality that
\[
|I_2 - II_2| = o_P(\log^2(n)n^{-1/4}).
\]

Next we analyze term $I_1 - II_1$. Note that the random variable inside the variance in $I_1$ can be upper bounded by $C \log^2 n$ almost surely with respect to $X_1$ with $C$ some generic positive constant. By the variance representation
\[
\text{Var}(R_1) - \text{Var}(R_2) = \mathbb{E}[(R_1 + R_2)(R_1 - R_2)] - (\mathbb{E}R_1 - \mathbb{E}R_2)(\mathbb{E}R_1 + \mathbb{E}R_2)
\]
for any random variables $R_1$ and $R_2$ and some basic calculations, we can deduce that
\[
|I_1 - II_1| \leq C(\log n)^2 \mathbb{E}\left[ \left| \left( \frac{T_1}{e_{D_1}(X_1)} - 1 \right) (m_1(X_1) - \hat{m}_{D_1,1}(X_1)) \right| \right] + \mathbb{E}\left[ \left( \frac{1 - T_1}{1 - e_{D_1}(X_1)} - 1 \right) (m_1(X_1) - \hat{m}_{D_1,1}(X_1)) \right] \left| D_1 \right|
\]
\[\leq C(\log n)^3 \left\{ \mathbb{E}[|m_1(X_1) - \hat{m}_{D_1,1}(X_1)||D_1] + \mathbb{E}[|m_0(X_1) - \hat{m}_{D_1,0}(X_1)||D_1] \right\}.
\]

In view of (40), it holds that
\[
|I_1 - II_1| = o_P(\log^5(n)n^{-\frac{1}{3+\rho}}).
\]

Thus, combining the above results leads to
\[
|\sigma_{DR}(\hat{e}_{D_1}, \hat{m}_{D_1}) - \sigma_{DR}^2| = o_P(\log^5(n)n^{-\frac{1}{3+\rho}} + \log^2(n)n^{-1/4}).
\]
Denote by $Z_{n,i} = \hat{\tau}_i(\hat{e}_{D_1}, \hat{m}_{D_1})$ (c.f. (30)). Then similar arguments as in the proof of Theorem 2 can be applied with the aid of the law of large numbers. In particular, we can obtain that

$$|\hat{\sigma}^2_{D_1,D_2}(\hat{e}_{D_1}, \hat{m}_{D_1}) - \sigma^2_{D_2}(\hat{e}_{D_1}, \hat{m}_{D_1})| = o_P\left(\frac{\log^7(n)}{\sqrt{n}}\right) + \sigma^2_{D_2}(\hat{e}, \hat{m}) \frac{n}{n-1}.$$

Together with bound (69), the above inequality yields that

$$|\hat{\sigma}^2_{D_1,D_2}(\hat{e}_{D_1}, \hat{m}_{D_1}) - \sigma^2_{D_2}| = o_P\left(\frac{\log^7(n)}{\sqrt{n}}\right) + o_P(\log^5(n)n^{-\frac{1}{2+s^*}}) + o_P(\log^2(n)n^{-1/4}).$$

With the assumption of $p^* < 2s$, the above bound can be further simplified as

$$|\hat{\sigma}^2_{D_1,D_2}(\hat{e}_{D_1}, \hat{m}_{D_1}) - \sigma^2_{D_2}| = o_P(\log^2(n)n^{-1/4}). \quad (70)$$

Therefore, the asymptotic normality of $\sqrt{n}(\hat{\tau}_{D_2,D_2}(\hat{e}_{D_1}, \hat{m}_{D_1}) - \tau) / \hat{\sigma}_{D_1,D_2}(\hat{e}_{D_1}, \hat{m}_{D_1})$ holds by Slutsky’s lemma, which concludes the proof of Theorem 4.

References

[1] Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

[2] Alberto Abadie and Matias D. Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10(1):465–503, 2018.

[3] Susan Athey. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 5–6. ACM, 2015.

[4] Susan Athey and Guido W. Imbens. Machine learning methods for estimating heterogeneous causal effects. *Stat*, 1050(5), 2015.

[5] Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.*, 47:2261–2285, 2019.

[6] Benedikt Bauer and Michael Kohler. Supplement to “on deep learning as a remedy for the curse of dimensionality in nonparametric regression”. *Annals of Statistics*, 2019.

[7] Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[8] Daniel J. Benjamin. Does 401(K) eligibility increase saving? evidence from propensity score subclassification. *Journal of Public Economics*, 87(5):1259–1290, 2003.
[9] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[10] Victor Chernozhukov and Christian Hansen. The effects of 401(K) participation on the wealth distribution: an instrumental quantile regression analysis. *Review of Economics and Statistics*, 86(3):735–751, 2004.

[11] Emre Demirkaya, Yingying Fan, Lan Gao, Jinchi Lv, Patrick Vossler, and Jingbo Wang. Nonparametric inference of heterogeneous treatment effects with two-scale distributional nearest neighbors. *arXiv preprint arXiv:1808.08469*, 2021.

[12] Jianqing Fan, Kosuke Imai, Han Liu, Yang Ning, and Xiaolin Yang. Improving covariate balancing propensity score: A doubly robust and efficient approach. *Working paper*, 2016.

[13] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 03 2011.

[14] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

[15] Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, March 2009.

[16] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard S. Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, 2017.

[17] James M. Poterba and Steven F. Venti. 401(K) plans and tax-deferred saving. In *Studies in the Economics of Aging*, NBER Chapters, pages 105–142. National Bureau of Economic Research, Inc, June 1994.

[18] James M. Poterba, Steven F. Venti, and David A. Wise. Do 401(K) contributions crowd out other personal saving? *Journal of Public Economics*, 58(1):1–32, 1995.

[19] Franco Scarselli and Ah Chung Tsoi. Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural Networks*, 11(1):15–37, 1998.

[20] Jasjeet S. Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *Oxford Handbook of Political Methodology, 2008*. 

35
[21] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning, pages 3076–3085, 2017.

[22] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113:1228–1242, 2018.

[23] Haixiang Zhang, Yinan Zheng, Zhou Zhang, Tao Gao, Brain Joyce, Grace Yoon, Wei Zhang, Joel Schwartz, Allan Just, Elena Colicino, Pantel Vokonas, Lihui Zhao, Jinchi Lv, Andrea Baccarelli, Lifang Hou, and Lei Liu. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics, 32:3150–3154, 2016.
This Supplementary Material contains the proofs of Corollary 1 and some technical lemmas, and additional numerical results for the simulation and real data examples in Sections 4–5. All the notation is the same as defined in the main body of the paper.

B Additional proofs and technical details

B.1 Proof of Corollary 1

The main idea of the proof is similar to that of the proof for Theorem 1. Using the same decomposition as in (49), it is seen that we only need to bound

\[ B_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m_{D_1}(X_i, 1) - m(X_i, 1)) \]

\[ B_0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m_{D_1}(X_i, 0) - m(X_i, 0)) \]

under the new conditions of Corollary 1. In the proof of Theorem 1, to bound terms \( B_1 \) and \( B_0 \) we have used Proposition 1 and the results established in [5]. We will establish parallel results under the conditions of Corollary 1 in the next subsection. Using Lemma 4 in Section B.3 (which contains parallel results to those in Proposition 1), we can deduce that

\[ |B_1| = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m_{D_1}(X_i, 1) - m(X_i, 1)) \]

\[ \leq o_P(\log(|D_1|)^2 + p'^{2}/2|D_1|^{-1/2}) + \sqrt{n} \|X(m_{D_1}(X, 1) - m(X, 1))\| \]

\[ \leq o_P((\log(|D_1|))^{2+p'/2}(|D_2|/|D_1|)^{1/2}) \]

\[ = o_P\left(\frac{(\log(n) + k\log(\log(n)))^{2+p'/2}}{(\log(n))^{k/2}}\right) = o_P(1). \]

Similarly, we can also obtain that \( |B_0| = o_P(1) \). Therefore, the asymptotic normality in Corollary 1 holds.

B.2 Some key lemmas for proving Corollary 1

In [5], \((s, C)\)-smooth functions for fixed \( s \) and \( C \) are investigated and the results derived therein involve several constants that depend on the smoothness parameter \( s \) implicitly. For \( m(x) \) satisfying Condition 2(i), it is also \((s, C)\)-smooth for any \( s \in \mathbb{N} \). Consequently, the result of Proposition 1 holds for each \( s \in \mathbb{N} \). However, since the constants involved in the proof of Proposition 1 are not uniform over all \( s \in \mathbb{N} \), the consistency rate therein may not hold uniformly over all \( s \in \mathbb{N} \). Because of this, we cannot simply send \( s \) to infinity to prove the results of Corollary 1. Instead, we adapt the proof ideas in [5] to establish our desired results.

The above arguments also help us understand the necessity of assuming the polynomial functional form in Condition 2(i). With such an assumption, the universal approximation
power of two-layer deep neural networks for polynomial functions established in [19] can be used to show that all the polynomial functions appearing in the definition of \( m(x) \) are uniformly well approximated. Hence, results parallel to Proposition 1, which are summarized in Lemma 4, can be obtained for \( m(x) \) satisfying Condition 2(i). Then Corollary 1 follows naturally. On the contrary, without the polynomial function assumption, we would likely encounter approximation errors that are nonuniform across \( s \) when using two-layer neural networks, which takes us back to the challenges discussed in the previous paragraph. This provides some justifications on Condition 2(i).

**Notation.** We first introduce some notation that will be used in our subsequent proofs. Let \( f^{(n)}(x) \) be the \( n \)th order derivative of function \( f : \mathbb{R} \rightarrow \mathbb{R} \) at \( x \). For a polytope \( K \subseteq \mathbb{R}^p \) bounded by hyperplanes \( u_j \cdot x + w_j \leq 0 \) \((j = 1, \cdots, H)\) with \( u_1, \cdots, u_H \in \mathbb{R}^p \) and \( w_1, \cdots, w_H \in \mathbb{R} \), define \( K_\delta^0 \) and \( K_\delta^C \) for \( \delta > 0 \) as

\[
K_\delta^0 = \{ x \in \mathbb{R}^p : u_j \cdot x + w_j \leq -\delta, \quad \forall j \in \{1, \cdots, H\} \},
\]

\[
K_\delta^C = \{ x \in \mathbb{R}^p : u_j \cdot x + w_j \geq \delta, \quad \text{for some } j \in \{1, \cdots, H\} \}.
\]

Denote by \( x^{(v)} \) the \( v \)th component of vector \( x \in \mathbb{R}^p \), and \( |x|_1 \) the \( L_1 \)-norm defined as \( |x|_1 = \sum_{v=1}^d |x^{(v)}| \).

The following lemma is adapted from Theorem 2 in [6].

**Lemma 1.** Let \( a \geq 1 \) and \( \lambda > 0 \) be two given constants. Assume that \( m : \mathbb{R}^p \rightarrow \mathbb{R} \) is a polynomial function defined as

\[
m(x) = \sum_{|\alpha| \leq q_0} r_\alpha x^\alpha \tag{A.1}
\]

with \( \max_{|\alpha| \leq q_0} |r_\alpha| = \bar{r}_m \), and \( \nu \) is an arbitrary probability measure on \( \mathbb{R}^p \). Let \( N \in \mathbb{N}_0 \) be chosen such that \( N \geq q_0 \) and \( \sigma : \mathbb{R} \rightarrow [0, 1] \) the sigmoid function. Then for any \( \eta \in (0, 1) \) and \( M \in \mathbb{N} \) such that \( M^\lambda \geq 2(N + |\nu\sigma|)(\frac{N+1}{\nu\sigma} + 1) \) in which \( \nu\sigma \in (0, 1) \) can be chosen such that \( \sigma^{(i)}(\nu\sigma) \neq 0 \) for all \( i \in \mathbb{N}_0 \), and \( M \geq a \), there exists a neural network of type

\[
t(x) = \sum_{i=1}^{(p+N)(N+1)(M+1)^p} \mu_i \sigma\left( \sum_{l=1}^{4d} \lambda_{i,l} \sigma\left( \sum_{v=1}^p \theta_{i,l,v} x^{(v)} + \theta_{i,l,0} \right) + \lambda_{i,0} \right) \tag{A.2}
\]

such that

\[
|t(x) - m(x)| \leq c_1 a^{N+q_0+3} M^{-\lambda} \tag{A.3}
\]

holds for all \( x \in [-a, a]^p \) up to a set of \( \nu \)-measure less than or equal to \( \eta \). The coefficients of \( t(x) \) can be bounded by

\[
|\mu_i| \leq c_1 a^{q_0} M^{N\lambda},
\]

\[
|\lambda_{i,l}| \leq M^{p+\lambda(N+2)},
\]

\[
|\theta_{i,l,v}| \leq \frac{p}{\eta} M^{p+\lambda(2N+3)+1}
\]

for all \( i \in \{1, \cdots, (p+N)/p\} \), \( l \in \{0, \cdots, 4d\} \), and \( v \in \{0, \cdots, p\} \), where the
positive constants $c_{13}$ and $c_{14}$ are free of $M$ and $\lambda$.

**Remark 1.** In the proof of Theorem 2 in [5], parameter $\lambda$ that controls both the bounds for the coefficients and affects the bound for the approximation error is chosen to be slightly greater than $q_0$ (to be exact $\lambda = q_0 + r$ for some $r \in (0,1)$). If we assume Condition 1(iii) instead of Condition 2(i) so that $m(x)$ does not take the polynomial form, for the Taylor expansion $p(x)$ of $m(x)$ at point $x_0$ to order $q_0$, it holds that

$$|t(x) - m(x)| \leq |t(x) - p(x)| + |p(x) - m(x)|.$$

The first term on the right-hand side above enjoys the same bound as in (A.3) because $p(x)$ is a polynomial function, while the second term can be bounded by $c\|x - x_0\|^{q_0}$ for some constant $c$ that depends only on $q_0$ and $p$, according to Lemma 8 in [6]. Within each cube $C_i$ that will be defined in (A.4), we have

$$c\|x - x_0\|^{q_0} \leq c p^{(q_0+1)/2} a^{q_0+1} M^{-q_0-1}.$$

Thus, setting $\lambda = q_0 + r$ makes the two bounds of roughly the same order and, meanwhile, minimizes the bounds for the coefficients, yielding the minimal complexity of the neural networks.

In contrast, by assuming Condition 2(i), the second term $|p(x) - m(x)|$ on the right-hand side above vanishes. Thus, we no longer require that $\lambda = q_0 + r$, and instead, $\lambda$ here can be some arbitrary positive number. In Lemmas 2 and 3 to be presented later, we will apply the result here by setting $\lambda = \lambda_n$ as specified in Condition 2(ii) to obtain the desired convergence rate.

**Proof.** The proof is adapted from that of Theorem 2 in [5]. It is presented here for the sake of completeness. Note that the existence of $t_\sigma$ is guaranteed by the discussion of $N$-admissible in the “Sigmoidal Squasher is $N$-admissible” section in [6]. Let $\{C_i : i = (i_1, i_2, \ldots, i_p) \in \{1, \ldots, M+1\}^p\}$ be a partition of the hypercube $C = [-a - \frac{2a}{M}, a]^p$, where $C_i$ is the subcube defined as

$$\left[-a + (i_1 - 2) \frac{2a}{M}, -a + (i_1 - 1) \frac{2a}{M}\right] \times \ldots$$

$$\times \left[-a + (i_p - 2) \frac{2a}{M}, -a + (i_p - 1) \frac{2a}{M}\right]. \quad (A.4)$$

Denote by $x_i$ the “bottom left” corner of cube $C_i$; that is, for $i = (i_1, \ldots, i_p)$,

$$x_i = \left(-a + (i_1 - 2) \frac{2a}{M}, \ldots, -a + (i_p - 2) \frac{2a}{M}\right).$$

We can extend the definition of $x_i$ to all $i \in \{1, 2, \ldots, M + 2\}^p$ with $C_i$ defined in (A.4).

For some $\lambda > 0$, we can apply Lemma 7 in [5] to function $m(x)$ and let $K$ defined therein be $C_i$. Then it follows that for $M$ large enough such that

$$\frac{Na}{(p+1)M^\lambda + |t_\sigma|} \frac{2^N M^{\lambda(N+1)}}{\sigma^N(t_\sigma)} + 1 \leq M^{p+\lambda(N+2)} \left(\frac{3}{4} - M^{-p-\lambda(2N+3)}\right)$$
and $M \geq a$, neural networks $t(x)$ of type

$$t(x) = \sum_{j=1}^{(p+N)\binom{N+1}{M+1}} \mu_i \sigma \left( \sum_{l=1}^{4d} \lambda_{i,l} \sigma \left( \sum_{v=1}^{p} \theta_{i,l,v} x^{(v)}(x) + \theta_{i,l,0} \right) + \lambda_{i,0} \right)$$  \hspace{1cm} (A.5) $$

exist with coefficients bounded as

$$|\mu_i| \leq c_{14}a^{q_0}M^{Np},$$
$$|\lambda_{i,l}| \leq M^{p+\lambda(N+2)},$$
$$|\theta_{i,l,v}| \leq 6^{p}M^{p+\lambda(2N+3)+1}$$

for all $i \in \{1,\ldots,(p+N)\binom{N+1}{M+1}\}$, $l \in \{0,\ldots,4p\}$, and $v \in \{0,\ldots,p\}$ such that

$$|t(x) - m(x)| \leq c_{22}\bar{r}(m)a^{N+3}M^{-\lambda} \hspace{1cm} \text{for} \hspace{0.5cm} x \in (C_i)^0 \cap [-a,a]^p,$$
$$|t(x)| \leq c_{23}\bar{r}(m)M^{-p-2\lambda} \hspace{1cm} \text{for} \hspace{0.5cm} x \in (C_i)^0 \cap [-a,a]^p,$$
$$|t(x)| \leq c_{24}\bar{r}(m)M^{N-\lambda} \hspace{1cm} \text{for} \hspace{0.5cm} x \in \mathbb{R}^p.$$  

Here, $\bar{r}(m)$ is some constant depending on $q_0$, the order of $m(x)$, and $(C_i)^0$ and $(C_i)^0$ are defined analogously to $K_\delta^0$ and $K_\delta^0$, respectively. The constants $c_{22}$, $c_{23}$, and $c_{24}$ depend only on $p$ and $N$. Since the polynomial functional form of $m(x)$ stays the same across different cubes, the result above holds for all cubes with all the constants remaining unchanged. That is, the above results hold for $K = C_i$ for any $i \in \{1,\ldots,M+1\}^p$.

By Lemma 3 in [5], $\bar{r}(m)$ in the representation above can be upper bounded as

$$\bar{r}(m) \leq c_{27}a^{q_0},$$

where constant $c_{27}$ here can be chosen as $c_{27}$ in [5] multiplied by $q_0!$ and it depends only on $q_0$. Recall that $(C_i)^0$ is defined similar to $K_\delta^0$. Then it holds that for $x \in (C_i)^0 \cap [-a,a]^p$,

$$|t(x) - m(x)| \leq c_{22}\bar{r}(m)a^{N+3}M^{-\lambda} = c_{13}a^{N+q_0+3}M^{-\lambda}.$$  \hspace{1cm} (A.6) $$

Since $m(x)$ takes the same functional form across different cubes, this bound holds for all $i \in \{1,\ldots,M+1\}^p$. That is, (A.6) holds for all $x$ in $[-a,a]^p$ except for set

$$\bigcup_{j=1}^{p} \bigcup_{i \in \{1,\ldots,M+2\}^p} \{ x \in \mathbb{R}^p : |x^{(j)} - x^{(j)}_i| \leq \delta \}  \hspace{1cm} (A.7)$$

because of the definition of $(C_i)^0$.

By slightly shifting the whole grid cubes along the $j$th component with the same value that is less than $\frac{2\delta}{M}$ for a fixed $j \in \{1,\ldots,p\}$, we can construct different versions of $t(x)$ that
still satisfy (A.6) for all \( x \in [-a, a]^p \) except for those \( x \) belonging to
\[
\bigcup_{i \in \{1, \ldots, M+2\}^p} \{ x \in \mathbb{R}^p : |x^{(j)} - x_i^{(j)}| \leq \delta \}.
\]  
(A.8)

Here, all the components of \( x_i \) increase by an amount less than \( \frac{2a}{M} \), and we have at least \( p/\eta \) choices to make the above different versions of sets in (A.8) pairwisely disjoint because
\[
\left[ \frac{2a}{M} \right] = \left[ \frac{2a}{2pM} \right] = \left[ \frac{p}{\eta} \right] \geq \frac{p}{\eta}.
\]

Since the sum of the \( \nu \)-measures of these sets is less than or equal to one, at least one of them must have measure less than or equal to \( \eta/p \). Thus we can shift the \( j \)th component of \( x_i \) accordingly so that the \( \nu \)-measure of (A.7) is less than \( \eta \) by the union bound. This completes the proof of Lemma 1.

The following lemma is adapted from Theorem 3 in [5].

**Lemma 2.** Let \( X \) be a \( \mathbb{R}^p \)-valued random variable and \( m : \mathbb{R}^p \to \mathbb{R} \) satisfy a generalized hierarchical interaction model of order \( p^* \) and finite level \( l \). For a nonnegative integer \( q_0 \), let \( N \in \mathbb{N}_0 \) with \( N \geq q_0 \). Assume that in Definition 3, all the functions \( g_b, f_{j,k} \) are polynomial functions up to order \( q_0 \) and all functions \( g_b \) are Lipschitz continuous with Lipschitz constant \( L > 0 \). Let the activation function be chosen as the sigmoid function and \( t_x \) as defined in Lemma 1. Let \( \lambda_n \in \mathbb{R}_+ \), \( M_n \in \mathbb{N} \) be such that \( M_n^{\lambda_n} \geq 2(N + |t_x|)(\frac{2N+1}{\sigma(\eta T)}+1) \) for \( n \) large enough, and let \( a_n \in [1, M_n] \) be an increasing sequence with condition \( a_n^{N+q_0+3} \leq M_n^{\lambda_n} \) satisfied for \( n \) sufficiently large. Assume that \( \eta_n \in (0, 1) \) and parameters in \( \mathcal{H}_{M^*,p^*,p-1,\alpha} \)
are defined as \( M^* = \left( \frac{p^*+N}{p^*} \right)(N+1)(M_n+1)^{p^*} \) and \( \alpha = \log(n)\frac{M_n^{p^*+\lambda_n(2N+3)+1}}{\eta_n} \). Then for arbitrary \( c > 0 \) and all \( n \) greater than a certain \( n_0(c) \in \mathbb{N} \), there exists a neural network \( t \in \mathcal{H}_{M^*,p^*,p-1,\alpha} \) such that outside of a set of \( \mathbb{P} \)-measure less than or equal to \( c \eta_n \), we have
\[
|t(x) - m(x)| \leq c_{29}a_n^{N+q_0+3}M_n^{\lambda_n}
\]
for all \( x \in [-a_n, a_n]^p \). Here, constant \( c_{29} \) depends on \( c, p, p^*, q_0, \) and \( N \), but not on \( n \). Moreover, \( t(x) \) can be chosen such that
\[
|t(x)| \leq c_{30}a_n^{q_0}M_n^{p^*+N\lambda_n}
\]
holds for all \( x \in \mathbb{R}^p \).

**Proof.** The proof is a simple modification of that of Theorem 3 in [5]. For completeness, we still present it here. The main idea is proof by induction. We only consider the case when \( c \eta_n < 1 \) because if \( c \eta_n \geq 1 \), then the assertion is automatically true.

For a function \( m(x) = f(b_1^T x, \ldots, b_p^T x) = f(h(x)) \) in which \( f : \mathbb{R}^{p^*} \to \mathbb{R} \) is a polynomial function up to \( q_0 \) order and \( h : \mathbb{R}^p \to \mathbb{R}^{p^*} \) is the mapping \( h(x) = (b_1^T x, \ldots, b_p^T x)^T \), one can apply Lemma 1 to \( f(y) \) to obtain a neural network approximation \( \tilde{f}(y) \) for \( y \in [-\max_{k=1,\ldots,p^*} |b_k|a_n, \max_{k=1,\ldots,p^*} |b_k|a_n]^{p^*} \) except for a set \( \tilde{D}_0 \) of \( \mathbb{P}(h(X)) \)-measure less than
or equal to $c_\eta n$ with an error of

$$|\hat{f}(y) - f(y)| \leq c_{13}(\max_{k=1,\ldots,p^*} |b_k|_1 a_n)^{N+q_0+3} M_n^{-\lambda_n}.$$  

The corresponding neural network approximation $t(x)$ of $m(x)$ can be obtained using the relationship of $t(x) = \hat{f}(y) = \hat{f}(h(x))$ due to the fact that $y = h(x)$ is a linear transformation with $\max_{k=1,\ldots,p^*} |b_k|_1$ contributing to the bounds of parameters $\mu_i$ and $\theta_{i,l,v}$. That is, to write $t(x)$ in the form of (A.5), we have

$$|\mu_i| \leq c_{14}(\max_{k=1,\ldots,p^*} |b_k|_1 a_n)^{q_0} M_n^{N\lambda_n} \leq \alpha,$$

$$|\lambda_{i,l}| \leq M^{p^*+\lambda_n(N+2)} \leq \alpha,$$

$$|\theta_{i,l,v}| \leq 6 \max_{k=1,\ldots,p^*} |b_k|_1 \frac{P^*}{\eta_n} M_n^{p^*+\lambda_n(2N+3)+1} \leq \alpha.$$

Then the $\mathbb{P}_x$-measure of the exception set $D_0 := \{x \in \mathbb{R}^p | h(x) \in \tilde{D}_0\}$ is also bounded by $c_\eta n$. Outside of $D_0$, it holds that

$$|t(x) - m(x)| \leq c_{13}(\max_{k=1,\ldots,p^*} |b_k|_1 a_n)^{N+q_0+3} M_n^{-\lambda_n}.$$

On the other hand, we can show that

$$|t(x)| \leq M^* \max_{i=1,\ldots,M^*} |\mu_i| \leq c_{31} a_n^{q_0} M_n^{p^*+N\lambda_n}$$

for all $x \in \mathbb{R}^p$. Thus, the conclusion is true for the case of $l = 0$.

When $l > 0$, let $m(x) = \sum_{k=1}^K g_k(f_{1,k}(x), \ldots, f_{p^*,k}(x)) = \sum_{k=1}^K g_k(h_k(x))$ with $h_k(x)$ the linear mapping defined analogously to $h(x)$, and the neural network approximation be $\hat{\hat{m}}(x) = \sum_{k=1}^K \hat{g}_k(\tilde{f}_{1,k}(x), \ldots, \hat{f}_{p^*,k}(x)) = \sum_{k=1}^K \hat{g}_k(\tilde{h}_k(x))$, where $\tilde{f}_{j,k} \in \mathcal{H}(l, \ldots, l, \ldots, l, 0)$ can be found according to the induction hypothesis with $\eta_n$ replaced by $\frac{\eta_n}{2^l K}$, since $f_{j,k}(x)$ are assumed to be polynomials up to order $q_0$ of $x$. Then each of the terms $|\tilde{f}_{j,k}(x) - \hat{f}_{j,k}(x)|$ can be bounded by $c_{32} a_n^{N+q_0+3} M_n^{-\lambda_n}$ for all $n$ sufficiently large and all $x \in [-a_n, a_n]^p$ outside of a set $D_{j,k}$ of $\mathbb{P}_x$-measure less than or equal to $\frac{c_{32}}{2^l K}$. Further, $\hat{g}_k$ can be chosen from Lemma 1 with $\eta = \frac{c_{32}}{2^l K}$ such that

$$|\hat{g}_k(y) - g_k(y)| \leq c_{33}(\max_{j=1,\ldots,p^*} \|f_{j,k}\|_{\infty} + c_{32})^{N+q_0+3} M_n^{-\lambda_n} \leq c_{33} M_n^{-\lambda_n}$$

holds for all $y \in [-\max_{j=1,\ldots,p^*} \|f_{j,k}\|_{\infty} - c_{32}, \max_{j=1,\ldots,p^*} \|f_{j,k}\|_{\infty} + c_{32}]^p$ except a set $\tilde{D}_k$ that satisfies $\mathbb{P}_{h_k(x)}(\tilde{D}_k) \leq \frac{\eta_n}{2^l K}$ ($c_{32}$ can be modified so that $\max_{j=1,\ldots,p^*} \|f_{j,k}\|_{\infty} + c_{32} \geq 1$ is
satisfied. Indeed, \( \hat{g}_k \) can be represented in the form of (A.5) with parameters satisfying

\[
|\mu_s| \leq c_{14} \left( \max_{j=1, \ldots, p^*} \|f_{j,k}\|_\infty + c_{32} \right)^{\gamma_0} M_n^{N\lambda_n} \leq \alpha, \\
|\lambda_{i,t}| \leq M^{p^* + \lambda_n(N+2)} \leq \alpha, \\
|\theta_{i,t,o}| \leq \frac{p}{\eta_n} M_n^{p^* + \lambda_n(2N+3)+1} \leq \alpha,
\]

which implies that \( \hat{g}_k \in \mathcal{H}^{(0)}_{M^*, p^*, p^* - 1, \alpha} \).

Let us define \( \hat{h}_k^{-1}(\tilde{D}_k) := \{x \in \mathbb{R}^p | \hat{h}_k(x) \in \tilde{D}_k \} \). Since \( \mathbb{P}^{\hat{h}_k(x)}(\hat{D}_k) = \mathbb{P}_x(\hat{h}_k^{-1}(\tilde{D}_k)) \), \( \hat{g}_k(\hat{h}_k(x)) \) approximates \( g_k(\hat{h}_k(x)) \) with the maximum approximation error given above for all

\[
x \in [-a_n, a_n]^p \setminus \bigcup_{j=1, \ldots, p^*} D_{j,k}
\]

outside of the set \( D_k := \hat{h}_k^{-1}(\tilde{D}_k) \) of \( \mathbb{P}_x \)-measure less than or equal to \( \frac{c_{19}}{2n} \). Denote by \( t(x) = \hat{m}(x) \). Then from the derivations above, we have that \( t(x) \in \mathcal{H}^{(i)}_{M^*, p^*, p^* - 1, \alpha} \) and

\[
|t(x) - m(x)| \leq \sum_{k=1}^K g_k(h_k(x)) - \sum_{k=1}^K g_k(\hat{h}_k(x)) + \left| \sum_{k=1}^K g_k(\hat{h}_k(x)) - \sum_{k=1}^K \hat{g}_k(\hat{h}_k(x)) \right| \\
\leq \sum_{k=1}^K L \sum_{j=1}^{p^*} |f_{j,k}(x) - \hat{f}_{j,k}(x)| + \left| \sum_{k=1}^K g_k(\hat{h}_k(x)) - \sum_{k=1}^K \hat{g}_k(\hat{h}_k(x)) \right| \\
\leq K L p^* c_{32} a_n^{N+q_0+3} M_n^{-\lambda_n} + K c_{33} M_n^{-\lambda_n} \leq c_{29} a_n^{N+q_0+3} M_n^{-\lambda_n}
\]

holds for all \( x \in [-a_n, a_n]^p \) outside of the set

\[
\bigcup_{j=1, \ldots, p^*} D_{j,k} \cup \bigcup_{k=1, \ldots, K} D_k.
\]

Meanwhile, the \( \mathbb{P}_x \)-measure of the set is bounded by \( p^* K \frac{c_{19}}{2p} + K \frac{c_{19}}{2n} = c_{19}n \) as desired.

On the other hand, for all \( x \in \mathbb{R}^p \), we can deduce that

\[
|t(x)| \leq K \left( p^* + N \right) (N + 1)(M_n + 1)^{p^*} \max_{k=1, \ldots, K} \left( \max_{j=1, \ldots, p^*} \|f_{j,k}\|_\infty + c_{32} \right)^{\gamma_0} M_n^{N\lambda_n} \\
\leq c_{34} M^{p^* + N\lambda_n},
\]

which concludes the proof of Lemma 2.

The following lemma is adapted from Theorem 1 in [5].

**Lemma 3.** Let \( \{(X_i, Y_i)\}_{i=1}^n \) be an i.i.d. sample collected from an underlying distribution such that \( \text{supp}(X) \) is bounded and \( \mathbb{E}\exp(c_1 Y^2) \leq \infty \) for some constant \( c_1 > 0 \). Assume that Condition 2 with the sigmoid function \( \sigma : \mathbb{R} \to (0, 1) \) is satisfied. Let \( t_\sigma \) be defined as in Lemma 1 and \( q_0 \) the highest order of all the polynomials appearing in Condition 2(i) with arbitrary constant \( N \in \mathbb{N}_0 \) such that \( N \geq q_0 \). Denote by \( m_n \) the least-squares estimate
defined in (8). Then it holds that
\[ \mathbb{E} \int |m_n(x) - m(x)|^2 \mathbb{P}_X(dx) \leq c_{60} \log^{p^*+3}(n)n^{-1} \]
for all sufficiently large \( n \), where constant \( c_{60} \) depends on \( N, q_0, t_\sigma, p \), and \( p^* \), but not on \( n \).

**Proof.** Let \( a_n = \log \frac{2^{N+3} + q_0 + 3}{2^{N+3} + q_0} (n) \). For a sufficiently large \( n \), it holds that \( \text{supp}(X) \subseteq [-a_n, a_n]^p \), which entails that \( \mathcal{N}(\delta, \mathcal{G}, \| \cdot \|_{\infty, \text{supp}(X)}) \leq \mathcal{N}(\delta, \mathcal{G}, \| \cdot \|_{\infty, [-a_n, a_n]^p}) \) for an arbitrary function space \( \mathcal{G} \) and \( \delta > 0 \). Then an application of Lemmas 1 and 2 in [5] gives
\[
\mathbb{E} \int |m_n(x) - m(x)|^2 \mathbb{P}_X(dx) \leq c_{7} \log^2(n) \frac{\log(n) M^*}{n} + 2 \inf_{h \in \mathcal{H}_{M^*+p^*,p-1,\alpha}} \int |h(x) - m(x)|^2 \mathbb{P}_X(dx). \tag{A.9}
\]

We next bound the second term on the right-hand side above by using Lemma 2. Note that the condition that all the \( g_k \)'s are Lipschitz continuous with some Lipschitz constant \( L > 0 \) can be guaranteed by the fact that they are polynomials on a bounded support.

We first define an integer-valued function \( n(\lambda) \)
\[ n(\lambda) = \inf \{ n \in \mathbb{N} : M_n = \left\lfloor n^{\frac{3}{2\lambda + p^*}} \right\rfloor, M_n^\lambda \geq 2(N + |t_\sigma|)(\frac{2^{N+1}}{\sigma(N)(t_\sigma)} + 1), n^{\frac{3}{2\lambda + p^*}} \geq a_n \} \]
for all \( \lambda \geq q_0 + 1 \). Clearly, \( n(\lambda) \) is finite and increasing with \( \lambda \). Indeed, for \( \lambda \) sufficiently large, it follows that
\[ n(\lambda) = \inf \{ n : \frac{\log(n)}{2\lambda + p^*} \geq \log(\frac{3}{2(N + q_0 + 3)}) + \log(\log(n)) \}. \]

Starting from \( n = n(q_0 + 1) \), let us define \( \lambda_n = \inf \{ \lambda \in \mathbb{N} : n(\lambda) \geq n + 1 \} \). Then we have \( n(\lambda_n) \geq n + 1 \). Since \( n(\lambda - 1) \) does not satisfy \( \frac{\log(n)}{2\lambda + p^*} \geq \log(\frac{3}{2(N + q_0 + 3)}) + \log(\log(n)) \), it holds that
\[ \frac{1}{2\lambda_n + p^*} \leq \frac{\log(\frac{3}{2(N + q_0 + 3)}) + \log(\log(n(\lambda_n) - 1))}{\log(n(\lambda_n) - 1)} \]
\[ \leq \frac{\log(\frac{3}{2(N + q_0 + 3)}) + \log(\log(n))}{\log(n)}, \]
where the second inequality follows from the monotonicity of function \( \frac{\log(\frac{3}{2(N + q_0 + 3)}) + \log(\log(n))}{\log(n)} \) with respect to \( n \) when \( n \) is large enough.

We set \( M_n = \left\lfloor n^{\frac{3}{2\lambda_n + p^*}} \right\rfloor \) and \( \eta_n = \log \frac{3(N + 3)}{N + q_0 + 3}(n)n^{-\frac{2\lambda_n(N + 1) + 2p^*}{2\lambda_n + p^*}} \). Denote by \( \alpha_0 = \log(n) \frac{M_n^{p^*} + \lambda_n(2N+3)+1}{\eta_n} \).
Then it is seen that
\[ \alpha_0 = \log_\alpha (\frac{2^{2N+30+6}}{\epsilon n^{4N+6}}) \leq \frac{\lambda_n(4N+5)+3p^*}{2\lambda_n+p^*} \leq n^{4N+6}. \]

Choosing constant \( c_2 \) in Condition 2(ii) to be larger than \( 4N + 6 \), we can obtain that \( H_{M^*,p^*,p-1,\alpha_0} \subset H_{M^*,p^*,p-1,\alpha} \) with \( \alpha = n^{c_2} \) since \( \alpha \geq \alpha_0 \). Consequently, it follows that
\[ \inf_{h \in H_{M^*,p^*,p-1,\alpha}} \int |h(x) - m(x)|^2 p_X(dx) \leq \inf_{h \in H_{M^*,p^*,p-1,\alpha_0}} \int |h(x) - m(x)|^2 p_X(dx). \quad (A.10) \]

Denote by \( t(x) \in H_{M^*,p^*,p-1,\alpha_0} \) the neural network characterized in Lemma 2 with \( \alpha \) therein set to be \( \alpha_0 \) defined above, and let \( D_n \) be the exception set in Lemma 2, outside of which \( |t(x) - m(x)| \leq c_2 \alpha_n^{N+90+3} M_n^{\lambda_n} \) holds with \( P_X(D_n) \leq c \eta_n \) for \( c = 1 \). Then we can deduce that
\[ \inf_{h \in H_{M^*,p^*,p-1,\alpha_0}} \int |h(x) - m(x)|^2 p_X(dx) \]
\[ \leq \int |t(x) - m(x)|^2 p_X(dx) \]
\[ = \int |t(x) - m(x)|^2 1_{D_n^c} p_X(dx) + \int |t(x) - m(x)|^2 1_{D_n} p_X(dx) \]
\[ \leq (c_2 \alpha_n^{N+90+3} M_n^{\lambda_n})^2 + (2c_3 \alpha_n^{90} M_n^{p^*+N\lambda_n})^2 \eta_n \]
\[ \leq c_{11} \log^3(n) n^{-\frac{2 \lambda_n}{2\lambda_n+p^*}}. \quad (A.11) \]

Therefore, in view of (A.9), (A.10), and (A.11), it holds for \( n \) sufficiently large that
\[ \mathbb{E} \int |m_n(x) - m(x)|^2 p_X(dx) \]
\[ \leq c_4 \log^3(n) n^\frac{p^*}{2\lambda_n+p^*} n^{-1} \]
\[ \leq c_4 \log^3(n) n^{p^* \frac{\log(\frac{3}{\epsilon} \log(\log(n)))}{\log(n)}} n^{-1} \]
\[ \leq c_{60} \log^3+p^*(n)n^{-1}, \]
which completes the proof of Lemma 3.

**B.3 Lemma 4 and its proof**

The following lemma gives parallel results to Proposition 1.

**Lemma 4.** Assume that (i) and (iv) of Condition 1 and Condition 2 hold with the sigmoid activation function \( \sigma(x) = \frac{e^x}{1+e^x} \) in \( H^{(l)} \). Then the estimator \( \hat{\tau}_D \) defined in (9) satisfies that
\[ |\hat{\tau}_D - \tau| = o_P(\log^{1+p^*}(n)n^{-1/2}) \]
as \( n_D := n \to \infty \).

**Proof.** The proof follows from similar arguments as in the proof of Proposition 1 using the newly established Lemmas 1–3 in Section B.2 with the caution that dimensionality \( p \)
in Lemmas 1–3 needs to be updated to $p + 1$ for proving Lemma 4 here. The details are omitted for simplicity.

C Additional numerical results

In this section, we present additional simulation and real data results corresponding to different numbers of training epochs. In particular, Figures 4–6 and Tables 4–6 summarize simulation results parallel to those in Section 4.1 with the number of epochs ranging from 100 to 400, Figures 7–9 and Tables 7–9 summarize simulation results parallel to those in Section 4.2 with the number of epochs ranging from 100 to 400, and Figures 10–11 and Tables 10–11 summarize real data results parallel to those in Section 5 with the number of epochs ranging from 200 to 400.

Figure 4: The scaled density of the ATE estimate over 200 replications for different training sample sizes and different activation functions. Here we use a fixed inference sample size of $n = 1000$ and train each network for 100 epochs. The true treatment effect of $\tau = 1$ is shown as a red vertical line.
| $n_1$ | Activation | Mean   | Median | SD      | MSE      |
|-------|------------|--------|--------|---------|----------|
| 1000  | ReLU       | 0.9808 | 0.9791 | 0.09563 | 0.00947  |
|       | Sigmoid    | 0.9239 | 0.9236 | 0.09009 | 0.01386  |
| 2000  | ReLU       | 0.9807 | 0.9815 | 0.07616 | 0.00614  |
|       | Sigmoid    | 0.9479 | 0.9503 | 0.04731 | 0.00494  |
| 3000  | ReLU       | 0.9867 | 0.9864 | 0.06919 | 0.00494  |
|       | Sigmoid    | 0.9612 | 0.9597 | 0.04027 | 0.00312  |
| 4000  | ReLU       | 0.9754 | 0.9756 | 0.05937 | 0.00411  |
|       | Sigmoid    | 0.9627 | 0.9601 | 0.03353 | 0.00251  |
| 5000  | ReLU       | 0.9883 | 0.9911 | 0.06214 | 0.00398  |
|       | Sigmoid    | 0.9636 | 0.9614 | 0.03735 | 0.00271  |

Table 4: Results of the first simulation setting in Section 4.1 aggregated over 200 replications. In each replication, the networks are trained for 100 epochs.

Figure 5: The scaled density of the ATE estimate over 200 replications for different training sample sizes and different activation functions. Here we use a fixed inference sample size of $n = 1000$ and train each network for 200 epochs. The true treatment effect of $\tau = 1$ is shown as a red vertical line.
### Table 5: Results of the first simulation setting in Section 4.1 aggregated over 200 replications.
In each replication, the networks are trained for 200 epochs.

| $n_1$ | Activation | Mean     | Median   | SD        | MSE       |
|-------|------------|----------|----------|-----------|-----------|
| 1000  | ReLU       | 0.9847   | 0.9844   | 0.11223   | 0.01277   |
|       | Sigmoid    | 0.9555   | 0.9566   | 0.06773   | 0.00654   |
| 2000  | ReLU       | 0.9823   | 0.9872   | 0.07756   | 0.00630   |
|       | Sigmoid    | 0.9672   | 0.9675   | 0.04974   | 0.00354   |
| 3000  | ReLU       | 0.9877   | 0.9851   | 0.07042   | 0.00508   |
|       | Sigmoid    | 0.9771   | 0.9760   | 0.03771   | 0.00194   |
| 4000  | ReLU       | 0.9837   | 0.9764   | 0.06929   | 0.00504   |
|       | Sigmoid    | 0.9779   | 0.9779   | 0.03706   | 0.00186   |
| 5000  | ReLU       | 0.9806   | 0.9850   | 0.06339   | 0.00437   |
|       | Sigmoid    | 0.9732   | 0.9725   | 0.03404   | 0.00187   |

### Table 6: Results of the first simulation setting in Section 4.1 aggregated over 200 replications.
In each replication, the networks are trained for 400 epochs.

| $n_1$ | Activation | Mean     | Median   | SD        | MSE       |
|-------|------------|----------|----------|-----------|-----------|
| 1000  | ReLU       | 0.9670   | 0.9656   | 0.10374   | 0.01180   |
|       | Sigmoid    | 0.9764   | 0.9726   | 0.06355   | 0.00457   |
| 2000  | ReLU       | 0.9692   | 0.9711   | 0.07743   | 0.00692   |
|       | Sigmoid    | 0.9917   | 0.9913   | 0.05239   | 0.00280   |
| 3000  | ReLU       | 0.9711   | 0.9687   | 0.07216   | 0.00602   |
|       | Sigmoid    | 0.9958   | 0.9974   | 0.04351   | 0.00190   |
| 4000  | ReLU       | 0.9903   | 0.9834   | 0.06273   | 0.00401   |
|       | Sigmoid    | 0.9930   | 0.9932   | 0.03933   | 0.00159   |
| 5000  | ReLU       | 0.9741   | 0.9773   | 0.06648   | 0.00507   |
|       | Sigmoid    | 0.9971   | 0.9986   | 0.03195   | 0.00102   |
| $n_1$ | Estimate Type                     | Activation | Mean     | Median   | SD       | MSE      |
|-------|-----------------------------------|------------|----------|----------|----------|----------|
|       | Difference of Means Estimate      | ReLU       | 0.9763   | 0.9767   | 0.09150  | 0.00889  |
| 1000  | Doubly Robust Estimate            | ReLU       | 0.9703   | 0.9659   | 0.16154  | 0.02684  |
|       | Sigmoid                           | ReLU       | 0.9808   | 0.9788   | 0.08293  | 0.00721  |
|       | Sigmoid                           | ReLU       | 0.9703   | 0.9659   | 0.16154  | 0.02684  |
|       | Doubly Robust Estimate            | Sigmoid    | 0.9808   | 0.9788   | 0.08293  | 0.00721  |
| 2000  | Difference of Means Estimate      | ReLU       | 0.9707   | 0.9692   | 0.07016  | 0.00575  |
|       | Doubly Robust Estimate            | ReLU       | 0.9891   | 0.9851   | 0.17272  | 0.02980  |
|       | Sigmoid                           | ReLU       | 0.9891   | 0.9851   | 0.17272  | 0.02980  |
|       | Sigmoid                           | ReLU       | 0.9742   | 0.9669   | 0.07878  | 0.00684  |
| 3000  | Difference of Means Estimate      | ReLU       | 0.9745   | 0.9852   | 0.06145  | 0.00400  |
|       | Doubly Robust Estimate            | Sigmoid    | 0.9626   | 0.9615   | 0.03774  | 0.00282  |
|       | Sigmoid                           | ReLU       | 0.9880   | 0.9795   | 0.14298  | 0.02049  |
|       | Doubly Robust Estimate            | Sigmoid    | 0.9678   | 0.9657   | 0.07579  | 0.00675  |
| 4000  | Difference of Means Estimate      | ReLU       | 0.9813   | 0.9812   | 0.06223  | 0.00420  |
|       | Doubly Robust Estimate            | Sigmoid    | 0.9597   | 0.9590   | 0.03335  | 0.00273  |
|       | Sigmoid                           | ReLU       | 0.9884   | 0.9975   | 0.13479  | 0.01821  |
|       | Doubly Robust Estimate            | Sigmoid    | 0.9695   | 0.9662   | 0.07448  | 0.00645  |
| 5000  | Difference of Means Estimate      | ReLU       | 0.9892   | 0.9930   | 0.05888  | 0.00357  |
|       | Doubly Robust Estimate            | Sigmoid    | 0.9639   | 0.9640   | 0.02923  | 0.00215  |
|       | Sigmoid                           | ReLU       | 0.9941   | 0.9972   | 0.11426  | 0.01302  |
|       | Doubly Robust Estimate            | Sigmoid    | 0.9759   | 0.9810   | 0.06615  | 0.00494  |

Table 7: The simulation results corresponding to Figure 7 for 100 training epochs.
| \( n_1 \) | Estimate Type | Activation | Mean  | Median | SD    | MSE   |
|--------|---------------|------------|-------|--------|-------|-------|
|        | Difference of Means Estimate | ReLU | 0.9735 | 0.9764 | 0.09660 | 0.00999 |
|        |               | Sigmoid   | 0.9553 | 0.9565 | 0.07190 | 0.00714 |
| 1000   | Doubly Robust Estimate | ReLU | 0.9602 | 0.9495 | 0.17878 | 0.03339 |
|        |               | Sigmoid   | 0.9654 | 0.9649 | 0.08281 | 0.00802 |
|        | Difference of Means Estimate | ReLU | 0.9767 | 0.9787 | 0.07471 | 0.00610 |
|        |               | Sigmoid   | 0.9626 | 0.9655 | 0.04696 | 0.00359 |
| 2000   | Doubly Robust Estimate | ReLU | 1.0010 | 0.9830 | 0.17861 | 0.03174 |
|        |               | Sigmoid   | 0.9718 | 0.9620 | 0.08075 | 0.00728 |
|        | Difference of Means Estimate | ReLU | 0.9862 | 0.9814 | 0.07825 | 0.00628 |
|        |               | Sigmoid   | 0.9749 | 0.9722 | 0.03776 | 0.00205 |
| 3000   | Doubly Robust Estimate | ReLU | 0.9711 | 0.9655 | 0.13425 | 0.01876 |
|        |               | Sigmoid   | 0.9688 | 0.9657 | 0.07743 | 0.00694 |
|        | Difference of Means Estimate | ReLU | 0.9766 | 0.9699 | 0.06312 | 0.00451 |
|        |               | Sigmoid   | 0.9695 | 0.9694 | 0.03313 | 0.00202 |
| 4000   | Doubly Robust Estimate | ReLU | 0.9826 | 0.9743 | 0.13235 | 0.01797 |
|        |               | Sigmoid   | 0.9699 | 0.9596 | 0.07169 | 0.00602 |
|        | Difference of Means Estimate | ReLU | 0.9854 | 0.9844 | 0.06355 | 0.00423 |
|        |               | Sigmoid   | 0.9729 | 0.9737 | 0.03164 | 0.00173 |
| 5000   | Doubly Robust Estimate | ReLU | 0.9893 | 0.9891 | 0.12253 | 0.01505 |
|        |               | Sigmoid   | 0.9752 | 0.9776 | 0.06742 | 0.00514 |

Table 8: The simulation results corresponding to Figure 8 for 200 training epochs.
| $n_1$ | Estimate Type   | Activation | Mean    | Median  | SD      | MSE     |
|-------|-----------------|------------|---------|---------|---------|---------|
| 1000  | Difference of Means Estimate | ReLU  | 0.9764  | 0.9798  | 0.09930 | 0.01037 |
|       |                  | Sigmoid    | 0.9833  | 0.9834  | 0.07282 | 0.00556 |
|       | Doubly Robust Estimate | ReLU  | 0.9809  | 0.9732  | 0.18509 | 0.03445 |
|       |                  | Sigmoid    | 0.9626  | 0.9633  | 0.08055 | 0.00785 |
| 2000  | Difference of Means Estimate | ReLU  | 0.9666  | 0.9646  | 0.08197 | 0.00780 |
|       |                  | Sigmoid    | 0.9869  | 0.9904  | 0.04714 | 0.00238 |
|       | Doubly Robust Estimate | ReLU  | 0.9948  | 0.9835  | 0.18331 | 0.03346 |
|       |                  | Sigmoid    | 0.9678  | 0.9692  | 0.08182 | 0.00770 |
| 3000  | Difference of Means Estimate | ReLU  | 0.9770  | 0.9660  | 0.06987 | 0.00538 |
|       |                  | Sigmoid    | 0.9959  | 0.9971  | 0.04165 | 0.00174 |
|       | Doubly Robust Estimate | ReLU  | 0.9689  | 0.9614  | 0.17931 | 0.03296 |
|       |                  | Sigmoid    | 0.9660  | 0.9647  | 0.08015 | 0.00755 |
| 4000  | Difference of Means Estimate | ReLU  | 0.9802  | 0.9814  | 0.06417 | 0.00449 |
|       |                  | Sigmoid    | 0.9888  | 0.9895  | 0.03778 | 0.00155 |
|       | Doubly Robust Estimate | ReLU  | 0.9723  | 0.9590  | 0.15317 | 0.02411 |
|       |                  | Sigmoid    | 0.9696  | 0.9696  | 0.07448 | 0.00645 |
| 5000  | Difference of Means Estimate | ReLU  | 0.9864  | 0.9919  | 0.06906 | 0.00493 |
|       |                  | Sigmoid    | 0.9931  | 0.9944  | 0.03044 | 0.00097 |
|       | Doubly Robust Estimate | ReLU  | 0.9928  | 0.9943  | 0.12544 | 0.01571 |
|       |                  | Sigmoid    | 0.9780  | 0.9852  | 0.06821 | 0.00511 |

Table 9: The simulation results corresponding to Figure 9 for 400 training epochs.
| Inference Proportion | Estimate Type                  | Activation | Median | Robust SD |
|----------------------|--------------------------------|------------|--------|-----------|
| 0.2                  | Difference of Means Estimate   | ReLU       | 7328   | 2008      |
|                      |                               | Sigmoid    | 6300   | 2261      |
|                      | Doubly Robust Estimate        | ReLU       | 8683   | 3528      |
|                      |                               | Sigmoid    | 8114   | 3052      |
| 0.3                  | Difference of Means Estimate   | ReLU       | 7624   | 2154      |
|                      |                               | Sigmoid    | 5960   | 2152      |
|                      | Doubly Robust Estimate        | ReLU       | 8159   | 2349      |
|                      |                               | Sigmoid    | 8281   | 2084      |
| 0.4                  | Difference of Means Estimate   | ReLU       | 7546   | 2428      |
|                      |                               | Sigmoid    | 6526   | 2443      |
|                      | Doubly Robust Estimate        | ReLU       | 8220   | 2301      |
|                      |                               | Sigmoid    | 8013   | 1689      |
| 0.5                  | Difference of Means Estimate   | ReLU       | 7472   | 1831      |
|                      |                               | Sigmoid    | 5819   | 2208      |
|                      | Doubly Robust Estimate        | ReLU       | 8292   | 1960      |
|                      |                               | Sigmoid    | 8184   | 1462      |

Table 10: The real data results corresponding to Figure 10 for 200 training epochs.
| Inference Proportion | Estimate Type          | Activation | Median | Robust SD |
|----------------------|------------------------|------------|--------|-----------|
|                      | Difference of Means Estimate | ReLU       | 7711   | 1664      |
|                      |                        | Sigmoid    | 6462   | 2212      |
|                      | Doubly Robust Estimate | ReLU       | 8200   | 3410      |
|                      |                        | Sigmoid    | 7495   | 3057      |
| 0.2                  | Difference of Means Estimate | ReLU       | 7761   | 2252      |
|                      |                        | Sigmoid    | 6650   | 2115      |
|                      | Doubly Robust Estimate | ReLU       | 7987   | 2547      |
|                      |                        | Sigmoid    | 8118   | 2252      |
| 0.3                  | Difference of Means Estimate | ReLU       | 8064   | 2518      |
|                      |                        | Sigmoid    | 6722   | 1942      |
|                      | Doubly Robust Estimate | ReLU       | 7970   | 2257      |
|                      |                        | Sigmoid    | 7840   | 1967      |
| 0.4                  | Difference of Means Estimate | ReLU       | 7676   | 2400      |
|                      |                        | Sigmoid    | 6571   | 2456      |
|                      | Doubly Robust Estimate | ReLU       | 7935   | 2332      |
|                      |                        | Sigmoid    | 7959   | 1505      |
| 0.5                  | Difference of Means Estimate | ReLU       | 7711   | 1664      |
|                      |                        | Sigmoid    | 6462   | 2212      |
|                      | Doubly Robust Estimate | ReLU       | 8200   | 3410      |
|                      |                        | Sigmoid    | 7495   | 3057      |

Table 11: The real data results corresponding to Figure 10 for 400 training epochs.
Figure 6: The scaled density of the ATE estimate over 200 replications for different training sample sizes and different activation functions. Here we use a fixed inference sample size of $n = 1000$ and train each network for 400 epochs. The true treatment effect of $\tau = 1$ is shown as a red vertical line.

Figure 7: The scaled density of the ATE estimate over 200 replications for different training sample sizes and different activation functions. The red curves correspond to the DNN estimate defined in (10) and the blue curves correspond to the doubly robust estimate defined in (15). Here we use a fixed inference sample size of $n = 1000$ and train each network for 100 epochs. From top to bottom, the training sample size $n_1$ increases from 1000 to 5000.
Figure 8: The scaled density of the ATE estimate over 200 replications for different training sample sizes and different activation functions. The red curves correspond to the DNN estimate defined in (10) and the blue curves correspond to the doubly robust estimate defined in (15). Here we use a fixed inference sample size of $n = 1000$ and train each network for 200 epochs. From top to bottom, the training sample size $n_1$ increases from 1000 to 5000.

Figure 9: The scaled density of the ATE estimate over 200 replications for different training sample sizes and different activation functions. The red curves correspond to the DNN estimate defined in (10) and the blue curves correspond to the doubly robust estimate defined in (15). Here we use a fixed inference sample size of $n = 1000$ and train each network for 400 epochs. From top to bottom, the training sample size $n_1$ increases from 1000 to 5000.
Figure 10: The scaled density of the ATE estimate over 100 replications for different training sample size proportions and different activation functions. The red curves correspond to the DNN estimate defined in (10) and the blue curves correspond to the doubly robust estimate defined in (15). The red vertical line is the ATE estimate reported in [7] from the quadratic spline specification without variable selection of 8093. The rows in the figure correspond to different sizes of the inference set varying from 20% to 50% of the data. In this figure, both estimates come from networks trained for 200 epochs.

Figure 11: The scaled density of the ATE estimate over 100 replications for different training sample size proportions and different activation functions. The red curves correspond to the DNN estimate defined in (10) and the blue curves correspond to the doubly robust estimate defined in (15). The red vertical line is the ATE estimate reported in [7] from the quadratic spline specification without variable selection of 8093. The rows in the figure correspond to different sizes of the inference set varying from 20% to 50% of the data. In this figure, both estimates come from networks trained for 400 epochs.