Journal of Econometrics 238 (2024) 105555

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Causal inference of general treatment effects using neural networks with a diverging number of confounders☆

Xiaohong Chen[a], Ying Liu[b], Shujie Ma[b,*], Zheng Zhang[c,*]

[a] Department of Economics, Yale University, United States
[b] Department of Statistics, University of California at Riverside, United States
[c] Center for Applied Statistics, Institute of Statistics & Big Data, Renmin University of China, China

ARTICLE INFO

JEL classification:
C01
C12
C14

Keywords:
Artificial neural networks
Barron space
Mixed smoothness class
ReLU
Diverging confounders
Propensity score
Quantile treatment effects
Weighted bootstrap

ABSTRACT

Semiparametric efficient estimation of various multi-valued causal effects, including quantile treatment effects, is important in economic, biomedical, and other social sciences. Under the unconfoundedness condition, adjustment for confounders requires estimating the nuisance functions relating outcome or treatment to confounders nonparametrically. This paper considers a generalized optimization framework for efficient estimation of general treatment effects using artificial neural networks (ANNs) to approximate the unknown nuisance function of growing-dimensional confounders. We establish a new approximation error bound for the ANNs to the nuisance function belonging to a mixed smoothness class without a known sparsity structure. We show that the ANNs can alleviate the "curse of dimensionality" under this circumstance. We establish the root-n consistency and asymptotic normality of the proposed general treatment effects estimators, and apply a weighted bootstrap procedure for conducting inference. The proposed methods are illustrated via simulation studies and a real data application.

1. Introduction

Estimation of and inference on various causal effects using observational data have been popular in economics and other sciences. Under the unconfoundedness assumption, semiparametric efficient estimation of multi-valued Treatment Effects (TEs), including quantile TEs and asymmetric TEs, requires nonparametric estimation of a nuisance function relating outcome and/or treatment to confounders. Various nonparametric linear smoothers such as kernels and splines have been used in Outcome Regression (OR) or Inverse Probability Weighting (IPW) based TE studies. In many applied works, researchers believe that the unconfoundedness assumption is likely to hold conditioning on many confounders/covariates. However, popular nonparametric linear smoothers estimated nuisance function(s) of many covariates suffer from the so-called "curse of dimensionality". Artificial Neural Networks (ANNs) are nonlinear sieves that can approximate an unknown function of high dimensional covariates better than linear sieves such as splines, cosines, and polynomials. Moreover, recent computation advances have made the implementation of ANNs easier. This motivates our

☆ This paper is a revised version of "Efficient Estimation of General Treatment Effects using Neural Networks with A Diverging Number of Confounders", arxiv:2009.07055 (15 Sep 2020). This work has been presented since 2020 at various workshops and seminars including: UCSB, USC, UC-Irvine, Guanghua School of Management at Peking University, University of Chicago, Tsinghua University, and Toulouse School of Economics. The authors are equal contributors and are alphabetically ordered.

* Corresponding authors.
E-mail addresses: shujie.ma@ucr.edu (S. Ma), zhengzhang@ruc.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.jeconom.2023.105555
Received 17 May 2023; Received in revised form 28 July 2023; Accepted 8 August 2023
Available online 8 November 2023

investigation of ANN-based efficient estimation and inference on general TEs with increasing dimensional confounders, without known sparsity.

In this paper, we propose an ANN-based, root-*n* consistent, asymptotic normal, and efficient estimator of general multi-valued TE. Our TE estimator is obtained by directly optimizing a generalized objective function that involves an ANN-approximated nonparametric Propensity Score (PS) function, which is the only nuisance function to be estimated. Theoretically, we derive a new convergence rate of the ANN estimator for the nuisance function under mild conditions when the number of confounders is allowed to grow with the sample size (n). Our method can be naturally used to estimate general TEs, including the average, quantile, and asymmetric least squares TEs. In addition, our optimization procedure enables us to easily construct convenient weighted bootstrapped confidence sets, without the need of estimating the asymptotic variances that are of complicated forms for quantile TEs and asymmetric TEs.[1]

Feedforward ANNs are effective tools for solving the classification and prediction problems with high dimensional covariates and big data sets. The basic idea is to extract linear combinations of the inputs as features, and then model the target as a nonlinear function of these features. It has been shown in the literature (Barron, 1993; Chen and White, 1999; Hornik et al., 1994; Klusowski and Barron, 2018) that when the unknown target function admits a Fourier representation with a bounded moment, its ANNs approximator enjoys a fast approximation rate, making ANNs a promising tool to potentially break the notorious "curse of dimensionality" in nonparametric multivariate regression. This Fourier function class is recently named "Barron class" by E et al. (2022), who claim it is one right function space to address the curse of dimensionality problem. Nevertheless, it is of interest to investigate how the Barron class is related to some classical function spaces such as the Sobolev space (Stone, 1994; Wasserman, 2006) commonly used in the nonparametric regression literature.[2] Moreover, it is still unclear how the moment of the Fourier transform appeared in the ANN approximation error bounds depends on the dimension of the covariates. This moment is implicitly treated as a constant in the existing works on ANNs (Barron, 1993; Chen and White, 1999; Klusowski and Barron, 2018) as they consider fixed dimensions.

In this paper, we introduce a mixed smoothness class, and show that it is a subset of the Barron class. For any function in the mixed smoothness class, we derive an upper bound for the moment of its Fourier transform in terms of the dimension of the covariates. Functions in this mixed smoothness class need to be at least one order smoother in each coordinate than those in the standard Sobolev ball. We show that the nonlinear ANN estimators for functions in the mixed smoothness class have fast convergence rates. We also show that the conventional linear sieve approximators still suffer the "curse of dimensionality" when the target function belongs to the mixed smoothness class. Our new theoretical results enhance readers' understanding why single hidden layer ANNs perform better than nonparametric linear smoothers when estimating functions in a mixed smoothness class with increasing dimensional covariates.

While the development of credible inferential theories for the ANN-based estimator of TEs is essential to test the significance of the various causal effects, it is also a daunting task because of the complex nonlinear structure of the ANNs. In this paper, we establish the root-*n* asymptotic normality of our ANN-based TE estimator when the number of the confounders is allowed to grow with the sample size. Different from the earlier works on semiparametric efficient estimation and inference for TEs (see, e.g., Ai et al., 2021; Chen et al., 2008; Hirano et al., 2003; Robins et al., 1994), our semiparametric inferential theory allows for settings with diverging dimensional confounders. To the best of our knowledge, our paper is the first to provide a thorough theoretical justification for the ANN-based inferential procedures for general TEs when the dimension of the confounders can grow with the sample size.

Our ANN-based TE estimator is obtained directly from a generalized optimization procedure without estimating the Efficient Influence Function (EIF). The estimated EIF approach requires estimating two nuisance functions nonparametrically, while our optimization based procedure involves estimating one nuisance function only. Recently, Farrell et al. (2021) proposed an ANN-based Doubly Robust (DR) (or EIF based) estimator of average TEs, which involves estimating both OR and PS nuisance functions via ANNs. They assume that both nuisance functions belong to the standard Sobolev (or Hölder) ball and the dimension of the confounders is fixed. Although the EIF estimation-based method is commonly used for estimating average TEs, see for example Cao et al. (2009), Tan (2010), van der Laan and Rose (2011), Rotnitzky et al. (2012), Kennedy et al. (2017), Chernozhukov et al. (2018), it can be more difficult to apply in quantile, asymmetric and other complex TE settings, as these TE parameters can enter the estimated EIF equation in a nonlinear and non-separable fashion. When the nuisance functions are trained by nonlinear machine learning algorithms, it becomes even more computationally challenging to estimate the EIF in these complex settings. Our TE estimator is obtained directly from optimizing an objective function with a plug-in ANN-based estimator of the PS nuisance function, so a weighted bootstrap procedure can be conveniently applied for conducting inference without estimating the EIF nor the asymptotic variance function. To better illustrate our ANN-based TE estimation and inference procedures, we focus on using the ANNs with one hidden layer in the main text, and discuss the extension to ANNs with multiple hidden layers in the online supplement.

Finally, for those readers who care about efficient estimation of the averaged treatment effect (ATE) only, we also propose an ANN-based efficient estimator of the ATE obtained from the ANN estimated OR nuisance function. Under standard regularity conditions, our proposed ANN-PS and ANN-OR estimators for ATE have the same asymptotic distribution, and are both asymptotic efficient when the number of covariates is fixed. We show that, unlike the estimators using IPW and DR methods, the ANN-OR based ATE estimator can achieve the root-*n* asymptotic normality without imposing the strict overlap condition on the PS function. Consequently, the ANN-OR based ATE estimator is more robust than the ANN-PS based estimator when the true unknown PS function is close to zero. In our Monte Carlo simulations, we also observe that ANN-OR based ATE estimator performs slightly better than ANN-DR ATE estimator, which in

[1] See our online supplement for consistent estimation of the asymptotic variances. Nevertheless, our simulation studies indicate that bootstrapped CSs are more accurate than the CSs based on estimated asymptotic variance.

[2] The minimax optimal rate in root-mean squared error norm for estimating a function in the standard Sobolev ball is known, and no nonparametric estimator can avoid the "curse of dimensionality" for the standard Sobolev ball.

turn performs slightly better than ANN-IPW ATE estimator. However, it is difficult to apply the ANN-OR based procedure to estimate other types of multi-valued TEs such as quantile TEs.

The rest of the paper is organized as follows. Section 2 provides a new approximation error rate result for ANNs to a mixed smoothness class of functions with diverging dimension. Section 3 introduces the general multi-valued TEs and our proposed ANN-based estimators for TEs. Section 4 establishes the large sample properties and Section 5 presents the inferential procedures. Section 6 extends the optimization procedures and the asymptotic properties to general multi-valued treatment effects for the treated subgroups. Section 7 reports simulation studies and Section 8 contains a real data application. Section 9 briefly concludes. All the technical proofs and additional simulation results are provided in the Appendix and the on-line Supplemental Materials.

2. ANNs approximation for functions in the Barron class

For nonparametric estimation of a target function of high dimensional covariates, in addition to specifying the approximation basis, identifying a "good" target function space is also crucial in machine learning literature, as E et al. (2022) write:

> Sobolev/Besov type spaces are not the right function spaces for studying machine learning models that can potentially address the curse of dimensionality problem.

In this section we present ANNs with one hidden layer and the related approximation results for a target function in the Barron class.

Let $\mathscr{X}$ denote the support of a random vector $\boldsymbol{X}$ which is compact in $\mathbb{R}^p$. Without loss of generality, we assume $\mathscr{X} = [0,1]^p$. Let F_X be the cumulative distribution function (CDF) of $\boldsymbol{X}$. Denote the $L^2(dF_X)$-norm of any function $f(\cdot)$ by $\| f \|_{L^2(dF_X)} := \{\int_{\mathscr{X}} |f(x)|^2 dF_X(x)\}^{1/2}$. Let $\widetilde{f}(\boldsymbol{a})$ be the Fourier transform of $f(x)$ defined by

$$\widetilde{f}(\boldsymbol{a}) := \frac{1}{(2\pi)^p}\int_{\mathbb{R}^p} \exp(-i\boldsymbol{a}^\top x) f(x)dx,$$

where $\boldsymbol{a} = (a_1, ..., a_p)^\top \in \mathbb{R}^p$. We define the mth moment of the Fourier transform of $f(x)$ as $v_{f,m} := \int_{\mathbb{R}^p} |\boldsymbol{a}|_1^m |\widetilde{f}(a)| da$, where $|\boldsymbol{a}|_1 := \sum_{i=1}^p |a_i|$. Barron (1993), Hornik et al. (1994), Chen and White (1999), and Klusowski and Barron (2018) considered the target function belonging to the Barron class $\mathscr{F}_p^m$:

$$\mathscr{F}_p^m := \left\{ f : \mathscr{X} \to \mathbb{R} : f(x) = \int_{\mathbb{R}^p} \exp(i\boldsymbol{a}^\top x)\widetilde{f}(\boldsymbol{a})d\boldsymbol{a},\ v_{f,m} < \infty \right\} \tag{2.1}$$

$\mathscr{F}_p^m$ contains a class of functions of p dimension that admit Fourier representations with the finite mth moment. The input variables of functions in $\mathscr{F}_p^m$ have dimension p, which is allowed to grow to infinity as the sample size n increases. It is worth noting that $v_{f,m}$ depends on the dimension p, and its value can increase with p. In the nonparametric regression literature, spaces with certain smoothness constraints such as the Hölder or Sobolev space are instead more commonly used (Chen, 2007; Stone, 1994; Wasserman, 2006). We will build a connection between the function class $\mathscr{F}_p^m$ given in (2.1) and a mixed smoothness ball, and will establish an upper bound for $v_{f,m}$ in terms of the dimension p in Theorem 1. To the best of our knowledge, our paper is the first one that builds such a connection between the Fourier function class used for ANNs and a mixed smoothness ball and establishes an upper bound for $v_{f,m}$ which appears in the approximation error bounds.

Consider to approximate a target function $f \in \mathscr{F}_p^m$ using the ANNs, belonging to the class

$$\begin{aligned}\mathscr{G}(\psi, B, \quad r, p) = \{g : g(x) = g_0(x;\boldsymbol{\gamma}_0) + \frac{B}{r}\sum_{j=1}^r \gamma_j \psi\left(\boldsymbol{a}_j^\top x\right),\ \boldsymbol{a}_j = \left(a_{j1}, ..., a_{jp}\right)^\top \in \mathbb{R}^p,\\ \| \boldsymbol{a}_j \|_2 = 1,\ |\gamma_j| \le 1,\ j \in \{1, ..., r\},\ B \in \mathbb{R}^+\},\end{aligned} \tag{2.2}$$

where $g_0(x;\boldsymbol{\gamma}_0)$ is a parametric function indexed by an unknown parameter vector $\boldsymbol{\gamma}_0$, and $\| \boldsymbol{a}_j \|_2 := \{|a_{j1}|^2 + ... + |a_{jp}|^2\}^{1/2}$. The structure of $g_0(x;\boldsymbol{\gamma}_0)$ depends on the type of the activation function that is used. For example, if the ReLU activation function is used, then $g_0(x;\boldsymbol{\gamma}_0) = \boldsymbol{\gamma}_0^\top x$. $\mathscr{G}(\psi, B, r, p)$ is the collection of output functions for neural networks with p-dimensional input feature x, a single hidden layer with r hidden units and an activation function ψ and real-valued input-to-hidden unit weights $(\boldsymbol{a}_j)$, and hidden-to-output weights $(\gamma_j^* := B\gamma_j)$. Note that for any outer parameter $\gamma_j^* \in \mathbb{R}$, it can be written as $\gamma_j^* = B\gamma_j$ with $|\gamma_j| \le 1$, and B is a scale factor of all γ_j^*'s. Both r and B are allowed to increase with the sample size n, which will be discussed in Section 3.1. The ANN class in (2.2) comes from the class given in Klusowski and Barron (2018). The difference is that we change their ℓ_1 constraint on the inner parameters $\boldsymbol{a}_j$ to a ℓ_2 normalization $\| \boldsymbol{a}_j \|_2 = 1$. This normalization has been commonly used in semiparametric index models, see Ma and He (2016).

The approximation error for a target function depends on the smoothness of the approximand, the dimension of the covariates, and the type of approximation basis. We first present the approximation results based on some popularly used neural networks, which have been established in the existing literature:

- (Sigmoid type activation function) Suppose that the function $f \in \mathscr{F}_p^1$, $g_0(x;\boldsymbol{\gamma}_0) \equiv 0$, the activation function $\psi(\cdot)$ is compactly supported, bounded, and uniformly Lipschitz continuous. If $B \leq 2v_{f,1} < \infty$, then Chen and White (1999, Theorem 2.1) show that the $L^2(dF_X)$-approximation rate of f based on ANN is

$$\inf_{g\in\mathscr{G}(\psi,B,r,p)}\left\{\int_{\mathscr{X}}|f(x)-g(x)|^2dF_X(x)\right\}^{1/2} \leq \text{const} \times v_{f,1}\cdot r^{-\frac{1}{2}-\frac{1}{p}}. \tag{2.3}$$

The activation functions ψ include the Heaviside, logistic, tanh, cosine squasher, and other sigmoid functions (Hornik et al., 1994; Makovoz, 1996), but do not include the ReLU and squared ReLU ridge functions stated below.

- (ReLU activation function) Suppose that the function $f \in \mathscr{F}_p^2$, $g_0(x;\boldsymbol{\gamma}_0) = \boldsymbol{\gamma}_0^\top \boldsymbol{x}$ for $\boldsymbol{\gamma}_0 \in \mathbb{R}^p$, and $\psi(\boldsymbol{a}^\top\boldsymbol{x}) = (\boldsymbol{a}^\top\boldsymbol{x})_+$. If $B \leq 2v_{f,2} < \infty$, then Klusowski and Barron (2018, Theorem 2 and its discussion on page 7651) show that the $L^2(dF_X)$-approximation rate based on ReLU ridge functions is

$$\inf_{g\in\mathscr{G}(\psi,B,r,p)}\left\{\int_{\mathscr{X}}|f(x)-g(x)|^2dF_X(x)\right\}^{1/2} \leq \text{const} \times v_{f,2}\cdot r^{-\frac{1}{2}-\frac{1}{p}}. \tag{2.4}$$

- (Squared ReLU activation function) Suppose that the function $f \in \mathscr{F}_p^3$, $g_0(x;\boldsymbol{\gamma}_0) = \gamma_{01}^\top \boldsymbol{x} + \boldsymbol{x}^\top\gamma_{02}\cdot\boldsymbol{x}$ for $\boldsymbol{\gamma}_0 = \{\gamma_{01},\gamma_{02}\} \in \mathbb{R}^p \times \mathbb{R}^{p\times p}$, and $\psi(\boldsymbol{a}^\top\boldsymbol{x}) = (\boldsymbol{a}^\top\boldsymbol{x})_+^2$. If $B \leq 2v_{f,3} < \infty$, then Klusowski and Barron (2018, Theorem 3 and its discussion on page 7651) show that the $L^2(dF_X)$-approximation rate based on squared ReLU ridge functions is

$$\inf_{g\in\mathscr{G}(\psi,B,r,p)}\left\{\int_{\mathscr{X}}|f(x)-g(x)|^2dF_X(x)\right\}^{1/2} \leq \text{const} \times v_{f,3}\cdot r^{-\frac{1}{2}-\frac{1}{p}}. \tag{2.5}$$

If the target function $f(x)$ is in a Barron class with a finite moment given in (2.1), then (2.3), (2.4) and (2.5) show that the $L^2(dF_X)$-approximation rates of ANNs are $O(v_{f,m}\cdot r^{-1/2-1/p}) = o(v_{f,m}\cdot r^{-1/2})$ for $m = 1, 2, 3$, in which $r^{-1/2}$ no longer depends on the dimension p. Thus, the resulting ANNs estimator can break the "curse of dimensionality" that typically arises in the nonparametric kernel and linear sieve estimation.

Remark 1. Recently, DeVore et al. (2023) propose weighted variation spaces that enlarge the second order Barron space $\mathscr{F}_p^2$. They show that the $L^2(dF_X)$-approximation rate based on the shallow ReLU neural networks for a function $f(\cdot)$ belonging to their weighted variation spaces achieves const $\times \parallel f \parallel_{\mathscr{V}_w}\cdot r^{-\frac{1}{2}-\frac{3}{2p}}$, where $\parallel\cdot\parallel_{\mathscr{V}_w}$ is the weighted variation norm defined in DeVore et al. (2023, Section 4, page 8). Our theoretical results for the classic Barron space presented in this article can also be adapted to the weighted variation spaces. We focus on the Barron space for easing the presentation.

2.1. *The mixed smoothness ball*

To the best of our knowledge, there are two questions that remain unanswered in the literature, including (1) how restrictive this Barron class $\mathscr{F}_p^m$ is compared to the conventional smoothness spaces such as the Sobolev ball typically assumed in the multivariate nonparametric regression literature, and (2) how the moment of the Fourier transform $v_{f,m}$ depends on the dimension p. To address these questions, we build a connection between the Barron class $\mathscr{F}_p^m$ and a mixed smoothness ball, and establish an upper bound for $v_{f,m}$.

Given a p-tuple $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ of nonnegative integers, set $|\boldsymbol{\alpha}|_1 := \sum_{j=1}^p \alpha_j$ and let $D^{\boldsymbol{\alpha}}$ denote the differential operator defined by

$$D^{\boldsymbol{\alpha}} := \frac{\partial^{|\boldsymbol{\alpha}|_1}}{\partial x_1^{\alpha_1}\cdots\partial x_p^{\alpha_p}}.$$

For an integer $s \in \mathbb{N}$, denote the class of all s-times continuously differentiable real-valued functions on $\mathscr{X}$ by $\mathscr{C}^{s,\infty}(\mathscr{X})$:

$$\mathscr{C}^{s,\infty}(\mathscr{X}) := \left\{f(\,\cdot\,):\ \sup_{|\boldsymbol{\alpha}|_1\leq s}|D^{\boldsymbol{\alpha}}f(x)| \leq 1\right\}. \tag{2.6}$$

The upper bound "1" used in the definition of $\mathscr{C}^{s,\infty}(\mathscr{X})$ is only for notational simplicity and it can be replaced by any finite positive constant. The function class $\mathscr{C}^{s,\infty}(\mathscr{X})$ is called the Sobolev ball (in L^∞-norm) of order s.

Let $\partial_{x_1}\cdots\partial_{x_p}$ be the partial derivative with respect to $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ defined by

$$\partial_{x_1}\cdots\partial_{x_p} := \frac{\partial^p}{\partial x_1\cdots\partial x_p}.$$

For any $\delta > 0$, define $\Delta^{\delta}_{x_i}$ to be the difference operator by

$$\Delta^{\delta}_{x_i} f\left(x_1, \dots, x_p\right) := f\left(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_p\right) - f\left(x_1, \dots, x_{i-1}, x_i - \delta, x_{i+1}, \dots, x_p\right),$$

for $i \in \{1, 2, \dots, p\}$. We have

$$\lim_{\delta_1, \dots, \delta_p \to 0} \frac{\Delta^{\delta_p}_{x_p} \cdots \Delta^{\delta_1}_{x_1} f(x)}{\delta_1 \cdots \delta_p} = \partial_{x_1} \cdots \partial_{x_p} f(x),$$

provided that the partial derivative $\partial_{x_1} \cdots \partial_{x_p} f(x)$ exists.

Being motivated by the application of sparse grids in dealing with high-dimensional partial differential equations (PDEs) (Bungartz and Griebel, 2004), we define $\mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X})$ for $m \in \mathbb{N}$ and $\epsilon \in (0,1]$ to be the *mixed smoothness ball* of order $(m, 1 + \epsilon)$:

$$\begin{aligned} \mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X}) := \Bigg\{ f : f(\cdot) \in \mathscr{C}^{m,\infty}(\mathscr{X}), \ & \sup_{\{\forall \boldsymbol{\alpha} : |\boldsymbol{\alpha}|_1 = m\}} \sup_{x \in \mathscr{X}} \left| \partial_{x_1} \cdots \partial_{x_p} D^{\boldsymbol{\alpha}} f(x) \right| \leq 1, \\ & \sup_{\{\forall \boldsymbol{\alpha} : |\boldsymbol{\alpha}|_1 = m\}} \sup_{\{x \in \mathscr{X}, \delta_1 > 0, \dots, \delta_p > 0\}} \frac{\left| \Delta^{\delta_p}_{x_p} \cdots \Delta^{\delta_1}_{x_1} \partial_{x_1} \cdots \partial_{x_p} D^{\boldsymbol{\alpha}} f(x) \right|}{\delta_1^{\epsilon} \cdots \delta_p^{\epsilon}} \leq 1 \Bigg\} \end{aligned} \tag{2.7}$$

The following result states that $\mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X})$ with $\epsilon \in (0,1]$ is a subspace of $\mathscr{F}^m_p$. The proof is relegated to Appendix B.

Theorem 1. *Let $f \in \mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X})$ for some $\epsilon \in (0,1]$, then $f \in \mathscr{F}^m_p$ and*

$$v_{f,m} \leq 2 \cdot \left(M_0 \cdot \frac{\pi}{2}\right)^m \cdot M^p$$

for some universal constant M defined by $M := M_0 \cdot \left(\frac{\pi}{2}\right) \cdot \left(\frac{1}{2\epsilon} + \frac{1}{2}\right)$, where the definition of M_0 is given in (B.3). To the best of our knowledge, Theorem 1 is the first result in the literature that provides a connection between the mixed smoothness ball $\mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X})$ and $\mathscr{F}^m_p$. Theorem 1 shows that the mixed smoothness ball $\mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X})$ is a subspace of the Barron class $\mathscr{F}^m_p$. Thus, for any functions in $\mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X})$, their ANNs approximators enjoy the nice approximation rates given in (2.3)–(2.5). Furthermore, Theorem 1 explicitly provides an upper bound for the mth moment of the Fourier transform $v_{f,m}$, which enables us to evaluate the effect of the dimension p on the approximation rates of ANNs. This upper bound has not been provided in the literature, and the theories in most existing works on ANNs are established by assuming that p is fixed.

Remark 2. In particular, if $\epsilon = 1$ in (2.7), we obtain

$$\mathscr{W}^{m,2,\infty}(\mathscr{X}) := \left\{ f(\cdot) : f(\cdot) \in \mathscr{C}^{m,\infty}(\mathscr{X}) \text{ and } \sup_{\{\forall \boldsymbol{\alpha} : |\boldsymbol{\alpha}|_1 = m\}} \sup_{x \in \mathscr{X}} \left| \partial^2_{x_1} \cdots \partial^2_{x_p} D^{\boldsymbol{\alpha}} f(x) \right| \leq 1 \right\},$$

where

$$\partial^2_{x_1} \cdots \partial^2_{x_p} := \frac{\partial^{2p}}{\partial x_1^2 \cdots \partial x_p^2}.$$

The functions in the mixed smoothness ball $\mathscr{W}^{m,2,\infty}(\mathscr{X})$ need to be 2-order smoother in each coordinate of X than the functions in the regular Sobolev ball of order $s = m$ given in (2.6), i.e. $\mathscr{C}^{m,\infty}(\mathscr{X})$. It is worth noting that $\mathscr{W}^{m,2,\infty}(\mathscr{X})$ is much broader than the Sobolev ball of order $s = m + 2p$; indeed, we have the following inclusion relation:

$$\mathscr{C}^{m+2p,\infty}(\mathscr{X}) \subsetneq \mathscr{W}^{m,2,\infty}(\mathscr{X}) \subset \mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X}) \subset \mathscr{F}^m_p, \quad \text{for } \epsilon \in (0,1].$$

The functions in this mixed smoothness ball do not need a compositional structure such as a hierarchical interaction structure considered in Bauer and Kohler (2019) and Schmidt-Hieber (2020). We should be mindful that breaking the curse of dimensionality happens at the cost of sacrificing flexibility. If a function is assumed to be in the Sobolev ball of order m, the nonparametric optimal minimax rates suffer from the curse of dimensionality, i.e., no nonparametric estimator can avoid the dimensionality problem under this condition (Schmidt-Hieber, 2020).

3. Parameters of interest and ANN-based estimators

In this section, we first define our general treatment effect parameters of interest, and then introduce our ANN optimization based estimators.

Let D denote a treatment variable taking value in $\mathscr{D} = \{0, 1, \dots, J\}$, where $J \geq 1$ is a positive integer. Let $Y^*(d)$ denote the potential outcome when the treatment status $D = d$ is assigned. The probability density of $Y^*(d)$ exists, denoted by $f_{Y^*(d)}$, is continuously differentiable. Let $\mathscr{L}(\cdot)$ denote a nonnegative and strictly convex loss function satisfying $\mathscr{L}(0) = 0$ and $\mathscr{L}(v) \geq 0$ for all $v \in \mathbb{R}$. The derivative of $\mathscr{L}(\cdot)$ exists almost everywhere and non-constant which is denoted by $\mathscr{L}^{'}(\cdot)$. Let $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_J^*)^\top \in \mathbb{R}^{J+1}$ be the parameter of interest which is uniquely identified through the following optimization problem:

$$\boldsymbol{\beta}^* := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{d=0}^{J} \mathbb{E}[\mathscr{L}(Y^*(d) - \beta_d)], \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)^\top \in \mathbb{R}^{J+1}$ and $J \in \mathbb{N}$. The formulation (3.1) permits various definitions of treatment effect (TE) parameters, some of which have been considered in the literature. For example,

- $\mathscr{L}(v) = v^2$ and $J = 1$, then $\beta_0^* = \mathbb{E}[Y^*(0)]$ and $\beta_1^* = \mathbb{E}[Y^*(1)]$, and $\beta_1^* - \beta_0^*$ is the average treatment effects (ATE) studied by Hahn (1998), Hirano et al. (2003), Chan et al. (2016) and many others. When $J \geq 2$, then $\beta_d^* = \mathbb{E}[Y^*(d)]$ is the multi-valued ATE first studied by Cattaneo (2010).
- $\mathscr{L}(v) = v \cdot \{\tau - \mathbb{1}(v \leq 0)\}$ for some $\tau \in (0,1)$ and $J = 1$, then $\beta_0^* = F_{Y^*(0)}^{-1}(\tau)$ and $\beta_1^* = F_{Y^*(1)}^{-1}(\tau)$, and $\beta_1^* - \beta_0^*$ is the quantile treatment effects (QTE, Chen et al., 2008; Firpo, 2007; Han et al., 2019).
- $\mathscr{L}(v) = v^2 \cdot |\tau - \mathbb{1}(v \leq 0)|$ is the asymmetric least square treatment effects (ALSTE, Newey and Powell, 1987). ALSTE estimators have properties analogue to QTE estimators, but they are easier to compute. ALSTE has a variety of applications, such as the study of racial/ethnic disparities in health care, in which the data are often skewed.

The problem with (3.1) is that the potential outcomes $(Y^*(0), Y^*(1), \ldots, Y^*(J))$ cannot all be observed. The observed outcome is denoted by $Y := Y^*(D) = \sum_{d=0}^{J+1} \mathbb{1}(D = d) Y^*(d)$. One may attempt to solve the problem:

$$\min_{\boldsymbol{\beta}} \sum_{d=0}^{J} \mathbb{E}[\mathscr{L}(Y - \beta_d)].$$

However, due to the selection in treatment, the true value $\boldsymbol{\beta}^*$ is not the solution of the above problems. To address this problem, most literature imposes the following *unconfoundedness* condition (Rosenbaum and Rubin, 1983):

Assumption 1. For each $d \in \mathscr{D}$, $Y^*(d) \perp D | \boldsymbol{X}$.

This condition is also maintained in our work. Nevertheless, we depart from the classical semiparametric estimation and inference for various TEs by allowing the dimension p of the confounders $\boldsymbol{X}$ to grow with sample size n. Specifically, we work with triangular array data $\{((D_{i,n}, \boldsymbol{X}_{i,n}, Y_{i,n}), i = 1, \ldots, n), n = 1, 2, \ldots\}$ defined on some common probability space $(\Omega, \mathscr{A}, \mathbb{P})$. Each $\boldsymbol{X}_{i,n}$ is a vector whose dimension p_n may grow with n, the support of $\boldsymbol{X}_{i,n}$ is assumed to be $[0,1]^{p_n}$. For each given n, these vectors are independent across i, but not necessarily identically distributed. The law $\mathbb{P}_n$ of $\{(D_{i,n}, \boldsymbol{X}_{i,n}, Y_{i,n}), i = 1, \ldots, n\}$ can change with n, though we do not make explicit use of $\mathbb{P}_n$. Thus, all parameters (including p_n) that characterize the distribution of $\{(D_{i,n}, \boldsymbol{X}_{i,n}, Y_{i,n}), i = 1, \ldots, n\}$ are implicitly indexed by the sample size n, but we omit the index n in what follows to simplify notation.

3.1. ANN-IPW estimator for general TEs

Under Assumption 1, the causal parameters $\boldsymbol{\beta}^*$ can be identified by the minimizer of the following optimization problem:

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{d=0}^{J} \mathbb{E}\left[\frac{\mathbb{1}(D_i = d)}{\pi_d^*(\boldsymbol{X}_i)} \mathscr{L}(Y_i - \beta_d)\right], \tag{3.2}$$

where $\pi_d^*(\boldsymbol{X}_i) := \mathbb{P}(D_i = d | \boldsymbol{X}_i)$ is the propensity score (PS) function which is unknown in practice.

Based on (3.2), existing approaches rely on parametric or nonparametric estimation of the PS function $\pi_d^*(\cdot)$. Parametric methods suffer from model misspecification problems, while conventional nonparametric methods, such as linear sieve or kernel regression, fail to work if the dimension of covariates p is large which is known as the "curse of dimensionality". The goal of this article is to efficiently estimate $\boldsymbol{\beta}^*$ under this general framework when the dimension of covariates p is large, and it possibly increases as the sample size n grows. We propose to estimate the PS function $\pi_d^*(\cdot)$ using feedforward ANNs with one hidden layer described below.

All three ANNs described in Section 2 can be applied to estimate the PS function $\pi_d^*(\cdot)$, and the resulting TE estimators have the same asymptotic properties based on the three ANNs. For convenience of presentation, we use the sigmoid type ANNs to present the theoretical results in this section. To facilitate our subsequent statistical applications, we allow $r = r_n$ and $B = B_n$ to depend on sample size n. We denote the resulting ANN sieve space as

$$\mathscr{G}_n := \mathscr{G}(\psi, B_n, r_n, p).$$

Denote $D_{di} := \mathbb{1}(D_i = d)$ for brevity. Let $L(a) := \exp(a)/(1 + \exp(a))$, for $a \in \mathbb{R}$, be the logistic function. The inverse logistic transform of the true PS is defined by

$$g_d^*(x) := L^{-1}\left(\pi_d^*(x)\right) = \log\left\{\pi_d^*(x) \,/\, \left(1 - \pi_d^*(x)\right)\right\},$$

and it satisfies $\mathbb{E}[\ell_d(D_{di}, \boldsymbol{X}_i; g_d^*)] \geq \mathbb{E}[\ell_d(D_{di}, \boldsymbol{X}_i; g_d)]$ for all $g_d \in \mathscr{G}_n$, where

$$\begin{aligned}\ell_d(D_{di}, \boldsymbol{X}_i; g_d) &:= D_{di}\log L(g_d(\boldsymbol{X}_i)) + \{1 - D_{di}\}\log(1 - L(g_d(\boldsymbol{X}_i))) \\ &= D_{di} g_d(\boldsymbol{X}_i) - \log[1 + \exp(g_d(\boldsymbol{X}_i))].\end{aligned}$$

Let $\widehat{g}_d$ be the ANN estimator of g_d^* based on the space $\mathscr{G}_n$, i.e.

$$L_{d,n}(\widehat{g}_d) \geq \sup_{g_d \in \mathscr{G}_n} L_{d,n}(g_d) - O\left(\epsilon_n^2\right), \tag{3.3}$$

where $L_{d,n}(g_d) := n^{-1}\sum_{i=1}^n \ell_d(D_{di}, \boldsymbol{X}_i; g_d)$ is the empirical criterion, and $\epsilon_n = o(n^{-1/2})$.

The ANN estimator of g_d^* depends on the sample size n. For notational simplicity, we omit the index n. (3.3) states that the ANN estimator $\widehat{g}_d$ of g_d^* does not need to be the global maximizer of the objective function $L_{d,n}(g_d)$, which may not be obtained in practice. It can be any local solutions satisfying (3.3), i.e., the values of the objective function evaluated at the local solutions and at the global maximizer cannot be far away from each other, and their difference needs to satisfy the order $O(\epsilon_n^2)$. This assumption is also imposed for sieve extreme estimation; see Shen (1997), Chen and Shen (1998) and Chen and White (1999). The estimator of π_d^* is defined by $\widehat{\pi}_d := L(\widehat{g}_d)$, then we use the empirical version of (3.2) to construct the estimator of $\boldsymbol{\beta}^*$, denoted by $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, ..., \widehat{\beta}_J)^\top$ where

$$\widehat{\beta}_d := \operatorname*{argmin}_{\beta \in \Theta} \frac{1}{n}\sum_{i=1}^n \frac{D_{di}}{\widehat{\pi}_d(\boldsymbol{X}_i)} \mathscr{L}(Y_i - \beta), \tag{3.4}$$

for every $d \in \mathscr{D} = \{0, 1, ..., J\}$. $\widehat{\boldsymbol{\beta}}$ is called the artificial neural networks-based inverse probability weighting (ANN-IPW) estimator.

3.2. ANN-OR estimator for ATE

In this subsection, we consider an alternative estimator for a particularly important parameter of interest, ATE, which corresponds to a loss function $\mathscr{L}(v) = v^2$. Using Assumption 1 and the property of conditional expectation, we can rewrite (3.1) as follows:

$$\boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{d=0}^J \mathbb{E}[\mathbb{E}[\mathscr{L}(Y_i - \beta_d)|\boldsymbol{X}_i, D_i = d]]. \tag{3.5}$$

Based on the above expression, an alternative estimation strategy for $\boldsymbol{\beta}^*$ is to first estimate the conditional expectation $\mathbb{E}[\mathscr{L}(Y_i - \beta_d)|\boldsymbol{X}_i, D_i = d]$ (with β_d being fixed), and then estimate $\boldsymbol{\beta}^*$ by minimizing the empirical version of (3.5) with $\mathbb{E}[\mathscr{L}(Y_i - \beta_d)|\boldsymbol{X}_i, D_i = d]$ replaced by its estimate. Unlike the estimator $\widehat{\mathscr{E}}_d(x)$ in (F.4) of the supplement, where $\widehat{\beta}_d$ involved in $\widehat{\mathscr{E}}_d(x)$ is separately obtained through (3.4), solving an empirical version of (3.5) is difficult for a general $\mathscr{L}(\cdot)$, since $\boldsymbol{\beta}^*$ is involved in the ANN estimator of $\mathbb{E}[\mathscr{L}(Y_i - \beta_d)|\boldsymbol{X}_i, D = d]$ which may not have a closed-form expression.

In this case, $\beta_d^* = \mathbb{E}[Y^*(d)] = \mathbb{E}[z_d^*(\boldsymbol{X}_i)]$, where $z_d^*(\boldsymbol{X}_i) := \mathbb{E}[Y_i|\boldsymbol{X}_i, D_i = d]$ is the outcome regression (OR) function and satisfies $\mathbb{E}[\ell_d^{OR}(D_{di}, \boldsymbol{X}_i, Y_i; z_d^*)] \geq \mathbb{E}[\ell_d^{OR}(D_{di}, \boldsymbol{X}_i, Y_i; z_d)]$ for all $z_d \in \mathscr{G}_n$, where

$$\ell_d^{OR}(D_{di}, \boldsymbol{X}_i, Y_i; z_d) := -D_{di}\{Y_i - z_d(\boldsymbol{X}_i)\}^2.$$

Let $\widehat{z}_d$ be the ANN estimator of z_d^* based on the space $\mathscr{G}_n$, i.e.

$$L_{d,n}^{OR}(\widehat{z}_d) \geq \sup_{z_d \in \mathscr{G}_n} L_{d,n}^{OR}(z_d) - O\left(\epsilon_n^2\right), \tag{3.6}$$

where $L_{d,n}^{OR}(z_d) := n^{-1}\sum_{i=1}^n \ell_d^{OR}(D_{di}, \boldsymbol{X}_i, Y_i; z_d)$ is the empirical criterion, and $\epsilon_n = o(n^{-1/2})$. Then the ANN-OR estimator of β_d^* is defined to be

$$\widehat{\beta}_d^{OR} = \frac{1}{n}\sum_{i=1}^n \widehat{z}_d(\boldsymbol{X}_i). \tag{3.7}$$

4. Large sample properties of estimators

4.1. Properties of the ANN-IPW estimator for general TEs

We first introduce sufficient conditions for the convergence rates of our ANN estimators $\{\widehat{\pi}_d\}_{d=0}^J$ for the unknown PS nuisance functions.

Assumption 2. For every $d \in \mathscr{D} = \{0, 1, ..., J\}$ and $m \geq 1$, we assume $g_d^*(\cdot) \in \mathscr{F}_p^m$.

Assumption 3. (i) The dimension of $\boldsymbol{X}_i$ is denoted by $p \in \mathbb{N}$ and the number of hidden units is denoted by $r_n \in \mathbb{N}$. They satisfy

$$\max\left\{ v_{g_d^*,m} \cdot r_n^{-\frac{1}{2}-\frac{1}{p}}, \sqrt{\frac{r_n \cdot p \cdot \log n}{n}} \right\} = o\left(n^{-\frac{1}{4}}\right).$$

(ii) The bound of the hidden-to-output weights, B_n, specified in (2.2) satisfies $B_n \leq 2v_{g_d^*,m}$. Assumption 2 is a smoothness condition imposed on the transformed PS functions. Assumption 3 (i) allows the dimension of covariates going to infinity as the sample size grows, while it imposes restrictions on the growth rate of the dimension of covariates and that of the number of hidden units to ensure that the $L^2(dF_X)$-convergence rate of estimated PS attains $o_P(n^{-1/4})$, which is needed to establish the $\sqrt{n}$-asymptotic normality for the proposed TE estimator.

Remark 3. As shown in Theorem 1, the mixed smoothness ball $\mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X})$ belongs to the Barron space $\mathscr{F}_p^m$. Assumption 3 (i) can be implied by the following primitive condition:

> **Assumption** 3’. (i) Suppose $g_d^* \in \mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X})$, p and r_n are allowed to grow to infinity as the sample size n increases, with the rates
>
> $$p \leq a_n \cdot (\log n)^{\frac{1}{2}} \text{ and } C_1 \cdot n^{\frac{p+1}{2(p+2)}} \leq r_n \leq C_2 \cdot (\log n)^{-\frac{3}{2}} \cdot n^{\frac{1}{2}},$$

where $a_n \to 0$ can be arbitrarily slow, and C_1 and C_2 are two positive constants. The following result establishes the convergence rates of g_d^* and π_d^*.

Theorem 2. *Suppose Assumptions 2 and 3 hold. Then*

$$\| \widehat{g}_d - g_d^* \|_{L^2(dF_X)} = O_P\left(\max\left\{ v_{g_d^*,m} \cdot r_n^{-\frac{1}{2}-\frac{1}{p}}, \sqrt{\frac{r_n \cdot p \cdot \log n}{n}} \right\} \right) = o_P\left(n^{-1/4}\right),$$

and

$$\| \widehat{\pi}_d - \pi_d^* \|_{L^2(dF_X)} = O_P\left(\max\left\{ v_{g_d^*,m} \cdot r_n^{-\frac{1}{2}-\frac{1}{p}}, \sqrt{\frac{r_n \cdot p \cdot \log n}{n}} \right\} \right) = o_P\left(n^{-1/4}\right),$$

where the constants hiding inside O_P and o_P do not depend on p and n. The proof of Theorem 2 is provided in Supplement B. Theorem 2 shows that under a suitable smoothness condition, the M-estimates based on ANNs with a single hidden layer circumvent the curse of dimensionality and achieve a desirable rate for establishing the asymptotic normality of plug-in estimators (Chen et al., 2003). Bauer and Kohler (2019) showed that their least squares estimator based on multilayer neural networks with a smooth activation function can achieve the convergence rate of $n^{-2s/(2s+d^*)}$ (up to a log factor), if the regression function satisfies a s-smooth generalized hierarchical interaction model of order d^*, where d^* is fixed. Schmidt-Hieber (2020) established a similar rate for the ReLU activation function. However, the target function class considered in Bauer and Kohler (2019) and Schmidt-Hieber (2020) is different from that used in our paper. The extension of our results for multilayer neural networks is beyond the scope of the current article. We refer to the Supplement for more discussion.

Let $\mathscr{E}_d(x,\beta_d^*) := \mathbb{E}[\mathscr{L}^{'}(Y_i^*(d) - \beta_d^*)|\boldsymbol{X}_i = x]$, $u_d^*(x) := \mathscr{E}_d(x;\beta_d^*)/\pi_d^*(x)$, and $\overline{g}(g_d,\epsilon_n) := (1-\epsilon_n)\cdot g_d + \epsilon_n \cdot \{u_d^* + g_d^*\}$ be a local alternative value around $g_d \in \mathscr{G}_n$. The directional derivative of $\ell_d(D_{di},\boldsymbol{X}_i;g_d)$ is given by

$$\begin{aligned}\frac{\partial}{\partial g_d}\ell_d(D_{di},\boldsymbol{X}_i;g_d)[u] &:= \lim_{t\to 0}\frac{\ell_d(D_{di},\boldsymbol{X}_i;g_d + t\cdot u) - \ell_d(D_{di},\boldsymbol{X}_i;g_d)}{t}\\ &= \{D_{di} - L(g_d(\boldsymbol{X}_i))\}u(\boldsymbol{X}_i),\ \text{ for } u \in L^2(dF_X).\end{aligned}$$

We now introduce sufficient conditions and additional notation for the asymptotic normality of our ANN-IPW estimators $\widehat{\boldsymbol{\beta}}$ for the general TE parameters.

Assumption 4. (i) Let Θ be a compact set of $\mathbb{R}^{J+1}$ containing the true parameters $\boldsymbol{\beta}^*$. (ii) The propensity scores are uniformly bounded away from zero, i.e., there exists a constant $\underline{c}$ such that $0 < \underline{c} \leq \pi_d^*(x)$ for all $x \in \mathscr{X}$ and $d \in \{0,1,\ldots,J\}$. (iii) For every $d \in \{0,1,\ldots,J\}$ and $m \geq 1$, we assume the function $\mathscr{E}_d(\cdot,\beta_d^*)$ is uniformly bounded.

Assumption 5. (Approximation error) We assume the following conditions hold:

$$\sup_{\left\{g_d \in \mathscr{G}_n : \|g_d - g_d^*\|_{L^2(dF_X)} \leq \delta_n\right\}} \| \operatorname{Proj}_{\mathscr{G}_n} \overline{g}(g_d, \epsilon_n) - \overline{g}(g_d, \epsilon_n) \|_{L^2(dF_X)} = O\left(\frac{\epsilon_n^2}{\delta_n}\right),$$

and

$$\sup_{\left\{g_d \in \mathscr{G}_n : \|g_d - g_d^*\|_{L^2(dF_X)} \leq \delta_n\right\}} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial}{\partial g_d} \ell_d\left(D_{di}, \boldsymbol{X}_i; g_d^*\right) \left[\overline{g}(g_d, \epsilon_n) - \operatorname{Proj}_{\mathscr{G}_n} \overline{g}(g_d, \epsilon_n)\right] \right) = O_P\left(\epsilon_n^2\right),$$

where $\operatorname{Proj}_{\mathscr{G}_n} \overline{g}(g_d, \epsilon_n)$ denotes the $L^2(dF_X)$-projection of $\overline{g}(g_d, \epsilon_n)$ on the ANN space $\mathscr{G}_n$ and δ_n is a sequence of positive real numbers satisfying $\| \widehat{g}_d - g_d^* \|_{L^2(dF_X)} = o_P(\delta_n)$.

Assumption 6. 1. There exists a finite positive constant $\kappa \geq 1/2$ such that for any $\beta \in \Theta$ and any $\delta > 0$ in a neighborhood of zero,

$$\left\{ \mathbb{E}\left[\sup_{\widetilde{\beta} : |\widetilde{\beta} - \beta| < \delta} \{\mathscr{L}'(Y - \widetilde{\beta}) - \mathscr{L}'(Y - \beta)\}^2 \right] \right\}^{1/2} \leq \text{const} \times \delta^{\kappa};$$

2. $\sup_{\beta \in \Theta} \mathbb{E}[|\mathscr{L}'(Y - \beta)|^2] \langle \infty$ and $\mathbb{E}[\sup_{\beta \in \Theta} |\mathscr{L}'(Y - \beta)|] \langle \infty$;
3. $\sup_{x \in \mathscr{X}} \mathbb{E}[|\mathscr{L}'(Y - \beta_d^*)| | \boldsymbol{X} = x] \langle C < \infty$ for some finite constant $C > 0$;
4. $H_d := -\partial_{\beta_d} \mathbb{E}[\mathscr{L}'(Y^*(d) - \beta_d^*)] \rangle 0$.

Assumption 4 (i) is a standard condition for the parameter space. Assumption 4 (ii) is a strict overlap condition ensuring the existence of participants at all treatment levels, which is commonly assumed in the literature. D'Amour et al. (2021) discussed the applicability of the strict overlap condition with high-dimensional covariates, and provided a variety of circumstances under which this condition holds. They also argued that the strict overlap condition may not be necessary if other smoothness conditions are imposed on the potential outcomes, or it can be technically relaxed with some non-standard asymptotic analyses (e.g. Hong et al., 2020; Ma and Wang, 2020) and the sacrifice of uniform inference on ATE. Assumption 4 (iii) is a smoothness condition for approximation. The functions $\{\pi_d^*(\cdot), \mathscr{E}_d(\cdot, \beta_d^*)\}_{d=0}^J$ generally depend on the sample size n. Assumption 5 specifies both approximation error and stochastic equicontinuity in neural network space, which is needed for establishing Lemma in the supplemental material. Such a condition is also imposed in Shen (1997, Condition (C)), Chen and Shen (1998, Condition (B.3)), and Chen and Liao (2015, Assumption 3.3 (ii)). Assumption 6 concerns L^2 continuity and envelope conditions, which are needed for the applicability of the uniform law of large numbers, establishing stochastic equicontinuity and weak convergence, see Chen et al. (2008). Again, they are satisfied by widely used loss functions such as $\mathscr{L}(v) = v^2$, $\mathscr{L}(v) = v\{\tau - \mathbb{1}(v \leq 0)\}$, and $\mathscr{L}(v) = v^2 \cdot |\tau - \mathbb{1}(v \leq 0)|$ discussed in Section 3. Assumption 6 (3) implies $\sup_{x \in \mathscr{X}} \left| \mathscr{E}(x; \beta_d^*) \right| < C < \infty$ by Jensen's inequality.

The following theorem shows the asymptotic distribution of the proposed estimator $\widehat{\boldsymbol{\beta}}$, whose proof is presented in Appendix C and Supplement D.

Theorem 3. *Under Assumptions 1–6, for any $d \in \{0, 1, .., J\}$, we have $\widehat{\beta}_d \xrightarrow{p} \beta_d^*$ and*

$$\sqrt{n}\left(\widehat{\beta}_d - \beta_d^*\right) = H_d^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_d\left(Y_i, D_{di}, \boldsymbol{X}_i; \beta_d^*\right) + o_P(1), \tag{4.1}$$

where $H_d = -\partial_{\beta_d} \mathbb{E}[\mathscr{L}'(Y^*(d) - \beta_d^*)]$ and

$$S_d = S_d\left(Y_i, D_{di}, \boldsymbol{X}_i; \beta_d^*\right) := \frac{D_{di}}{\pi_d^*(\boldsymbol{X}_i)} \mathscr{L}'\left\{Y_i - \beta_d^*\right\} - \left\{\frac{D_{di} - \pi_d^*(\boldsymbol{X}_i)}{\pi_d^*(\boldsymbol{X}_i)}\right\} \mathscr{E}_d\left(\boldsymbol{X}_i, \beta_d^*\right).$$

Consequently,

$$\boldsymbol{V}^{-1/2} \cdot \sqrt{n}\{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\} \xrightarrow{d} \mathscr{N}\left(0, I_{(J+1)\times(J+1)}\right),$$

where $I_{(J+1)\times(J+1)}$ is the $(J+1) \times (J+1)$ identity matrix, $\boldsymbol{V} = \boldsymbol{H}^{-1} \mathbb{E}[\boldsymbol{S}\boldsymbol{S}^\top] \boldsymbol{H}^{-1}$, $\boldsymbol{H} = \mathrm{Diag}\{H_0, ..., H_J\}$ and $\boldsymbol{S} = (S_0, ..., S_J)^\top$. Based on the strict overlap condition and the integrability of the outcome, Assumption 4 (ii) and Assumption 6 (ii), we have that the asymptotic variance is finite, which implies that the proposed estimator $\widehat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent. In addition, when p is a fixed number, our estimator attains the semiparametric efficiency bound given in Ai et al. (2021).

4.2. Property of the ANN-OR estimator for ATE

The asymptotic normality of the ANN-IPW estimator requires the strict overlap condition, i.e. Assumption 4 (ii). In this section, we prove that such a condition can be possibly relaxed for the ANN-OR estimator of ATE defined in (3.7). From both the theoretical analysis and the numerical comparison in Section 7, we recommend the use of ANN-OR estimator for estimating ATE in practice and the use of ANN-IPW estimator for estimating other types of causal effects such as QTE.

Let $w_d^*(x) := f_X(x)/f_{X|D}(x|d)$, $z_d^*(x) = \mathbb{E}[Y|X=x, D=d]$, and $\overline{z}(z_d, \epsilon_n) := (1-\epsilon_n)\cdot z_d + \epsilon_n \cdot \{w_d^* + z_d^*\}$ be a local alternative value around $z_d \in \mathscr{G}_n$.

Assumption 7. For every $d \in \mathscr{D} = \{0,1,\dots,J\}$ and $m \geq 1$, we assume $z_d^*(\cdot) \in \mathscr{F}_p^m$.

Assumption 8. (i) We assume

$$\max\left\{ v_{z_d^*,m}\cdot r_n^{-\frac{1}{2}-\frac{1}{p}}, \sqrt{\frac{r_n\cdot p\cdot \log n}{n}} \right\} = o\left(n^{-\frac{1}{4}}\right).$$

(ii) The bound of the hidden-to-output weights, B_n, specified in (2.2) satisfies $B_n \leq 2v_{z_d^*,m}$.

Assumption 9. (i) $\mathbb{P}(D_{di}=1) \in (0,1)$ and $\pi_d^*(X) \in (0,1)$; (ii) There exists a constant $\overline{c}$ such that

$$\mathbb{E}\left[\left\{w_d^*(\boldsymbol{X})\right\}^2\right] = \mathbb{E}\left[\left\{\frac{f_X(\boldsymbol{X})}{f_{X|D}(\boldsymbol{X}|d)}\right\}^2\right] \Big\langle \overline{c} < \infty.$$

Assumption 10. (Approximation error) We assume the following conditions hold:

$$\sup_{\left\{z_d \in \mathscr{G}_n : \|z_d - z_d^*\|_{L^2(dF_X)} \leq \delta_n\right\}} \| \operatorname{Proj}_{\mathscr{G}_n}\overline{z}(z_d,\epsilon_n) - \overline{z}(z_d,\epsilon_n) \|_{L^2(dF_X)} = O\left(\frac{\epsilon_n^2}{\delta_n}\right), \textit{and}$$

$$\sup_{\left\{z_d \in \mathscr{G}_n : \|z_d - z_d^*\|_{L^2(dF_X)} \leq \delta_n\right\}} \frac{1}{n}\sum_{i=1}^n \left(\overline{z}(z_d,\epsilon_n)(\boldsymbol{X}_i) - \operatorname{Proj}_{\mathscr{G}_n}\overline{z}(z_d,\epsilon_n)(\boldsymbol{X}_i)\right) = O_P\left(\epsilon_n^2\right),$$

where $\operatorname{Proj}_{\mathscr{G}_n}\overline{z}(z_d,\epsilon_n)$ denotes the $L^2(dF_X)$-projection of $\overline{z}(z_d,\epsilon_n)$ on the ANN space $\mathscr{G}_n$.

Assumption 11. $\sup_{x\in\mathscr{X}} \mathbb{E}[\{Y^*(d)\}^2 \big| X = x] \langle C < \infty$ for some finite constant $C > 0$.

Assumptions 7 –11 are comparable to Assumptions 2–6. It's worth noting that Assumption 9 does not restrict the propensity score $\pi_d^*(X)$ to be uniformly bounded below by a constant.

Theorem 4. *Under Assumptions 1, 7–11, for every $d \in \{0,1,\dots,J\}$, we have $\widehat{\beta}_d^{OR} \xrightarrow{p} \beta_d^*$ and*

$$\sqrt{n}\left(\widehat{\beta}_d^{OR} - \beta_d^*\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n S_d^{OR}\left(Y_i, D_{di}, \boldsymbol{X}_i; \beta_d^*\right) + o_P(1),$$

where

$$S_d^{OR} = S_d^{OR}\left(Y_i, D_{di}, \boldsymbol{X}_i; \beta_d^*\right) = \frac{D_{di}}{\pi_d^*(\boldsymbol{X}_i)}Y_i - \left\{\frac{D_{di} - \pi_d^*(\boldsymbol{X}_i)}{\pi_d^*(\boldsymbol{X}_i)}\right\}\cdot z_d^*(\boldsymbol{X}_i) - \mathbb{E}\left[z_d^*(\boldsymbol{X}_i)\right].$$

Consequently,

$$\left\{\boldsymbol{V}^{OR}\right\}^{-1/2}\cdot\sqrt{n}\left\{\widehat{\boldsymbol{\beta}}^{OR} - \boldsymbol{\beta}^*\right\} \xrightarrow{d} \mathscr{N}\left(0, I_{(J+1)\times(J+1)}\right),$$

where $I_{(J+1)\times(J+1)}$ is the $(J+1)\times(J+1)$ identity matrix, $\boldsymbol{V}^{OR} = \mathbb{E}[\boldsymbol{S}^{OR}\cdot(\boldsymbol{S}^{OR})^\top]$, and $\boldsymbol{S}^{OR} = (S_0^{OR},\dots,S_J^{OR})^\top$. The proof of Theorem 4 is provided in Supplement E.

Note that $1/\pi_d^*(\boldsymbol{X}) = w_d^*(\boldsymbol{X})/\mathbb{P}(D_{di}=1)$, with Assumptions 9 and 11, we have that the asymptotic variance is finite, which implies the proposed estimator $\widehat{\boldsymbol{\beta}}^{OR}$ is $\sqrt{n}$-consistent. Moreover, the ANN-OR estimator $\widehat{\boldsymbol{\beta}}^{OR}$ has the same asymptotic variance as the ANN-IPW estimator $\widehat{\boldsymbol{\beta}}$ when $\mathscr{L}(v) = v^2$ for ATE. We can take the same inferential strategies as given in Section 5 to conduct inference based on the ANN-OR estimator.

5. Statistical inference

This section presents a weighted bootstrap procedure to conduct statistical inference for $\boldsymbol{\beta}^*$. Our TE estimator is obtained from directly optimizing an objective function, so a weighted bootstrap procedure can be performed to conduct inference without the need of estimating the asymptotic variance function. Estimation of the variance function can be challenging in the quantile TE setting. In Supplement F.2, we discuss a possible method for the estimation of the asymptotic variance based on the asymptotic formula given in Theorem 3.

Let $\{\omega_{d1}, \ldots, \omega_{dn}\}$ be *i.i.d.* positive random weights that are independent of data satisfying $\mathbb{E}[\omega_{di}] = 1$ and $Var(\omega_{di}) = 1$, where $d \in \{0, 1, \ldots, J\}$. The weighted bootstrap estimator of the inverse logistic PS g_d^* is defined by satisfying

$$L_{d,n}^B\left(\widehat{g}_d^B\right) \geq \sup_{g_d \in \mathscr{G}_n} L_{d,n}^B(g_d) - O\left(\epsilon_n^2\right),$$

where $L_{d,n}^B(g_d) := n^{-1}\sum_{i=1}^n \omega_{di}\ell_d(D_{di}, \boldsymbol{X}_i; g_d(\cdot))$ is the bootstrapped empirical criterion, and $\epsilon_n = o(n^{-1/2})$. Let $\widehat{\pi}_d^B := L(\widehat{g}_d^B)$. Then the weighted bootstrap IPW estimator of β_d^* is given by

$$\widehat{\beta}_d^B = \operatorname*{argmin}_{\beta \in \Theta} \frac{1}{n}\sum_{i=1}^n \frac{\omega_{di} D_{di}}{\widehat{\pi}_d^B(\boldsymbol{X}_i)} \mathscr{L}(Y_i - \beta),\ d \in \{0, 1, \ldots, J\}.$$

The weighted bootstrap OR estimator of ATE can be derived similarly. The weighted bootstrap estimator of the OR function z_d^* is defined by satisfying

$$L_{d,n}^{OR,B}\left(\widehat{z}_d^B\right) \geq \sup_{z_d \in \mathscr{G}_n} L_{d,n}^{OR,B}(z_d) - O\left(\epsilon_n^2\right),$$

where $L_{d,n}^{OR,B}(z_d) := n^{-1}\sum_{i=1}^n \omega_{di}\ell_d^{OR}(D_{di}, \boldsymbol{X}_i, Y_i; z_d(\cdot))$. Then the weighted bootstrap OR estimator of $\mathbb{E}[Y^*(d)]$ is given by

$$\widehat{\beta}_d^{OR,B} = \frac{1}{n}\sum_{i=1}^n \omega_{d,i}\widehat{z}_d^B(\boldsymbol{X}_i),\ d \in \{0, 1, \ldots, J\}.$$

Let $\widehat{\boldsymbol{\beta}}^B := \left(\widehat{\beta}_0^B, \ldots, \widehat{\beta}_{J+1}^B\right)^\top$ and $\widehat{\boldsymbol{\beta}}^{OR,B} := \left(\widehat{\beta}_0^{OR,B}, \ldots, \widehat{\beta}_{J+1}^{OR,B}\right)^\top$. The following theorem justifies the validation of the proposed bootstrap inference.

Theorem 5. *(i) Under Assumptions 1–6, for any $d \in \{0, 1, .., J\}$, then conditionally on the data we have*

$$\boldsymbol{V}^{-1/2} \cdot \sqrt{n}\left(\widehat{\boldsymbol{\beta}}^B - \widehat{\boldsymbol{\beta}}\right) \xrightarrow{d} \mathscr{N}\left(0, I_{(J+1)\times(J+1)}\right).$$

(ii) Under Assumptions 1, 3, 7-11, for any $d \in \{0, 1, .., J\}$, then conditionally on the data we have

$$\left\{\boldsymbol{V}^{OR}\right\}^{-1/2} \cdot \sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{OR,B} - \widehat{\boldsymbol{\beta}}^{OR}\right) \xrightarrow{d} \mathscr{N}\left(0, I_{(J+1)\times(J+1)}\right).$$

The proof of Theorem 5 is presented in Supplement F.

5.1. *Possible challenge of applying the EIF based method to quantile TE estimation*

The EIF can be applied to different loss functions. When EIF is given, the estimator of β_d^* can be obtained by solving the estimated efficient score function (Tsiatis, 2007). For example, when the loss function $\mathscr{L}(v) = v^2$ corresponding to ATE, the EIF of $\beta_d^* = \mathbb{E}[Y^*(d)]$ is

$$\frac{D_{di}}{\pi_d^*(\boldsymbol{X}_i)}Y_i - \left\{\frac{D_{di}}{\pi_d^*(\boldsymbol{X}_i)} - 1\right\}\mathbb{E}[Y_i|D_{di} = 1, \boldsymbol{X}_i] - \beta_d^*. \tag{5.1}$$

It involves the PS function $\pi_d^*(x)$ and the OR function $\mathbb{E}[Y_i|D_{di} = 1, \boldsymbol{X}_i = x]$ that can be estimated separately. As a result, the ATE of β_d^* can be obtained with the estimated PS and OR functions directly plug into the function given in (5.1).

When the loss function $\mathscr{L}(v) = v \cdot \{\tau - \mathbb{1}(v \leq 0)\}$ corresponding to the τth-quantile TE, the specific form of EIF for $\beta_d^* = F_{Y^*(d)}^{-1}(\tau)$ can also be derived from $H_d^{-1}S_d(Y_i, D_{di}, \boldsymbol{X}_i; \beta_d^*)$. As a result, its estimator can be obtained from solving the estimated efficient score equation

$$\sum_{i=1}^n \left[\frac{D_{di}}{\widehat{\pi}_d(\boldsymbol{X}_i)}\{\tau - \mathbb{1}(Y_i \leq \beta)\} - \left\{\frac{D_{di}}{\widehat{\pi}_d(\boldsymbol{X}_i)} - 1\right\}\{\tau - \widehat{\mathbb{E}}[\mathbb{1}(Y_i \leq \beta)|D_{di} = 1, \boldsymbol{X}_i]\}\right] = 0, \tag{5.2}$$

where $\widehat{\pi}_d(x)$ and $\widehat{\mathbb{E}}[\mathbb{1}(Y_i \leq \beta)|D_{di} = 1, \boldsymbol{X}_i = x]$ are estimates of $\pi_d^*(x)$ and $\mathbb{E}[\mathbb{1}(Y_i \leq \beta)|D_{di} = 1, \boldsymbol{X}_i = x]$, respectively. We can see that the estimation of quantile TEs from (5.2) is challenging when ANNs or other nonlinear machine learning methods are employed to obtain $\widehat{\mathbb{E}}[\mathbb{1}(Y_i \leq \beta)|D_{di} = 1, \boldsymbol{X}_i = x]$, as it intertwines with the unknown quantile TE parameter β nonlinearly.

Different from the aforementioned estimators constructed based on the estimated EIF, our TE estimators are directly obtained from optimizing an objective function that only involves the ANN-based estimated PS function. This approach greatly facilitates the computation of obtaining TE estimates and conducting causal inference without the need to estimate the EIF.

6. Extension to the general treatment effect on the treated

The above results can be easily extended to other multi-valued causal parameters defined on the treated subgroup. Let

$$\boldsymbol{\beta}^*_{d'} := \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{d=0}^{J} \mathbb{E}[\mathscr{L}(Y^*(d) - \beta_d) | D = d'], \tag{6.1}$$

for some fixed $d' \in \{0, 1, ..., J\}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)$ and $\boldsymbol{\beta}^*_{d'} = (\beta^*_{0,d'}, \beta^*_{1,d'}, ..., \beta^*_{J,d'})$. The formulation (6.1) includes the following important cases discussed in Lee (2018):

- $\mathscr{L}(v) = v^2$, then $\beta^*_{d,d'} = \mathbb{E}[Y^*(d)|D = d']$ is the average treatment effects on the treated.
- $\mathscr{L}(v) = v\{\tau - \mathbb{1}(v \leq 0)\}$, then $\beta^*_{d,d'} = F^{-1}_{Y^*(d)|D}(\tau | d')$ is the τth quantile of $Y^*(d)$ conditioned on the treated group $\{D = d'\}$.

Under Assumption 1, using the property of conditional expectation, the parameter of interest $\boldsymbol{\beta}^*_{d'}$ is identified by

$$\begin{aligned}
\boldsymbol{\beta}^*_{d'} :=\; & \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{d=0}^{J} \frac{1}{p_{d'}} \mathbb{E}[\mathbb{1}(D = d')\mathscr{L}(Y^*(d) - \beta_d)] \\
=\; & \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{d=0}^{J} \frac{1}{p_{d'}} \mathbb{E}\left[\pi^*_{d'}(\boldsymbol{X}) \cdot \mathbb{E}[\mathscr{L}(Y^*(d) - \beta_d)|\boldsymbol{X}]\right] \\
=\; & \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{d=0}^{J} \frac{1}{p_{d'}} \mathbb{E}\left[\pi^*_{d'}(\boldsymbol{X}) \cdot \mathbb{E}[\mathscr{L}(Y^*(d) - \beta_d)|\boldsymbol{X}] \cdot \mathbb{E}\left[\frac{\mathbb{1}(D = d)}{\pi^*_d(X)} | \boldsymbol{X}\right]\right] \\
=\; & \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{d=0}^{J} \frac{1}{p_{d'}} \cdot \mathbb{E}\left[\mathbb{1}(D = d) \cdot \frac{\pi^*_{d'}(\boldsymbol{X})}{\pi^*_d(\boldsymbol{X})} \cdot \mathscr{L}(Y - \beta_d)\right],
\end{aligned}$$

where $p_{d'} := \mathbb{P}(D = d')$. The estimator of $\boldsymbol{\beta}^*_{d'}$ is obtained by minimizing the empirical analogue of the above equation:

$$\widehat{\boldsymbol{\beta}}_{d'} = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{d=0}^{J} \frac{\sum_{i=1}^{n} D_{di} \widehat{\pi}_{d'}(\boldsymbol{X}_i) \mathscr{L}(Y_i - \beta_d) \Big/ \widehat{\pi}_d(\boldsymbol{X}_i)}{\sum_{i=1}^{n} D_{d'i}},$$

where $\widehat{\pi}_d$ is the ANN estimator of π^*_d. The estimator of $\beta^*_{d,d'}$ for $d \in \{0, 1, ..., J\}$ can be defined as

$$\widehat{\beta}_{d,d'} = \operatorname*{argmin}_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \frac{D_{di}}{\widehat{\pi}_d(\boldsymbol{X}_i)} \widehat{\pi}_{d'}(\boldsymbol{X}_i) \mathscr{L}(Y_i - \beta).$$

Similar to the proof of Theorem 3 we obtain the following result for $\boldsymbol{\beta}^*_{d'}$.

Theorem 6. *Under Assumptions 1–6, for any $d, d' \in \{0, 1, .., J\}$, we have that*

$$\sqrt{n}\left(\widehat{\beta}_{d,d'} - \beta^*_{d,d'}\right) = H^{-1}_{d,d'} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_{d,d'}\left(\boldsymbol{X}_i, D_{di}, Y_i; \beta^*_{d,d'}\right) + o_P(1),$$

where $H_{d,d'} = -\partial_{\beta_d} \mathbb{E}[\pi^*_{d'}(\boldsymbol{X}_i) \mathscr{L}'(Y^*_i(d) - \beta^*_d)]$ and

$$S_{d,d'}\left(Y_i, D_{di}, \boldsymbol{X}_i; \beta^*_{d,d'}\right) := \frac{D_{di}}{\pi^*_d(\boldsymbol{X}_i)} \pi^*_{d'}(\boldsymbol{X}_i) \mathscr{L}'\left(Y_i - \beta^*_{d,d'}\right) - \left\{\frac{D_{di}}{\pi^*_d(\boldsymbol{X}_i)} \pi^*_{d'}(\boldsymbol{X}_i) - D_{d'i}\right\} \mathscr{E}_{d'}\left(\boldsymbol{X}_i, \beta^*_{d,d'}\right).$$

7. Simulation studies

7.1. Background and methods used

In this section, we illustrate the finite sample performance of our proposed methods via simulations in which we generate data from models in Section 7.2. Our proposed IPW estimator can be applied to various types of treatment effects. We use ATE, ATT (average treatment effects on the treated), QTE and QTT (quantile treatment effects on the treated) for illustration of the performance of the IPW estimator. For QTE and QTT, we consider the 25th (Q1), 50th (Q2) and 75th (Q3) percentiles. We also illustrate the performance of the OR estimator for ATE and ATT. To obtain the IPW and OR estimators, we estimate the PS and OR functions by using our proposed ANN method as well as five other popular methods, including the generalized linear models (GLM), the generalized additive models (GAM),

Table 1
The summary statistics of the estimated ATEs for Model 1 with $p = 5$.

	IPW						OR						DR	Oracle
	ANN	GLM	GAM	RF	GBM	DNN	ANN	GLM	GAM	RF	GBM	DNN		
$n = 1000$														
bias	**0.0008**	0.0584	0.0594	0.0523	0.0562	0.0015	**0.0010**	0.0586	0.0574	0.0151	0.0144	0.0012	0.0026	0.0006
emp_sd	**0.0758**	0.0907	0.0924	0.0830	0.0844	0.0789	**0.0713**	0.0906	0.0930	0.0769	0.0828	0.0726	0.0752	0.0668
est_sd	**0.0701**	0.0923	0.0903	0.0511	0.0483	0.0715	**0.0701**	0.0923	0.0903	0.0511	0.0483	0.0715	0.0701	0.0686
cover_rate	**0.9275**	0.9025	0.8750	0.6700	0.6400	0.9250	**0.9425**	0.9050	0.8900	0.8000	0.7450	0.9425	0.9325	0.9600
est_sd_boot	**0.0771**	0.0901	0.0922	0.0821	0.0853	0.0791	**0.0751**	0.0902	0.0924	0.0773	0.0841	0.0737		
cover_rate_boot	**0.9375**	0.8850	0.0888	0.9025	0.8950	0.9375	**0.9475**	0.8875	0.8875	0.9275	0.9350	0.9500		
$n = 2000$														
bias	**0.0007**	0.0578	0.0585	0.0495	0.0425	0.0008	**0.0009**	0.0591	0.0571	0.0127	0.0131	0.0010	0.0007	0.0010
emp_sd	**0.0524**	0.0684	0.0675	0.0620	0.0634	0.0563	**0.0499**	0.0685	0.0677	0.0564	0.0575	0.0502	0.0523	0.0498
est_sd	**0.0488**	0.0652	0.0637	0.0371	0.0370	0.0508	**0.0488**	0.0652	0.0637	0.0371	0.0370	0.0508	0.0493	0.0485
cover_rate	**0.9375**	0.8175	0.8100	0.6250	0.6500	0.9175	**0.9500**	0.8175	0.8050	0.7875	0.7575	0.9475	0.9475	0.9475
est_sd_boot	**0.0573**	0.0679	0.0665	0.0612	0.0645	0.0599	**0.0493**	0.0685	0.0697	0.0563	0.0572	0.0505		
cover_rate_boot	**0.9475**	0.8275	0.8400	0.8400	0.8575	0.9400	**0.9500**	0.8275	0.8225	0.9200	0.9075	0.9500		
$n = 5000$														
bias	**0.0001**	0.0558	0.0545	0.0407	0.0225	0.0002	**0.0003**	0.0558	0.0545	0.0043	0.0073	0.0003	0.0009	0.0010
emp_sd	**0.0334**	0.0397	0.0390	0.0362	0.0376	0.0335	**0.0312**	0.0397	0.0389	0.0324	0.0327	0.0305	0.0322	0.0309
est_sd	**0.0309**	0.0413	0.0403	0.0244	0.0258	0.0307	**0.0309**	0.0413	0.0403	0.0244	0.0258	0.0307	0.0306	0.0307
cover_rate	**0.9350**	0.7350	0.7175	0.5700	0.7400	0.9250	**0.9475**	0.7375	0.7150	0.8500	0.8450	0.9525	0.9400	0.9600
est_sd_boot	**0.0331**	0.0402	0.0401	0.0355	0.0365	0.0337	**0.0308**	0.0399	0.0402	0.0327	0.0331	0.0305		
cover_rate_boot	**0.9475**	0.7350	0.7225	0.7850	0.8725	0.9475	**0.9575**	0.7350	0.7250	0.9325	0.9250	0.9525		

Table 2
The summary statistics of the estimated ATEs for Model 1 with $p = 10$.

	IPW						OR						DR	Oracle
	ANN	GLM	GAM	RF	GBM	DNN	ANN	GLM	GAM	RF	GBM	DNN		
$n = 1000$														
bias	**0.0010**	0.0635	0.0655	0.0574	0.0536	0.0010	**0.0064**	0.0637	0.0640	0.0264	0.0279	0.0037	0. 0073	0.0077
emp_sd	**0.0855**	0.0900	0.1127	0.0887	0.0915	0.0859	**0.0793**	0.0901	0.1041	0.0825	0.0832	0.0785	0.0823	0.0666
est_sd	**0.0789**	0.0943	0.0982	0.0547	0.0512	0.0784	**0.0789**	0.0943	0.0982	0.0547	0.0512	0.0784	0.0789	0.0695
cover_rate	**0.9250**	0.9075	0.8675	0.6725	0.6500	0.9275	**0.9400**	0.9075	0.8800	0.7775	0.7375	0.9450	0.9325	0.9550
est_sd_boot	**0.0884**	0.0923	0.1089	0.0893	0.0921	0.0864	**0.0798**	0.0913	0.0965	0.0833	0.0856	0.0778		
cover_rate_boot	**0.9350**	0.8950	0.8825	0.8925	0.8950	0.9375	**0.9475**	0.8950	0.8650	0.9150	0.9125	0.9450		
$n = 2000$														
bias	**0.0034**	0.0659	0.0661	0.0584	0.0462	0.0032	**0.0001**	0.0660	0.0675	0.0227	0.0266	0.0009	0.0012	0.0050
emp_sd	**0.0597**	0.0660	0.0692	0.0634	0.0629	0.0657	**0.0543**	0.0661	0.0683	0.0588	0.0590	0.0519	0.0558	0.0501
est_sd	**0.0532**	0.0668	0.0667	0.0397	0.0382	0.0539	**0.0532**	0.0668	0.0667	0.0397	0.0382	0.0539	0.0530	0.0492
cover_rate	**0.9250**	0.8275	0.8225	0.6050	0.6400	0.9250	**0.9400**	0.8250	0.8275	0.7825	0.7475	0.9650	0.9400	0.9500
est_sd_boot	**0.0604**	0.0668	0.0687	0.0624	0.0628	0.0649	**0.0560**	0.0664	0.0679	0.0590	0.0595	0.0530		
cover_rate_boot	**0.9475**	0.8250	0.8250	0.8125	0.8375	0.9425	**0.9500**	0.8275	0.8350	0.9025	0.8875	0.9525		
$n = 5000$														
bias	**0.0015**	0.0679	0.0670	0.0565	0.0362	0.0012	**0.0001**	0.0679	0.0670	0.0158	0.0177	0.0009	0. 0017	0.0020
emp_sd	**0.0363**	0.0439	0.0439	0.0416	0.0419	0.0357	**0.0332**	0.0440	0.0439	0.0376	0.0376	0.0319	0.0345	0.0319
est_sd	**0.0322**	0.0422	0.0418	0.0258	0.0267	0.0319	**0.0322**	0.0422	0.0418	0.0258	0.0267	0.0319	0.0323	0.0311
cover_rate	**0.9275**	0.6100	0.5975	0.4225	0.6150	0.9250	**0.9425**	0.6125	0.6000	0.7875	0.7700	0.9500	0.9475	0.9550
est_sd_boot	**0.0368**	0.0441	0.0443	0.0420	0.0423	0.0349	**0.0333**	0.0440	0.0442	0.0376	0.0378	0.0323		
cover_rate_boot	**0.9475**	0.6550	0.6475	0.6575	0.7875	0.9450	**0.9525**	0.6250	0.6175	0.8275	0.8150	0.9525		

the random forests (RF), the gradient boosted machines (GBM) and the deep neural networks with three hidden layers (DNN). We make a comparison of the performance of the resulting TE estimators with the nuisance functions estimated by the aforementioned six methods. Moreover, we compare our IPW and OR estimators with the doubly robust (DR) estimator (Farrell et al., 2021) and the Oracle estimator for ATE. For the DR estimator, the IPW and OR functions are also approximated by ANNs. The Oracle estimator is constructed based on the efficient influence function with the true PS and OR functions plugged in, see Hahn (1998). The Oracle estimators are infeasible in practice, but they are expected to perform the best for the estimation of ATE, and serve as a benchmark to compare with. In the quantile TE settings, both DR (EIF-based) and OR estimators are difficult to obtain, so we only show the performance of the IPW estimator.

We use the Rectified Linear Unit (ReLU) as the activation function for both ANN and DNN. We use cubic regression spline basis functions for GAM. We apply grid search with 5-fold cross-validation to select hyperparameters for all methods, including the number of neurons for ANN DNN, the number of trees and max depths of trees for RF and GBM, and the learning rate for GBM. All the simulation studies are implemented in Python 3.9. The DNN, GLM, GAM, RF and GBM methods are implemented using the packages tensorflow, statsmodel, pyGAM and scikit-learn, respectively.

7.2. *Data generating process*

We generate the treatment and outcome variables from a nonlinear model and a linear model, respectively, given as follows.
Model 1 (nonlinear model) :

$$\begin{aligned}
\text{logit}\{\mathbb{E}(D_i|\boldsymbol{X}_i)\} &= 0.5\big(X^*_{i1}X^*_{i2} - 0.7sin\big(\big(X^*_{i3}+X^*_{i4}\big)\big(X^*_{i5}-0.2\big)\big) - 0.1\big),\\
\mathbb{E}\big(Y^*_i(1)\big|\boldsymbol{X}_i\big) &= \mathbb{E}(Y_i|\boldsymbol{X}_i, D_i=1) = 0.3\big(X^*_{i1}-0.9\big)^2 + 0.1\big(X^*_{i2}-0.5\big)^2\\
&\quad -0.6X^*_{i2}X^*_{i3} + sin\big(-1.7\big(X^*_{i1}+X^*_{i3}-1.1\big) + X^*_{i4}X^*_{i5}\big) + 1,\\
\mathbb{E}\big(Y^*_i(0)\big|\boldsymbol{X}_i\big) &= \mathbb{E}(Y_i|\boldsymbol{X}_i, D_i=0) = 0.64\big(X^*_{i1}-0.9\big)^2 + 0.16\big(X^*_{i2}+0.2\big)^2\\
&\quad -0.6X^*_{i2}X^*_{i3} + sin\big(-1.7\big(X^*_{i1}+X^*_{i3}-1.1\big) + X^*_{i4}X^*_{i5}\big) - 1;
\end{aligned}$$

Model 2 (linear model) :

$$\begin{aligned}
\text{logit}\{\mathbb{E}(D_i|\boldsymbol{X}_i)\} &= 0.1\big(X^*_{i1}+X^*_{i2}-2X^*_{i3}+3X^*_{i4}-3X^*_{i5}\big),\\
\mathbb{E}\big(Y^*_i(1)\big|\boldsymbol{X}_i\big) &= \mathbb{E}(Y_i|\boldsymbol{X}_i, D_i=1) = 4X^*_{i1}+3X^*_{i2}-X^*_{i3}-5X^*_{i4}+7X^*_{i5}+1,\\
\mathbb{E}\big(Y^*_i(0)\big|\boldsymbol{X}_i\big) &= \mathbb{E}(Y_i|\boldsymbol{X}_i, D_i=0) = 4X^*_{i1}+3X^*_{i2}-X^*_{i3}-5X^*_{i4}+7X^*_{i5}-1,
\end{aligned}$$

where $X^*_{ij'} = c_p\frac{5}{p}\sum_{j=p(j'-1)/5+1}^{pj'/5}X_{ij}$ for $1 \le j' \le 5, 1 \le i \le n$, and $Y^*_i(d) = \mathbb{E}(Y^*_i(d)\big|\boldsymbol{X}_i) + \epsilon_i$, $d = \{0,1\}$, $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ for $1 \le i \le n$.

We generate the confounders from $X_{ij} = 2(F(Z_{ij}) - 0.5)$, where $Z_i = (Z_{i1},\ldots,Z_{ip})^\top \overset{i.i.d.}{\sim} \mathcal{N}(0,\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \{\sigma_{kk'}\}$, $\sigma_{kk'} = 0.2^{|k-k'|}$ for $1 \le k$, $k' \le p$, and $F(\cdot)$ is the cumulative distribution function of the standard normal for $1 \le i \le n, 1 \le j \le p$. Let c_p=1. We partition the confounders into 5 subgroups, and $X^*_{ij'}$ is the average of the $p/5$ confounders in the j'th subgroup for $j' = 1, \ldots, 5$, so that every confounder is included in our models. We consider $p = 5$, 10 and $n = 1000$, 2000, 5000. All simulation results are based on 400 realizations.

We also use the nonlinear model (Model 1) to illustrate the performance of our proposed methods for $p = 100$ and $n = 2000$, with the confounders $X_{ij} = 2(F(Z_{ij}/\sigma_j) - 0.5)$, where σ_j is the standard deviation of Z_{ij}, and $Z_i = (Z_{i1},\ldots,Z_{ip})^\top$ are generated from two designs.

- Design 1 (factor model): $Z_{ij} = F_i^\top L_j + \eta_{ij}$, where $F_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\boldsymbol{\Sigma}^*)$, $\boldsymbol{\Sigma}^* = \{\sigma_{kk'}\}$, in which $\sigma_{kk'} = 0.2^{|k-k'|}$ for $1 \le k, k' \le 10$, L_j is a constant vector kept fixed for each realization and is generated from $L_j \overset{i.i.d.}{\sim} \mathcal{N}(0,\boldsymbol{\Sigma}^*)$, and $\eta_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Let c_p=7.
- Design 2 (multivariate normal): $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \{\sigma_{kk'}\}$, $\sigma_{kk'} = 0.2^{m(|k-k'|)}$ for $1 \le k, k' \le p$, where $m(x) = \lceil x/10 \rceil$, and $\lceil a \rceil$ denotes the smallest integer no less than a. Let c_p=4.

7.3. *Simulation results*

To compare the performance of different methods for estimating the TEs, we report the following statistics: the absolute value of bias (bias), the empirical standard deviation (emp_sd), the average value of the estimated standard deviations based on the asymptotic formula (est_sd) and obtained from the weighted bootstrapping (est_sd_boot), and the empirical coverage rates of the 95% confidence intervals based on the estimated asymptotic standard deviations (cover_rate) and the weighted bootstrapping method (cover_rate_boot). The 95% confidence intervals based on bootstrapping are obtained from the 2.5th percentile and 97.5th percentile of the weighted bootstrapping estimates. The bootstrap confidence intervals and the estimated standard deviations (est_sd_boot) are obtained based on 400 bootstrap replicates for each simulation sample. The bootstrap weights are randomly generated from the exponential distribution with mean 1 according to Ma and Kosorok (2005).

Tables 1–2 report the numerical results for different estimators of ATE for Model 1 with $p = 5$, 10, respectively. We see that as n increases, the empirical coverage rates (cover_rate and cover_rate_boot) based on our proposed ANN-based IPW and OR estimates

Table 3
The summary statistics of the estimated QTEs by the IPW method for Model 1 with p=5.

	Q1						Q2						Q3					
	ANN	GLM	GAM	RF	GBM	DNN	ANN	GLM	GAM	RF	GBM	DNN	ANN	GLM	GAM	RF	GBM	DNN
n=1000																		
bias	**0.0045**	0.0601	0.0589	0.0630	0.0684	0.0036	**0.0012**	0.0534	0.0574	0.0507	0.0539	0.0025	**0.0098**	0.0363	0.0407	0.0286	0.0302	0.0067
emp_sd	**0.1266**	0.1358	0.1418	0.1274	0.1301	0.1235	**0.1075**	0.1116	0.1166	0.1055	0.1074	0.1099	**0.1196**	0.1151	0.1201	0.1121	0.1130	0.1206
est_sd	**0.1414**	0.1403	0.1596	0.1398	0.1405	0.1404	**0.1238**	0.1240	0.1447	0.1241	0.1249	0.1222	**0.1251**	0.1232	0.1412	0.1237	0.1247	0.1243
cover_rate	**0.9500**	0.9275	0.9375	0.9350	0.9250	0.9525	**0.9675**	0.9550	0.9500	0.9675	0.9650	0.9575	**0.9600**	0.9650	0.9650	0.9650	0.9625	0.9625
est_sd_boot	**0.1259**	0.1378	0.1432	0.1275	0.1305	0.1278	**0.1091**	0.1154	0.1171	0.1087	0.1072	0.1093	**0.1231**	0.1131	0.1241	0.1172	0.1187	0.1236
cover_rate_boot	**0.9350**	0.9225	0.9300	0.9275	0.9175	0.9375	**0.9600**	0.9500	0.9475	0.9500	0.9475	0.9500	**0.9575**	0.9450	0.9500	0.9575	0.9550	0.9600
$n\ = 2000$																		
bias	**0.0046**	0.0604	0.0560	0.0631	0.0646	0.0037	**0.0004**	0.0547	0.0523	0.0454	0.0467	0.0014	**0.0056**	0.0494	0.0506	0.0372	0.0369	0.0043
emp_sd	**0.0928**	0.1029	0.1012	0.0944	0.0944	0.0954	**0.0819**	0.0883	0.0884	0.0828	0.0844	0.0831	**0.0852**	0.0872	0.0888	0.0848	0.0860	0.0864
est_sd	**0.0900**	0.0986	0.0976	0.0984	0.0989	0.0933	**0.0791**	0.0870	0.0872	0.0871	0.0875	0.0804	**0.0814**	0.0860	0.0871	0.0862	0.0867	0.0824
cover_rate	**0.9325**	0.8650	0.8750	0.9000	0.8875	0.9425	**0.9350**	0.9050	0.9050	0.9300	0.9225	0.9325	**0.9550**	0.9100	0.9175	0.9425	0.9350	0.9375
est_sd_boot	**0.0911**	0.1014	0.1023	0.0954	0.0968	0.0961	**0.0811**	0.0889	0.0892	0.0857	0.0853	0.0842	**0.0841**	0.0872	0.0891	0.0857	0.0871	0.0858
cover_rate_boot	**0.9400**	0.8675	0.8925	0.8750	0.8650	0.9450	**0.9425**	0.9100	0.9025	0.9125	0.9200	0.9500	**0.9375**	0.9150	0.9200	0.9375	0.9350	0.9450
$n\ = 5000$																		
bias	**0.0034**	0.0594	0.0560	0.0456	0.0255	0.0029	**0.0033**	0.0489	0.0478	0.0367	0.0188	0.0036	**0.0026**	0.0454	0.0460	0.0320	0.0176	0.0014
emp_sd	**0.0560**	0.0600	0.0591	0.0555	0.0559	0.0570	**0.0514**	0.0545	0.0539	0.0520	0.0522	0.0517	**0.0494**	0.0513	0.0508	0.0497	0.0503	0.0490
est_sd	**0.0558**	0.0624	0.0607	0.0624	0.0639	0.0531	**0.0493**	0.0546	0.0537	0.0546	0.0557	0.0510	**0.0508**	0.0542	0.0540	0.0542	0.0551	0.0513
cover_rate	**0.9575**	0.8625	0.8600	0.9125	0.9575	0.9575	**0.9400**	0.8675	0.8675	0.9200	0.9525	0.9475	**0.9625**	0.8900	0.8850	0.9300	0.9575	0.9675
est_sd_boot	**0.0560**	0.0614	0.0603	0.0576	0.0578	0.0560	**0.0509**	0.0553	0.5420	0.0534	0.0539	0.0509	**0.5010**	0.0517	0.0515	0.0508	0.0512	0.5010
cover_rate_boot	**0.9575**	0.8575	0.8550	0.8825	0.9325	0.9575	**0.9500**	0.8725	0.8725	0.9075	0.9500	0.9500	**0.9600**	0.8575	0.8550	0.9200	0.9325	0.9625

Table 4
The summary statistics of the estimated QTEs by the IPW method for Model 1 with p=10.

	Q1						Q2						Q3					
	ANN	GLM	GAM	RF	GBM	DNN	ANN	GLM	GAM	RF	GBM	DNN	ANN	GLM	GAM	RF	GBM	DNN
n=1000																		
bias	**0.0019**	0.0717	0.0750	0.0700	0.0696	0.0023	**0.0047**	0.0544	0.0541	0.0507	0.0472	0.0033	**0.0030**	0.0488	0.0476	0.0394	0.0367	0.0036
emp_sd	**0.1585**	0.1345	0.1634	0.1325	0.1364	0.1592	**0.1437**	0.1195	0.1485	0.1179	0.1206	0.1437	**0.1332**	0.1185	0.1529	0.1193	0.1208	0.1331
est_sd	**0.2530**	0.1444	0.4770	0.1442	0.1487	0.2630	**0.2639**	0.1269	0.4708	0.1269	0.1312	0.2684	**0.2034**	0.1253	0.4458	0.1255	0.1294	0.2109
cover_rate	**0.9675**	0.9200	0.9950	0.9350	0.9375	0.9675	**0.9425**	0.9450	1.0000	0.9475	0.9525	0.9450	**0.9575**	0.9325	1.0000	0.9350	0.9500	0.9575
est_sd_boot	**0.1623**	0.1345	0.1552	0.1376	0.1425	0.1687	**0.1523**	0.1231	0.1253	0.1198	0.1225	0.1523	**0.1376**	0.1203	0.1623	0.1198	0.1256	0.1392
cover_rate_boot	**0.9325**	0.9150	0.9125	0.9300	0.9250	0.9350	**0.9350**	0.9350	0.9375	0.9400	0.9475	0.9500	**0.9325**	0.9225	0.9375	0.9000	0.9225	0.9350
n = 2000																		
bias	**0.0030**	0.0663	0.0678	0.0632	0.0535	0.0023	**0.0015**	0.0576	0.0580	0.0529	0.0421	0.0005	**0.0105**	0.0613	0.0620	0.0527	0.0426	0.0087
emp_sd	**0.1070**	0.1040	0.1083	0.1008	0.1016	0.1109	**0.0937**	0.0902	0.0943	0.0874	0.0872	0.0954	**0.0873**	0.0855	0.0890	0.0833	0.0830	0.0923
est_sd	**0.1057**	0.1021	0.1149	0.1019	0.1052	0.1134	**0.0919**	0.0889	0.1026	0.0889	0.0917	0.0939	**0.0913**	0.0878	0.1001	0.0878	0.0904	0.0919
cover_rate	**0.9425**	0.8775	0.8850	0.8900	0.9150	0.9475	**0.9200**	0.8900	0.9175	0.9100	0.9250	0.9350	**0.9425**	0.9125	0.9325	0.9275	0.9400	0.9525
est_sd_boot	**0.1072**	0.1043	0.1097	0.1011	0.1045	0.1122	**0.0932**	0.0901	0.0994	0.0883	0.0892	0.0952	**0.0885**	0.0861	0.0923	0.0869	0.0873	0.0925
cover_rate_boot	**0.9450**	0.8800	0.8000	0.8900	0.9125	0.9450	**0.9325**	0.8975	0.9150	0.9100	0.9175	0.9450	**0.9375**	0.9050	0.9125	0.9250	0.9275	0.9575
n = 5000																		
bias	**0.0065**	0.0774	0.0764	0.0696	0.0472	0.0055	**0.0046**	0.0637	0.0633	0.0547	0.0356	0.0051	**0.0021**	0.0537	0.0535	0.0433	0.0273	0.0031
emp_sd	**0.0657**	0.0682	0.0680	0.0650	0.0656	0.0634	**0.0568**	0.0580	0.0584	0.0560	0.0567	0.0532	**0.0554**	0.0555	0.0562	0.0549	0.0563	0.0531
est_sd	**0.0629**	0.0644	0.0638	0.0644	0.0659	0.0615	**0.0537**	0.0556	0.0555	0.0556	0.0568	0.0531	**0.0542**	0.0550	0.0551	0.0549	0.0560	0.0522
cover_rate	**0.9275**	0.7675	0.7725	0.8025	0.8875	0.9275	**0.9350**	0.7800	0.7725	0.8225	0.9075	0.9450	**0.9375**	0.8375	0.8300	0.8650	0.9250	0.9375
est_sd_boot	**0.0655**	0.0686	0.0678	0.0648	0.0658	0.0635	**0.0565**	0.0577	0.0579	0.0552	0.0571	0.0545	**0.0548**	0.0561	0.0568	0.0552	0.0561	0.0528
cover_rate_boot	**0.9325**	0.7825	0.7875	0.8125	0.8875	0.9450	**0.9475**	0.7900	0.7850	0.8200	0.9100	0.9525	**0.9400**	0.8500	0.8525	0.8675	0.9275	0.9425

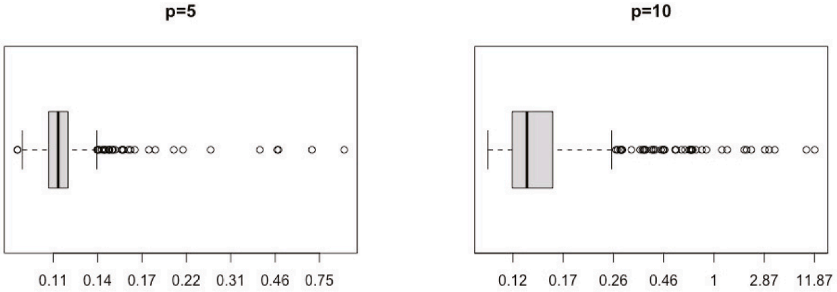

Fig. 1. Boxplots of the estimated asymptotic standard deviations of QTE(Q1) for Model 1, $n = 1000$.

become closer to the nominal level 95%. The biases are close to zero, and the values of emp_sd, est_sd and est_sd_boot decrease as n increases. These results corroborate our asymptotic theories. We observe that our ANN-based IPW and OR estimators have comparable performance to the DR and the Oracle estimators when estimating ATE. The proposed ANN-based OR estimator slightly outperforms the ANN-based IPW and DR estimators in the sense that it has the smallest emp_sd value. It is possible that the estimated PS functions have a few values close to zero. This can affect the emp_sd value of the IPW estimate for ATE. The DR estimator which is constructed based on the estimates of both IPW and OR functions has larger emp_sd values than the OR estimator, but it yields smaller emp_sd values than the IPW estimator. Our numerical results suggest that the proposed ANN-based OR estimator is preferred for the estimation of ATE. However, in practice, it can be difficult to construct OR and DR estimators for other types of TEs, such as quantile TEs. Then the proposed ANN-based IPW estimator becomes a more appealing tool. Moreover, our numerical results given in Tables 3–4 show that the performance of the ANN-based IPW estimators for quantile TEs is less influenced by the small values of the estimated PS functions because of the robustness of the quantile objective functions. For our proposed ANN-based IPW and OR estimators, it is convenient to apply the proposed weighted bootstrap procedure for conducting inference. We find that the empirical coverage rates of 95% confidence intervals obtained from the weighted bootstrapping are closer to the nominal level than those obtained from the estimated asymptotic standard deviations.

Next, we compare the performance of different machine learning (ML) methods for the estimation of ATE. We see that the GLM and GAM methods yield large estimation biases due to the model misspecification problem. Our numerical results show that the proposed ANN method outperforms the other two ML methods, RF and GBM, for the estimation of TEs. The empirical coverage rates based on the ANN method are closer to the nominal level in all cases than the rates obtained from RF and GBM. It is worth noting that our ANN-based TE estimators enjoy the properties of root-n consistency and semiparametric efficiency. In general, our numerical results corroborate those theoretical properties. Moreover, for RF and GBM, the OR estimator also performs better than the IPW estimator for ATE estimation. The empirical coverage rates of the 95% confidence intervals obtained from the weighted bootstrapping are improved compared to the rates obtained from the estimated asymptotic standard deviation. The DNN method has comparable performance to the ANN method.

Tables 3–4 show the numerical results of different methods for the estimation of QTEs for Model 1 with $p = 5$, 10, respectively. It is difficult to construct OR and DR estimators for QTEs, so we only report the results for the IPW estimators, which are very convenient to be obtained in this context. The PS functions are estimated by different ML methods, and the numerical results of the resulting IPW estimates are summarized in Tables 3–4. In general, we observe similar patterns of numerical performance of different methods as shown in Tables 1–2. It is worth noting that the proposed ANN-based IPW method has very stable performance for the estimation of QTEs. The resulting emp_sd values are not influenced by possibly small values of the estimated PS functions because of the robustness nature of the quantile objective function. Moreover, in the QTE settings, estimation of the asymptotic standard deviations can involve a complicated procedure, and several approximations are needed. As a result, the estimation is not guaranteed to perform well. Fig. 1 shows the boxplots of the estimated asymptotic standard deviations of QTE (Q1) for Model 1 with $p = 5, 10$, $n = 1000$. We see that the estimated values are large for some simulation replicates. In contrast, the estimated standard deviations obtained from the weighted bootstrapping have more reliable performance. In complex TE settings such as QTEs, the proposed weighted bootstrap method that avoids the estimation of the asymptotic variance provides a robust way to conduct statistical inference, and thus it is recommended in practice. It is convenient to apply the weighted bootstrap method in our proposed TE estimation procedure, as the TE estimators are obtained from optimizing a general objective function. We apply different ML methods to estimate the PS function. The numerical results show that the ANN and DNN methods have comparable performance, and they still outperform other methods for the estimation and inference of QTEs.

To save space, the numerical results of different methods for ATTs and QTTs for Model 1 and all the numerical results for Model 2 are presented in Tables 1–12 of Section I in the Supplementary Materials. Tables 1–4 show that the numerical results of different methods for estimating ATTs and QTTs have similar patterns as those given in Tables 1–4 for ATEs and QTEs. In Model 2, both PS and OR functions are generated from linear models, so the GLM and GAM methods no longer have the model misspecification problem, and GLM is expected to have the best performance. However, we can see from Tables 5–12 that the ANN and DNN methods have comparable performance to GLM for the estimation of TEs in all cases. It is worth noting that our proposed method can also be applied to the estimation of asymmetric least squares TEs and other types of TEs, and it has similar patterns of numerical performance as the estimation of ATEs and QTEs. The numerical results are not presented due to space limitations.

Table 5
The summary statistics of the estimated ATEs and ATTs for Model 1 with $p = 100$ and $n = 2000$.

		Design 1				Design 2			
		ATE		ATT		ATE		ATT	
		ANN	ANN-PCA	ANN	ANN-PCA	ANN	ANN-PCA	ANN	ANN-PCA
IPW	bias	0.0104	0.0444	0.0205	0.0516	0.0163	0.0446	0.0197	0.0578
	emp_sd	0.0983	0.0719	0.1293	0.0790	0.1080	0.0776	0.1673	0.1005
	est_sd	0.0512	0.0651	0.0764	0.0789	0.0618	0.0733	0.0782	0.0865
	cover_rate	0.8225	0.8650	0.8025	0.8875	0.8100	0.8800	0.7650	0.8500
	est_sd_boot	0.0694	0.0668	0.0899	0.0796	0.0798	0.0737	0.1036	0.0934
	cover_rate_boot	0.8650	0.8675	0.8350	0.8900	0.8550	0.8925	0.8100	0.8825
OR	bias	0.0097	0.0379	0.0132	0.0356	0.0222	0.0388	0.0193	0.0511
	emp_sd	0.0892	0.0722	0.1163	0.0870	0.0988	0.0760	0.1356	0.0961
	est_sd	0.0512	0.0651	0.0764	0.0789	0.0618	0.0733	0.0782	0.0865
	cover_rate	0.8350	0.8900	0.8475	0.8950	0.7975	0.9150	0.8150	0.8400
	est_sd_boot	0.0625	0.0639	0.0871	0.0847	0.0714	0.0698	0.0939	0.0942
	cover_rate_boot	0.8775	0.8850	0.8575	0.9150	0.8450	0.8950	0.8575	0.8850
DR	bias	0.0101	0.0382	0.0158	0.0457	0.0195	0.0390	0.0204	0.0548
	emp_sd	0.0894	0.0723	0.1207	0.0862	0.0992	0.0765	0.1397	0.0964
	est_sd	0.0521	0.0669	0.0785	0.0807	0.0633	0.0724	0.0801	0.0849
	cover_rate	0.8375	0.8850	0.8500	0.9050	0.8025	0.8975	0.8275	0.8375

Table 6
The summary statistics of the estimated QTEs and QTTs by the IPW method for Model 1 with $p = 100$ and $n = 2000$.

		Design 1				Design 2			
		QTE		QTT		QTE		QTT	
		ANN	ANN-PCA	ANN	ANN-PCA	ANN	ANN-PCA	ANN	ANN-PCA
Q1	bias	0.0157	0.0259	0.0139	0.0313	0.0132	0.0337	0.0111	0.0418
	emp_sd	0.1379	0.1101	0.1917	0.1257	0.1988	0.1210	0.2551	0.1542
	est_sd_boot	0.1305	0.1004	0.1812	0.1234	0.1753	0.1190	0.2236	0.1497
	cover_rate_boot	0.9050	0.8850	0.9125	0.8875	0.9125	0.8775	0.9050	0.8575
Q2	bias	0.0131	0.0283	0.0249	0.0302	0.0154	0.0326	0.0055	0.0386
	emp_sd	0.1256	0.1017	0.1687	0.1123	0.1680	0.1101	0.2056	0.1315
	est_sd_boot	0.1104	0.0977	0.1447	0.1026	0.1544	0.1041	0.1869	0.1278
	cover_rate_boot	0.9225	0.9150	0.9075	0.9125	0.9300	0.9025	0.9225	0.9050
Q3	bias	0.0265	0.0324	0.0446	0.0301	0.0248	0.0354	0.0272	0.0314
	emp_sd	0.1250	0.0966	0.1539	0.1041	0.1784	0.1077	0.2196	0.1192
	est_sd_boot	0.1213	0.0935	0.1495	0.1066	0.1647	0.1003	0.1967	0.1138
	cover_rate_boot	0.9175	0.9100	0.9025	0.8925	0.9200	0.9050	0.9175	0.8975

At last, we evaluate the performance of our proposed TE estimators in the settings with $p = 100$ and $n = 2000$. In this scenario, the number of confounders is very large compared to the sample size, and it does not satisfy the order requirement given in Assumption 4. Note that when dealing with high-dimensional covariates, one often assumes a parametric structure on the regression model and imposes a sparsity condition such that a small number of covariates are useful for the prediction The sparsity assumption and the parametric structure are not required in our setting. For the purpose of dimensionality reduction, we apply Principal Component Analysis (PCA) to extract the first 20 leading principal components, and use them to estimate the PS and OR functions via ANNs. For comparison, we also use the original covariates matrix without PCA to fit the nuisance models via ANNs. The resulting TE estimators with and without the PCA procedure are called ANN-PCA and ANN, respectively. Tables 5–6 report the summary statistics of the ANN-based TE estimators for ATE, ATT, QTE and QTT for Model 1 with $p = 100$ and $n = 2000$, based on 400 simulation realizations, when the confounders are generated from Designs 1 & 2 given in Section 7.2. For QTE and QTT, we only report the estimated standard deviations and empirical coverage rates from the weighted bootstrapping, as it is difficult to estimate the asymptotic standard deviations in the quantile settings. The ATE and ATT are estimated by the IPW, OR and DR methods, respectively, while the QTE and QTT are only estimated by the IPW method.

From Table 5, for the estimation of ATE and ATT, we see that the empirical coverage rates obtained from all of the three methods, IPW, OR and DR, are smaller than the nominal level 0.95, and the values of bias and emp_sd are larger than those values given in Tables 1–2 for $p = 5, 10$. The ANN-PCA method yields larger biases but smaller emp_sd than the ANN method. The empirical coverage rates from the ANN-PCA method are closer to the nominal level than those from the ANN method for both designs, but they still cannot reach the nominal level. It is expected that these ANN-based methods have inferior performance for $p = 100$ compared to the $p = 5, 10$ settings, as the order assumption on the dimension p required for ANN approximations does not hold anymore when $p = 100$. As a result, the ANN-based estimators of the nuisance functions (OR and PS functions) are not guaranteed to be consistent estimators, yielding deteriorated performance, and those estimates further affect the estimation of ATE and ATT. The formula of est_sd involves the estimates of both OR and PS functions, so it is not surprising that its value is also affected. From Table 6 for the estimation of QTE and

Table 7
Group comparisons

Covariates		Non-smoker (N_{ns}=3288)		Smoker (N_{ns}=3359)		Std. Dif.	*p*-value
Gender	1 = Male	1404	(41.8%)	2019	(61.41%)	−15.99	<0.001
	0 = Female	1955	(58.2%)	1269	(38.59%)		
Age	Mean(SD)	48.97	(19)	51.73	(17.57)	−6.14	<0.001
Marital	1 = Yes	1989	(59.21%)	1867	(56.78%)	2.01	0.0446
	0 = No	1370	(40.79%)	1421	(43.22%)		
Education	1 = College or above	1626	(48.41%)	1297	(39.45%)	7.36	<0.001
	0 = Less than college	1733	(51.59%)	1991	(60.55%)		
Family PIR	Mean(SD)	2.79	(1.63)	2.57	(1.6)	5.62	<0.001
Alcohol	1 = Yes	1897	(56.48%)	2708	(82.36%)	−22.87	<0.001
	0 = No	1462	(43.52%)	580	(17.64%)		
PHSVIG	1 = Yes	1102	(32.81%)	908	(27.62%)	4.61	<0.001
	0 = No	2257	(67.19%)	2380	(72.38%)		
PHSMOD	1 = Yes	1491	(44.39%)	1376	(41.85%)	2.09	0.0366
	0 = No	1868	(55.61%)	1912	(58.15%)		
SBP	Mean(SD)	126.42	(21.04)	126.63	(19.98)	−0.43	0.6684
DBP	Mean(SD)	72.1	(13.56)	71.61	(14.1)	1.44	0.15

Table 8
The estimates and standard errors of ATE and QTE.

	ATE		QTE		
	IPW	OR	Q1	Q2	Q3
estimate	−0.224	−0.241	−0.400	−0.269	−0.040
est_sd	0.154	0.154	0.157	0.184	0.247
z-value	−1.454	−1.564	−2.547	−1.467	−0.162
p-value	0.073	0.058	0.005	0.071	0.436
est_sd_boot	0.162	0.149	0.156	0.187	0.254
z-value_boot	−1.383	−1.617	−2.564	−1.443	−0.157
p-value_boot	0.083	0.053	0.005	0.074	0.437

QTT, we can observe similar patterns as the results in Table 5, except that the ANN method has slightly larger empirical coverage rates than the ANN-PCA method. In sum, the TE estimation using ANNs in the context of ultra-high dimensional covariates is a challenging task. A sparse model assumption may be needed for high-dimensional settings. The investigation of its methodology and theories is beyond the scope of this paper, and it can be an interesting topic to pursue in the future.

8. Application

In this section, we apply the proposed methods to the data from the National Health and Nutrition Examination Survey (NHANES) to investigate the causal effect of smoking on body mass index (BMI). The collected data consist of 6647 subjects, including 3359 smokers and 3288 nonsmokers. The confounding variables include four continuous variables: age, family poverty income ratio (Family PIR), systolic blood pressure (SBP), and diastolic blood pressure (DBP); six binary variables: gender, marital status, education, alcohol use, vigorous activity over past 30 days (PHSVIG), and moderate activity over past 30 days (PHSMOD). Table 7 presents the group comparisons of all confounding variables in the full dataset. Mean and standard deviation (SD) are presented for continuous variables, while the count and percentage (%) of observations for each group are presented for categorical variables. Standardized difference(Std. Dif.) is calculated as $(\bar{x}_{ns} - \bar{x}_s)/\sqrt{s_{ns}^2/n_{ns} + s_s^2/n_s}$ for continuous variables, and $(p_{ns} - p_s)/\sqrt{pq/n_{ns} + pq/n_s}$ for categorical variables, where $\bar{x}$, s^2 and p denote sample mean, sample variance and sample proportion, and the subscripts ns and s refer to nonsmokers and smokers respectively, and p, q are the overall proportions. The last column shows the p-value of group comparison for each covariate. We notice that the smoking group and nonsmoking group differ greatly in their group characteristics. A naive comparison of the sample mean between smoking and nonsmoking groups will lead to a biased estimation of the smoking effects on BMI.

We apply our proposed ANN methods to estimate the PS and OR functions, respectively. We estimate ATE by the proposed IPW and OR methods, and estimate QTE by the IPW method only. The number of neurons is selected using grid search with 5-fold cross-validation.

Table 8 reports the estimates of ATE and QTE, the estimated standard deviations based on the asymptotic formula (est_sd) and obtained from the weighted bootstrapping (est_sd_boot), and the corresponding z-values and p-values for testing ATE and QTE. The negative values of the estimates indicate that smoking has adverse effects on BMI. From the numerical results based on the estimated asymptotic standard deviations, we see that the p-values of testing ATE are 0.073 and 0.058 by the IPW and OR methods, respectively. We also notice that the p-value for testing QTE at the 25% quantile is very small, which is 0.005. However, the p-value increases to 0.071 at the 50% quantile (median), and further to 0.436 at the 75% quantile. This indicates that smoking has a more prominent effect on the population with smaller BMI, and its effect diminishes as BMI increases; i.e., the effect of smoking becomes less significant as the

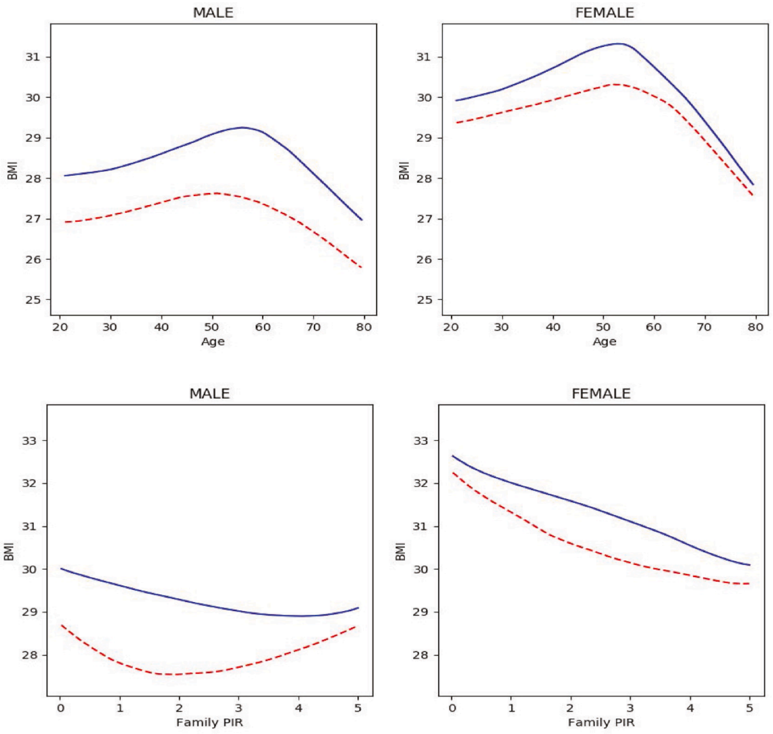

Fig. 2. The plots of $\tau_1(\cdot)$ and $\tau_0(\cdot)$ versus two continuous variables for the smoking and nonsmoking groups, and for males and females, respectively, where the blue solid curves represent nonsmoking group and red dashed line represent smoking group.

value of BMI becomes larger. This interesting pattern cannot be reflected in ATE. We can draw the same inferential conclusions as above when the weighted bootstrap method is applied.

We also examine the relationship between BMI and two continuous confounding variables, age and family poverty income ratio (Family PIR). Fig. 2 depicts the estimated conditional mean functions (OR functions) $\tau_1(\cdot)$ and $\tau_0(\cdot)$ versus the two continuous variables for the smoking and nonsmoking groups, and for males and females, respectively. For each comparison, all the other confounding variables are fixed as constants: the continuous variables take the values of their means while the categorical variables are kept as married, college or above, drinks alcohol, no vigorous activity and no moderate activity. It is interesting to notice that for the same age or Family PIR, the estimated conditional mean in the smoking group is smaller than that in the nonsmoking group for both males and female, and the estimated conditional mean in the male group is also smaller than that in the female group for both smoker and nonsmoker. We can clearly see nonlinear relationships between age and BMI as well as between Family PIR and BMI. Age is positively associated with BMI when it is less than 50, and the association between age and BMI becomes more negative as people get older. We also see that the smoking effects on BMI are very different between the male group and the female group. Smoking has a more significant effect on BMI for males than for females at the same age. In the male group, the BMI decreases as family income increases until it reaches the poverty threshold, and then the BMI increases with family income for smokers. For nonsmokers, it shows a relatively flatter trend. In the female group, the BMI keeps decreasing as family income increases for both smokers and nonsmokers.

9. Conclusion

In this paper, we provide a unified framework for efficient estimation of various types of TEs in observational data using ANNs with a diverging number of covariates/confounders. Our framework allows for settings with binary or multi-valued treatment variables, and includes the average, quantile, and asymmetric least squares TEs as special cases. We estimate the TEs through a generalized optimization, which involves an ANN estimation of one nuisance function only. When the unknown nuisance function is approximated by ANNs with one hidden layer, we show that the number of confounders is allowed to increase with the sample size. We further investigate how fast the number of confounders can grow with the sample size (n) to ensure root-n consistency, asymptotic normality and efficiency of the resulting TE estimator. These statistical properties are essential for inferring causations. We also show that a simple weighted bootstrap provides consistent confidence sets for the general TEs without the need to estimate the asymptotic variance. Compared to other approaches based on efficient influence functions, our general optimization-based estimation and inference methods are especially attractive for efficient estimation of complex TEs such as quantile and asymmetric TEs. Practically, we illustrate our proposed method through simulation studies and a real data example. The numerical studies support our theoretical findings.

We have shown that the ANNs with one hidden layer can circumvent the "curse of dimensionality" and the resulting TE estimators enjoy root-n consistency under the condition that the target function is in a mixed smoothness class. Our new results advance the understanding of the required conditions and the statistical properties for ANNs in causal inference, and lay a theoretical foundation to demonstrate that ANNs are promising tools for causality analysis when the dimension is allowed to diverge, whereas most existing works on ANNs estimation still assume the dimension of co to be fixed. In the online supplemental materials, we discuss the extension of our method for efficient estimation of and inference on general TEs when the nuisance function is approximated by fully-connected ANNs with multiple hidden layers. Finally, our optimization-based method can be also extended to causal analysis with continuous treatment variables and with longitudinal data designs. Thorough investigations are needed to develop the computational algorithms and establish the theoretical properties of the resulting estimators in these settings.

Acknowledgments

The authors sincerely thank the editor Elie Tamer and the referees for their constructive suggestions and comments. The research of Chen is partially supported by Cowles Foundation. The research of Liu and Ma is supported in part by the U.S. NSF grants DMS-17-12558, DMS-20-14221 and DMS-23-10288 and the UCR Academic Senate CoR Grant. Zheng Zhang is supported by the fund from National Key R&D Program of China [grant number 2022YFA1008300], the fund from the National Natural Science Foundation of China [grant number 12371284], and the fund for building world-class universities (disciplines) of Renmin University of China [project number KYGJC2023011].

Appendix A. Polynomial approximation and curse of dimensionality

Let $\mathscr{X} = [0,1]^p$ and $s_0 \in \mathbb{N}$, Lorentz (1966, Theorem 8) states that for any s_0-times continuously differentiable function $f(\cdot) : \mathscr{X} \to \mathbb{R}$, i.e. $\sup_{|\boldsymbol{\alpha}|_1 \le s_0} \sup_{x \in \mathscr{X}} |D^{\boldsymbol{\alpha}} f(x)| \le 1$, there exist polynomials $P_{n_1,\dots,n_p}(x)$, of degree n_i in x_i, such that

$$\sup_{x \in \mathscr{X}} \left| f(x) - P_{n_1,\dots,n_p}(x) \right| \le C_p \cdot \sum_{i=1}^{p} \frac{1}{n_i^{s_0}},$$

where C_p is a constant depending on p.

Consider a K-dimensional polynomial sieve $\{u_K(x)\}$ of the form:

$$u_1(x) = 1,\ u_2(x) = (1, x_1)^\top, \dots, u_{p+1}(x) = \left(1, x_1, \dots, x_p\right)^\top,\ u_{p+2}(x) = \left(1, x_1, \dots, x_p, x_1^2\right)^\top, \dots.$$

To ensure all degrees of $(x_1, \dots, x_p)$ get up to some order $n_0 \in \mathbb{N}$, i.e. $\min\{n_1, \dots, n_p\} \ge n_0$, we require $K = (n_0+1)^p$. Therefore, for any function $f(\cdot)$ satisfying $\sup_{|\boldsymbol{\alpha}|_1 \le s_0} \sup_{x \in \mathscr{X}} |D^{\boldsymbol{\alpha}} f(x)| \le 1$, the approximation rate based on the polynomial sieve $\{u_K(x)\}$ is

$$\inf_{\lambda_K \in \mathbb{R}^K} \sup_{x \in \mathscr{X}} \left| f(x) - \lambda_K^\top u_K(x) \right| \le C_p \cdot p \cdot K^{-\frac{s_0}{p}}.$$

Note that for any function in the mixed smoothness ball defined in Section 2.1, i.e. $f(\cdot) \in \mathscr{W}^{m,1+\epsilon,\infty}(\mathscr{X})$ for $\epsilon \in (0,1)$, we only have $\sup_{|\boldsymbol{\alpha}|_1 \le m+1} \sup_{x \in \mathscr{X}} |D^{\boldsymbol{\alpha}} f(x)| \le 1$. In light of the compactness of $\mathscr{X}$, the $L^2(dF_X)$-approximation error based on the polynomial sieve $\{u_K(x)\}$ is $C_p \cdot p \cdot K^{-\frac{m+1}{p}}$, which severely suffers from the curse of dimensionality.

Appendix B. Proof of Theorem 1

For a regular function $f(\cdot) : \mathscr{X} \to \mathbb{R}$ whose Fourier transform is denoted by $\widetilde{f}(\cdot)$, by using the identity $e^{-\pi i} = -1$ and a change of variables, we have

$$\begin{aligned}
\widetilde{f}\quad (t_1, t_2, \dots, t_p) &= \frac{1}{\{2\pi\}^p} \int_{\mathbb{R}^p} f\left(x_1, x_2, \dots, x_p\right) e^{-it_1 x_1} \cdot e^{-it_2 x_2} \cdots e^{-it_p x_p} dx_1 dx_2 \cdots dx_p \\
&= -\frac{1}{\{2\pi\}^p} \int_{\mathbb{R}^p} f\left(x_1, x_2, \dots, x_p\right) e^{-it_1 x_1 - i\pi} \cdot e^{-it_2 x_2} \cdots \cdot e^{-it_p x_p} dx_1 dx_2 \cdots dx_p \\
&= -\frac{1}{\{2\pi\}^p} \int_{\mathbb{R}^p} f\left(x_1, x_2, \dots, x_p\right) e^{-it_1 \left(x_1 + \frac{\pi}{t_1}\right)} \cdot e^{-it_2 x_2} \cdots e^{-it_p x_p} dx_1 dx_2 \cdots dx_p \\
&= -\frac{1}{\{2\pi\}^p} \int_{\mathbb{R}^p} f\left(x_1 - \frac{\pi}{t_1}, x_2, \cdots, x_p\right) e^{-it_1 x_1} \cdot e^{-it_2 x_2} \cdots e^{-it_p x_p} dx_1 dx_2 \cdots dx_p.
\end{aligned}$$

Then

$$
\begin{aligned}
& 2\widetilde{f}\left(t_1, t_2, \ldots, t_p\right) \\
= \quad & \frac{1}{\{2\pi\}^p}\int_{\mathbb{R}^p}\left[f\left(x_1, x_2, \cdots, x_p\right) - f\left(x_1 - \frac{\pi}{t_1}, x_2, \cdots, x_p\right)\right] e^{-it_1x_1}\cdots e^{-it_px_p}dx_1dx_2\cdots dx_p \\
= \quad & \frac{1}{\{2\pi\}^p}\int_{\mathbb{R}^p}\frac{\Delta_{x_1}^{\frac{\pi}{t_1}} f\left(x_1, x_2, \ldots, x_p\right)}{\left(\frac{\pi}{t_1}\right)} \cdot \left(\frac{\pi}{t_1}\right) e^{-it_1x_1} \cdot e^{-it_2x_2}\cdots e^{-it_px_p}dx_1dx_2\cdots dx_p,
\end{aligned}
$$

where $\Delta_{x_i}^{\delta}$ is the finite difference operator defined by

$$
\Delta_{x_i}^{\delta} f\left(x_1, \ldots, x_p\right) := f\left(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_p\right) - f\left(x_1, \ldots, x_{i-1}, x_i - \delta, x_{i+1}, \ldots, x_p\right),
$$

for $i \in \{1, 2, \ldots, p\}$ and $\delta > 0$. Inductively, we have that for any nonnegative integer $s_1 \in \mathbb{N}$ and a constant $\epsilon \in (0, 1]$:

$$
\begin{aligned}
& \widetilde{f}\left(t_1, t_2, \ldots, t_p\right) \\
= \quad & \frac{1}{2}\cdot\frac{1}{\{2\pi\}^p}\int_{\mathbb{R}^p}\frac{\Delta_{x_1}^{\frac{\pi}{t_1}} f(\boldsymbol{x})}{\left(\frac{\pi}{t_1}\right)} \cdot \left(\frac{\pi}{t_1}\right) e^{-it_1x_1} \cdot e^{-it_2x_2}\cdots e^{-it_px_p}dx_1dx_2\cdots dx_p \\
= \quad & \frac{1}{2^2}\cdot\frac{1}{\{2\pi\}^p}\int_{\mathbb{R}^p}\frac{\left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^2 f(\boldsymbol{x})}{\left(\frac{\pi}{t_1}\right)^2} \cdot \left(\frac{\pi}{t_1}\right)^2 e^{-it_1x_1} \cdot e^{-it_2x_2}\cdots e^{-it_px_p}dx_1dx_2\cdots dx_p \\
\vdots \quad & \\
= \quad & \frac{1}{2^{s_1}}\cdot\frac{1}{\{2\pi\}^p}\int_{\mathbb{R}^p}\frac{\left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^{s_1} f(\boldsymbol{x})}{\left(\frac{\pi}{t_1}\right)^{s_1}} \cdot \left(\frac{\pi}{t_1}\right)^{s_1} e^{-it_1x_1} \cdot e^{-it_2x_2}\cdots e^{-it_px_p}dx_1dx_2\cdots dx_p \\
= \quad & \frac{1}{2^{s_1+1}}\cdot\frac{1}{\{2\pi\}^p}\int_{\mathbb{R}^p}\frac{\left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^{s_1+1} f(\boldsymbol{x})}{\left(\frac{\pi}{t_1}\right)^{s_1+1}} \cdot \left(\frac{\pi}{t_1}\right)^{s_1+1} e^{-it_1x_1} \cdot e^{-it_2x_2}\cdots e^{-it_px_p}dx_1dx_2\cdots dx_p \\
= \quad & \frac{1}{2^{s_1+2}}\cdot\frac{1}{\{2\pi\}^p}\int_{\mathbb{R}^p}\frac{\left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^{s_1+2} f(\boldsymbol{x})}{\left(\frac{\pi}{t_1}\right)^{s_1+1+\epsilon}} \cdot \left(\frac{\pi}{t_1}\right)^{s_1+1+\epsilon} e^{-it_1x_1} \cdot e^{-it_2x_2}\cdots e^{-it_px_p}dx_1dx_2\cdots dx_p.
\end{aligned}
$$

For a vector of nonnegative integers $(s_1, \ldots, s_p)$, by applying the same argument to $\{t_2, \ldots, t_p\}$ gives:

$$
\begin{aligned}
& \widetilde{f}\left(t_1, t_2, \ldots, t_p\right) \\
= \quad & \left(\frac{1}{2}\right)^{\sum_{j=1}^{p}\left(s_j+2\right)} \cdot \frac{1}{\{2\pi\}^p}\int_{\mathbb{R}^p}\frac{\left(\Delta_{x_p}^{\frac{\pi}{t_p}}\right)^{s_p+2}\cdots\left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^{s_1+2} f(\boldsymbol{x})}{\left(\frac{\pi}{t_p}\right)^{s_p+1+\epsilon}\cdots\left(\frac{\pi}{t_1}\right)^{s_1+1+\epsilon}} \cdot \left(\frac{\pi}{t_1}\right)^{s_1+1+\epsilon}\cdots\left(\frac{\pi}{t_p}\right)^{s_p+1+\epsilon} \\
& \qquad \times e^{-it_1x_1} \cdot e^{-it_2x_2}\cdots e^{-it_px_p}dx_1dx_2\cdots dx_p \\
= \quad & \left(\frac{\pi}{2}\right)^{|s|_1+2p} \cdot \frac{1}{\{2\pi\}^p}\int_{[0,1]^p}\frac{\left(\Delta_{x_p}^{\frac{\pi}{t_p}}\right)^{s_p+2}\cdots\left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^{s_1+2} f(\boldsymbol{x})}{\left(\frac{\pi}{t_p}\right)^{s_p+1+\epsilon}\cdots\left(\frac{\pi}{t_1}\right)^{s_1+1+\epsilon}} \cdot \frac{1}{t_1^{s_1+1+\epsilon}\cdots t_p^{s_p+1+\epsilon}} \\
& \qquad \times e^{-it_1x_1} \cdot e^{-it_2x_2}\cdots e^{-it_px_p}dx_1dx_2\cdots dx_p.
\end{aligned}
$$

Then for a vector of nonnegative integers $(s_1, \ldots, s_p)$ and a constant $\epsilon \in (0, 1]$:

$$
|\widetilde{f}\left(t_1, t_2, \ldots, t_p\right)| \tag{B.1}
$$

$$\begin{aligned} \leq\ & \left(\frac{\pi}{2}\right)^{|\boldsymbol{s}|_1+2p} \cdot \frac{1}{\{2\pi\}^p} \int_{[0,1]^p} \left| \frac{\left(\Delta_{x_p}^{\frac{\pi}{t_p}}\right)^{s_p+2} \cdots \left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^{s_1+2} f(\boldsymbol{x})}{\left(\frac{\pi}{t_p}\right)^{s_p+1+\epsilon} \cdots \left(\frac{\pi}{t_1}\right)^{s_1+1+\epsilon}} \right| dx_1 dx_2 \cdots dx_p \\ & \times \frac{1}{|t_1|^{s_1+1+\epsilon} \cdots |t_p|^{s_p+1+\epsilon}}. \end{aligned} \tag{B.2}$$

Note that

$$\lim_{t_1,\ldots,t_p\to\infty} \frac{\left(\Delta_{x_p}^{\frac{\pi}{t_p}}\right)^{s_p+1} \cdots \left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^{s_1+1} f(\boldsymbol{x})}{\left(\frac{\pi}{t_p}\right)^{s_p+1} \cdots \left(\frac{\pi}{t_1}\right)^{s_1+1}} = \partial_{x_p}^{s_p+1} \cdots \partial_{x_1}^{s_1+1} f(\boldsymbol{x}),$$

provided that the limit exists.

For any $f(\cdot) \in \mathscr{W}^{m,1+\epsilon,\infty}([0,1]^p)$, by definition we have

$$\sup_{\{\forall \boldsymbol{\alpha}: |\boldsymbol{\alpha}|_1 \leq m\}} \sup_{\{\boldsymbol{x}\in[0,1]^p, t_1>0,\ldots,t_p>0\}} \left| \frac{\left(\Delta_{x_p}^{\frac{\pi}{t_p}}\right) \cdots \left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right) \partial_{x_1} \cdots \partial_{x_p} D^{\boldsymbol{\alpha}} f(\boldsymbol{x})}{\left(\frac{\pi}{t_p}\right)^{\epsilon} \cdots \left(\frac{\pi}{t_1}\right)^{\epsilon}} \right| \leq 1.$$

Then we can find a large enough constant $M_0 > 0$ such that $\min_{\{j\in 1,\ldots,p\}} |t_j| \geq M_0$, we have

$$\sup_{\{\forall \boldsymbol{\alpha}: |\boldsymbol{\alpha}|_1 \leq m\}} \sup_{\boldsymbol{x}\in[0,1]^p} \left| \frac{\left(\Delta_{x_p}^{\frac{\pi}{t_p}}\right)^{\alpha_p+2} \cdots \left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^{\alpha_1+2} f(\boldsymbol{x})}{\left(\frac{\pi}{t_p}\right)^{\alpha_p+1+\epsilon} \cdots \left(\frac{\pi}{t_1}\right)^{\alpha_1+1+\epsilon}} \right| \leq 2,$$

namely,

$$M_0 := \inf\left\{ C \in \mathbb{R} : \sup_{\{|t_j|\geq C\}_{j=1}^p} \sup_{\{\forall \boldsymbol{\alpha}: |\boldsymbol{\alpha}|_1 \leq m\}} \sup_{x\in\mathscr{X}} \frac{\left| \left(\Delta_{x_p}^{\frac{\pi}{t_p}}\right)^{\alpha_p+2} \cdots \left(\Delta_{x_1}^{\frac{\pi}{t_1}}\right)^{\alpha_1+2} f(x) \right|}{\left(\frac{\pi}{t_p}\right)^{\alpha_p+1+\epsilon} \cdots \left(\frac{\pi}{t_1}\right)^{\alpha_1+1+\epsilon}} \leq 2 \right\}. \tag{B.3}$$

Hence, by (B.1), for any $f(\cdot) \in \mathscr{W}^{m,1+\epsilon,\infty}([0,1]^p)$, we have that for any $k \in \{1,..,p\}$:

$$\begin{aligned} & \sup_{t_1,\ldots,t_k} |\widetilde{f}(t_1,\ldots,t_k,t_{k+1},\ldots,t_p)| \cdot \prod_{j=k+1}^{p} I(|t_j| \geq M_0) \\ \leq\ & \left(\frac{\pi}{2}\right)^{m+2p} \cdot \frac{2}{\{2\pi\}^p} \cdot \frac{1}{|t_{k+1}|^{\gamma_{k+1}+1+\epsilon} \cdots |t_p|^{\gamma_p+1+\epsilon}}, \text{ where } \sum_{j=k+1}^{p} \gamma_j \leq m. \end{aligned} \tag{B.4}$$

We emphasize that (B.4) holds for an *arbitrarily* fixed $k \in \{1,\ldots,p\}$ and for *all* $\{(\gamma_{k+1},\ldots,\gamma_p) : \sum_{j=k+1}^{p} \gamma_j \leq m\}$.

We next find the bound for $v_{f,m}$. Without loss of generality, we assume the function $\widetilde{f}(t_1,..,t_p)$ is symmetric in $\boldsymbol{t} = (t_1,\ldots,t_p)$. Note that

$$
\begin{aligned}
v_{f,m} &= \int_{\mathbb{R}^p} |\boldsymbol{t}|_1^m \cdot |\widetilde{f}(\boldsymbol{t})| d\boldsymbol{t} \\
&= \Big(\int_{|t_p|\in[0,M_0]} + \int_{|t_p|\in[M_0,\infty)}\Big)\cdots\Big(\int_{|t_1|\in[0,M_0]} + \int_{|t_1|\in[M_0,\infty)}\Big)\Big\{\sum_{k=1}^{p}|t_k|\Big\}^m \cdot |\widetilde{f}(\boldsymbol{t})| dt_1\cdots dt_p \\
&= \sum_{i=0}^{p}\binom{p}{i}\int_{|t_p|\in[M_0,\infty)}\cdots\int_{|t_{i+1}|\in[M_0,\infty)}\int_{|t_i|\in[0,M_0]}\cdots\int_{|t_1|\in[0,M_0]}\Big\{\sum_{k=1}^{p}|t_k|\Big\}^m \cdot |\widetilde{f}(\boldsymbol{t})| dt_1\cdots dt_p \\
&= \sum_{\sum_{j=1}^p \alpha_j = m}\binom{m}{\alpha_1,\dots,\alpha_p}\sum_{i=0}^{p}\binom{p}{i} \\
&\quad \times\int_{|t_p|\in[M_0,\infty)}\cdots\int_{|t_{i+1}|\in[M_0,\infty)}\int_{|t_i|\in[0,M_0]}\cdots\int_{|t_1|\in[0,M_0]}|t_1|^{\alpha_1}\cdots|t_p|^{\alpha_p}\cdot|\widetilde{f}(\boldsymbol{t})| dt_1\cdots dt_p \\
&\le \sum_{\sum_{j=1}^p \alpha_j = m}\binom{m}{\alpha_1,\dots,\alpha_p}\sum_{i=0}^{p}\binom{p}{i}\cdot M_0^{\alpha_1+\cdots\alpha_i+i} \\
&\quad \times\int_{|t_p|\in[M_0,\infty)}\cdots\int_{|t_{i+1}|\in[M_0,\infty)}|t_{i+1}|^{\alpha_{i+1}}\cdots|t_p|^{\alpha_p}\cdot\sup_{t_1,\dots,t_i}\big|\widetilde{f}\big(t_1,\dots,t_i,t_{i+1}\dots,t_p\big)\big| dt_{i+1}\cdots dt_p.
\end{aligned}
$$

For every $i \in \{0, 1, ..p\}$ and every $(\alpha_{i+1}, \dots, \alpha_p)$ in the summand, since $\alpha_{i+1} + \cdots\alpha_p \le m$, by applying (B.4) we have

$$
\begin{aligned}
&\sup_{t_1,\dots,t_i}|\widetilde{f}\big(t_1,\dots,t_i,t_{i+1},\dots,t_p\big)|\cdot\prod_{j=i+1}^{p} I(|t_j|\ge M_0) \\
\le\ & \Big(\frac{\pi}{2}\Big)^{m+2p}\cdot\frac{2}{\{2\pi\}^p}\cdot\frac{1}{|t_{i+1}|^{\alpha_{i+1}+1+\epsilon}\cdots|t_p|^{\alpha_p+1+\epsilon}}.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
v_{f,m} \le\ & \sum_{\sum_{j=1}^p \alpha_j = m}\binom{m}{\alpha_1,\dots,\alpha_p}\sum_{i=0}^{p}\binom{p}{i}M_0^{\alpha_1+\cdots+\alpha_i+i}\cdot\Big(\frac{\pi}{2}\Big)^{m+2p}\cdot\frac{2}{\{2\pi\}^p} \\
&\quad\times\int_{|t_p|\in[M_0,\infty)}\cdots\int_{|t_{i+1}|\in[M_0,\infty)}\frac{1}{|t_{i+1}|^{1+\epsilon}\cdots t_p^{1+\epsilon}}dt_{i+1}\cdots dt_p \\
\le\ & \sum_{\sum_{j=1}^p \alpha_j = m}\binom{m}{\alpha_1,\dots,\alpha_p}\sum_{i=0}^{p}\binom{p}{i}M_0^{m+i}\cdot\Big(\frac{\pi}{2}\Big)^{m+2p}\cdot\frac{2\cdot 2^p}{\{2\pi\}^p}\cdot\Big(\frac{M_0^{-\epsilon}}{\epsilon}\Big)^{p-i} \\
=\ & \sum_{\sum_{j=1}^p \alpha_j = m}\binom{m}{\alpha_1,\dots,\alpha_p}\sum_{i=0}^{p}\binom{p}{i}\Big(M_0\cdot\frac{\pi}{2}\Big)^m\cdot\big(M_0^{1+\epsilon}\big)^i\cdot\Big(\frac{\pi}{2}\Big)^{2p}\cdot\frac{2}{\pi^p}\cdot\frac{\big(M_0^{-\epsilon}\big)^p}{\epsilon^{p-i}} \\
\le\ & \sum_{\sum_{j=1}^p \alpha_j = m}\binom{m}{\alpha_1,\dots,\alpha_p}\sum_{i=0}^{p}\binom{p}{i}\Big(M_0\cdot\frac{\pi}{2}\Big)^m\cdot\Big(\frac{\pi}{2}\Big)^{2p}\cdot\frac{2}{\pi^p}\cdot\frac{M_0^p}{\epsilon^{p-i}} \\
=\ & \Big(\frac{\pi}{2}\Big)^{2p}\cdot\frac{2}{\pi^p}\cdot M_0^p\cdot\Big(M_0\cdot\frac{\pi}{2}\Big)^m\sum_{\sum_{j=1}^p \alpha_j = m}\binom{m}{\alpha_1,\dots,\alpha_p}\sum_{i=0}^{p}\binom{p}{i}\cdot\Big[\frac{1}{\epsilon}\Big]^{p-i} \\
=\ & \frac{2}{2^p}\cdot M_0^{p+m}\cdot\Big(\frac{\pi}{2}\Big)^{m+p}\cdot p^m\cdot\Big[\frac{1}{\epsilon}+1\Big]^p \\
\le\ & 2\cdot\Big(M_0\cdot\frac{\pi}{2}\Big)^m\cdot\Big[M_0\cdot\Big(\frac{\pi}{2}\Big)\cdot\Big(\frac{1}{2\epsilon}+\frac{1}{2}\Big)\Big]^p = 2\cdot\Big(M_0\cdot\frac{\pi}{2}\Big)^m\cdot M^p,
\end{aligned}
$$

for some universial large constant M defined by

$$
M := M_0\cdot\Big(\frac{\pi}{2}\Big)\cdot\Big(\frac{1}{2\epsilon}+\frac{1}{2}\Big).
$$

Appendix C. Proof of Theorem 3

We first show $\widehat{\beta}_d \xrightarrow{p} \beta_d^*$ for any $d \in \{0,1,\ldots,J\}$. The details of the proof are given in the on-line supplemental materials. Next, we establish the asymptotic normality for $\sqrt{n}\{\widehat{\beta}_d - \beta_d^*\}$. Since the loss function $\mathscr{L}(\cdot)$ may not be smooth (e.g. $\mathscr{L}(v) = v\{\tau - \mathbb{1}(v \le 0)\}$ in quantile regression), the Delta method for deriving the large sample property is not applicable in our case. To circumvent this problem, we apply the *nearness of arg mins* argument. Define

$$G_{d,n}(\beta, \widehat{\pi}_d) := \frac{1}{n}\sum_{i=1}^{n} \frac{D_{di}}{\widehat{\pi}_d(\boldsymbol{X}_i)} \mathscr{L}(Y_i - \beta). \tag{C.1}$$

By definition

$$\widehat{\beta}_d = \underset{\beta\in\Theta}{\operatorname{argmin}} G_{d,n}(\beta, \widehat{\pi}_d) = \underset{\beta\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \frac{D_{di}}{\widehat{\pi}_d(\boldsymbol{X}_i)} \mathscr{L}(Y_i - \beta), \tag{C.2}$$

then

$$\begin{aligned}
\widehat{\beta}_d = \;& \underset{\beta\in\Theta}{\operatorname{argmin}} n\left\{G_{d,n}(\beta, \widehat{\pi}_d) - G_{d,n}\left(\beta_d^*, \widehat{\pi}_d\right)\right\} \\
= \;& \underset{\beta\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{D_{di}}{\widehat{\pi}_d(\boldsymbol{X}_i)} \left\{\mathscr{L}(Y_i - \beta) - \mathscr{L}\left(Y_i - \beta_d^*\right)\right\} \\
= \;& \underset{\beta\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{D_{di}}{\widehat{\pi}_d(\boldsymbol{X}_i)} \Bigg[-\mathscr{L}^{'}\left(Y_i - \beta_d^*\right)\left(\beta - \beta_d^*\right) \\
& \qquad + \left\{\mathscr{L}(Y_i - \beta) - \mathscr{L}\left(Y_i - \beta_d^*\right) + \mathscr{L}^{'}\left(Y_i - \beta_d^*\right)\left(\beta - \beta_d^*\right)\right\}\Bigg]
\end{aligned}$$

By using change of variables and defining the following functions:

$$\begin{aligned}
&\widehat{u}_d := \sqrt{n}\left(\widehat{\beta}_d - \beta_d^*\right),\; u := \sqrt{n}\left(\beta - \beta_d^*\right), \\
&R_d(Y_i, u) := \mathscr{L}\left(Y_i - \left\{\beta_d^* + \frac{u}{\sqrt{n}}\right\}\right) - \mathscr{L}\left(Y_i - \beta_d^*\right) + \mathscr{L}^{'}\left(Y_i - \beta_d^*\right)\cdot \frac{u}{\sqrt{n}}, \\
&Q_{d,n}(u, \widehat{\pi}_d) := \sum_{i=1}^{n} \frac{D_{di}}{\widehat{\pi}_d(\boldsymbol{X}_i)}\left[-\mathscr{L}^{'}\left(Y_i - \beta_d^*\right)\cdot\frac{u}{\sqrt{n}} + R_d(Y_i, u)\right] = n\cdot\left[G_{d,n}(\beta, \widehat{\pi}_d) - G_{d,n}\left(\beta_d^*, \widehat{\pi}_d\right)\right].
\end{aligned}$$

Then we get

$$\widehat{u}_d = \underset{u}{\operatorname{argmin}} Q_{d,n}(u, \widehat{\pi}_d).$$

Next, we define the following quadratic function

$$\begin{aligned}
\widetilde{Q}_{d,n}(u) := \;& \frac{u}{\sqrt{n}}\sum_{i=1}^{n}\left[-\frac{D_{di}}{\pi_d^*(\boldsymbol{X}_i)}\mathscr{L}^{'}\left(Y_i - \beta_d^*\right) + \left(\frac{D_{di}}{\pi_d^*(\boldsymbol{X}_i)} - 1\right)\mathscr{E}_d\left(\boldsymbol{X}_i, \beta_d^*\right)\right] \\
& -\partial_{\beta_d}\mathbb{E}\left[\mathscr{L}^{'}\left(Y_i^*(d) - \beta_d^*\right)\right]\cdot\frac{u^2}{2},
\end{aligned}$$

which does not depend on $\widehat{\pi}_d$, and its minimizer is defined by

$$\widetilde{u}_d := \underset{u}{\operatorname{argmin}} \widetilde{Q}_{d,n}(u).$$

Since $\widetilde{Q}_{d,n}(u)$ is strictly convex and $\partial_{\beta}\mathbb{E}[\mathscr{L}^{'}(Y_i^*(d) - \beta_d^*)]\langle 0$, then the minimizer $\widetilde{u}_d$ is equal to

$$\widetilde{u}_d = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} H_d^{-1}\cdot S_d\left(Y_i, D_{di}, \boldsymbol{X}_i; \beta_d^*\right),$$

where

$$S_d\left(Y_i, D_{di}, \boldsymbol{X}_i; \beta_d^*\right) := \frac{D_{di}}{\pi_d^*(\boldsymbol{X}_i)}\mathscr{L}^{'}\left(Y_i - \beta_d^*\right) - \left(\frac{D_{di}}{\pi_d^*(\boldsymbol{X}_i)} - 1\right)\mathscr{E}_d\left(\boldsymbol{X}_i, \beta_d^*\right)$$

is the influence function of β_d^* and $H_d := -\partial_{\beta_d}\mathbb{E}[\mathscr{L}^{'}(Y_i^*(d) - \beta_d^*)]$.

The desired result can be obtained via the following steps:

- Step I: showing $\xi_{d,n}(u,\widehat{\pi}_d) := \widetilde{Q}_{d,n}(u) - Q_{d,n}(u,\widehat{\pi}_d) = o_p(1)$ for every fixed u;
- Step II: showing $|\widehat{u}_d - \widetilde{u}_d| = o_P(1)$;
- Step III: obtaining the desired result: $\sqrt{n}\{\widehat{\beta}_d - \beta_d^*\} = \widetilde{u}_d + \{\widehat{u}_d - \widetilde{u}_d\} = n^{-1/2}\sum_{i=1}^n H_d^{-1} \cdot S_d(Y_i, D_{di}, X_i; \beta_d^*) + o_P(1)$.

Note that both objective functions $\widetilde{Q}_{d,n}(u)$ and $Q_{d,n}(u,\widehat{\pi}_d)$ are convex in u with probability approaching to one, the pointwise convergence in Step I is sufficient for establishing Step II. The technical proofs of Step I-Step III are provided in the on-line supplemental materials.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jeconom.2023.105555.

References

Ai, C., Linton, O., Motegi, K., Zhang, Z., 2021. A unified framework for efficient estimation of general treatment models. Quant. Econ. 12 (3), 779–816.
Barron, A.R., 1993. Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theory 39 (3), 930–945.
Bauer, B., Kohler, M., 2019. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. Ann. Stat. 47 (4), 2261–2285.
Bungartz, H.-J., Griebel, M., 2004. Sparse grids. Acta Numer. 13, 147–269.
Cao, W., Tsiatis, A.A., Davidian, M., 2009. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 96, 723–734.
Cattaneo, M.D., 2010. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. J. Econom. 155 (2), 138–154.
Chan, K.C.G., Yam, S.C.P., Zhang, Z., 2016. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. J. R. Stat. Soc. Ser. B (Statistical Methodology) 78 (3), 673–700.
Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. Handb. Econom. 6 (B), 5549–5632.
Chen, X., Hong, H., Tarozzi, A., 2008. Semiparametric efficiency in GMM models with auxiliary data. Ann. Stat. 36 (2), 808–843.
Chen, X., Liao, Z., 2015. Sieve semiparametric two-step GMM under weak dependence. J. Econom. 189 (1), 163–186.
Chen, X., Linton, O., Van Keilegom, I., 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica 71 (5), 1591–1608.
Chen, X., Shen, X., 1998. Sieve extremum estimates for weakly dependent data. Econometrica 289–314.
Chen, X., White, H., 1999. Improved rates and asymptotic normality for nonparametric neural network estimators. IEEE Trans. Inf. Theory 45 (2), 682–691.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econom. J. 21, C1–C68.
D'Amour, A., Ding, P., Feller, A., Lei, L., Sekhon, J., 2021. Overlap in observational studies with high-dimensional covariates. J. Econom. 221 (2), 644–654.
DeVore, R., Nowak, R. D., Parhi, R., Siegel, J. W., 2023. Weighted variation spaces and approximation by shallow reLU networks. arXiv preprint arXiv:2307.15772.
E, W., Ma, C., Wu, L., 2022. The barron space and the flow-induced function spaces for neural network models. Constr. Approx. 55 (1), 369–406.
Farrell, M.H., Liang, T., Misra, S., 2021. Deep neural networks for estimation and inference. Econometrica 89 (1), 181–213.
Firpo, S., 2007. Efficient semiparametric estimation of quantile treatment effects. Econometrica 75 (1), 259–276.
Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66 (2), 315–331.
Han, P., Kong, L., Zhao, J., 2019. A general framework for quantile estimation with incomplete data. J. R. Stat. Soc. Ser. B 81, 305–333.
Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161–1189.
Hong, H., Leung, M.P., Li, J., 2020. Inference on finite-population treatment effects under limited overlap. Econom. J. 23 (1), 32–47.
Hornik, K., Stinchcombe, M., White, H., Auer, P., 1994. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Neural Comput. 6 (6), 1262–1275.
Kennedy, E., Ma, Z., McHugh, M., Small, D., 2017. Nonparametric methods for doubly robust estimation of continuous treatment effects. J. R. Stat. Soc. Ser. B 79, 1229–1245.
Klusowski, J.M., Barron, A.R., 2018. Approximation by combinations of reLU and squared reLU ridge functions with l1 and l0 controls. IEEE Trans. Inf. Theory 64 (12), 7649–7656.
Lee, Y.-Y., 2018. Efficient propensity score regression estimators of multivalued treatment effects for the treated. J. Econom. 204 (2), 207–222.
Lorentz, G.G., 1966. Approximation of Functions, Athena Series. Holt, Rinehart and Winston, New York.
Ma, He, 2016. Inference for single-index quantile regression models with profile optimization. Ann. Stat. 44, 1234–1268.
Ma, S., Kosorok, M.R., 2005. Robust semiparametric m-estimation and the weighted bootstrap. J. Multivar. Anal. 96 (1), 190–217.
Ma, X., Wang, J., 2020. Robust inference using inverse probability weighting. J. Am. Stat. Assoc. 115 (532), 1851–1860.
Makovoz, Y., 1996. Random approximants and neural networks. J. Approx. Theory 85 (1), 98–109.
Newey, W.K., Powell, J.L., 1987. Asymmetric least squares estimation and testing. Econometrica 55, 819–847.
Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc. 89 (427), 846–866.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.
Rotnitzky, A., Lei, Q., Sued, M., Robins, J.M., 2012. Improved double-robust estimation in missing data and causal inference models. Biometrika 99, 439–456.
Schmidt-Hieber, J., 2020. Nonparametric regression using deep neural networks with reLU activation function. Ann. Stat. 48, 1875–1897.
Shen, X., 1997. On methods of sieves and penalization. Ann. Stat. 25 (6), 2555–2591.
Stone, C.J., 1994. The use of polynomial splines and their tensor products in multivariate function estimation. Ann. Stat. 22, 118–184.
Tan, Z., 2010. Bounded, efficient and doubly robust estimation with inverse weighting. Biometrika 97, 661–682.
Tsiatis, A., 2007. Semiparametric Theory and Missing Data. Springer Science & Business Media.
van der Laan, M.J., Rose, S., 2011. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer, New York.
Wasserman, L., 2006. All of Nonparametric Statistics. In: Springer Texts in Statistics. Springer, New York.