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
0304-4076/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
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.\footnote{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.}

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.\footnote{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.}

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

\[
\tilde{f}(a) := \frac{1}{(2\pi)^p} \int_{\mathcal{X}} \exp(-ia^\top x)f(x)dx,
\]

where \( a = (a_1, \ldots, a_p)\top \in \mathbb{R}^p \). We define the \( m \)th moment of the Fourier transform of \( f(x) \) as \( \psi_{m} := \int_{\mathcal{X}} |a|^m |\tilde{f}(a)|da \), where \( |a|_1 := \sum_{j=1}^p |a_j| \). Barron (1993), Hornik et al. (1994), Chen and White (1999), and Klusowski and Barron (2018) considered the target function belonging to the Barron class \( \mathcal{F}_m \).

\[
\mathcal{F}_m := \left\{ f : \mathcal{X} \rightarrow \mathbb{R} : f(x) = \int_{\mathcal{X}} \exp(ia^\top x)\tilde{f}(a)da, \quad \psi_{m} < \infty \right\}
\] (2.1)

\( \mathcal{F}_m \) contains a class of functions of \( p \) dimension that admit Fourier representations with the finite \( m \)th moment. The input variables of functions in \( \mathcal{F}_m \) have dimension \( p \), which is allowed to grow to infinity as the sample size \( n \) increases. It is worth noting that \( \psi_m \) 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 \( \mathcal{F}_m \) given in (2.1) and a mixed smoothness ball, and will establish an upper bound for \( \psi_m \) 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 \( \psi_m \) which appears in the approximation error bounds.

Consider to approximate a target function \( f \in \mathcal{F}_m \) using the ANNs, belonging to the class

\[
\mathcal{E}(\psi, B, r, p) = \left\{ g : g(x) = g_0(x; \gamma_0) + \frac{B}{r} \sum_{j=1}^p \gamma_j \psi_j(a_j^\top x), \quad a_j = (a_{j1}, \ldots, a_{jp})^\top \in \mathbb{R}^p, \right. \\
\left. \quad \| a_j \|_2 = 1, \quad |\gamma_j| \leq 1, \quad j \in \{1, \ldots, r\}, \quad B \in \mathbb{R}^+ \right\},
\] (2.2)

where \( g_0(x; \gamma_0) \) is a parametric function indexed by an unknown parameter vector \( \gamma_0 \), and \( \| a_j \|_2 := (|a_{j1}|^2 + \ldots + |a_{jp}|^2)^{1/2} \). The structure of \( g_0(x; \gamma_0) \) depends on the type of the activation function that is used. For example, if the ReLU activation function is used, then \( g_0(x; \gamma_0) = \gamma_0^\top x \). \( \mathcal{E}(\psi, B, r, p) \) 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 \( \psi \) and real-valued input-to-hidden unit weights \( (a_j) \), 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 \( |\gamma_j| \leq 1 \), and \( B \) is a scale factor of all \( \gamma \)’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 \( \ell_2 \) constraint on the inner parameters \( a_j \) to a \( \ell_2 \) normalization \( \| a_j \|_2 = 1 \). 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 \( f \in \mathcal{S}_p^2 \), \( g_0(x; \gamma_0) \equiv 0 \), the activation function \( \psi(\cdot) \) is compactly supported, bounded, and uniformly Lipschitz continuous. If \( B \leq 2v_{j,1} < \infty \), then Chen and White (1999, Theorem 2.1) show that the \( L^2(dF_x) \)-approximation rate of \( f \) based on ANN is

\[
\inf_{g \in \mathcal{S}(w, b, p)} \left\{ \int_x |f(x) - g(x)|^2 dF_x(x) \right\}^{1/2} \leq \text{const} \times v_{j,1} \cdot r^{-1/2}.
\]

Remark 1. Recently, DeVore et al. (2023) propose weighted variation spaces that enlarge the second order Barron space \( \mathcal{S}_p^2 \). 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 \( \text{const} \times \| f \|_{\mathcal{V}_w} \cdot r^{-1/2} \), where \( \| \cdot \|_{\mathcal{V}_w} \) is the weighted variation norm defined in DeVore et al. (2023, Section 4, page 8).

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 \( \mathcal{S}_p^m \) 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 \( \mathcal{V}_{j,m} \) depends on the dimension \( p \). To address these questions, we build a connection between the Barron class \( \mathcal{S}_p^m \) and a mixed smoothness ball, and establish an upper bound for \( \mathcal{V}_{j,m} \).

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

\[
D^{\alpha} := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_p}}{\partial x_p^{\alpha_p}}.
\]

For an integer \( s \in \mathbb{N}_0 \), denote the class of all \( s \)-times continuously differentiable real-valued functions on \( \mathcal{X} \) by \( \mathcal{C}^{s,\infty}(\mathcal{X}) \):

\[
\mathcal{C}^{s,\infty}(\mathcal{X}) := \left\{ f(\cdot) : \sup_{|\alpha| \leq s} |D^\alpha f(x)| \leq 1 \right\}.
\]

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

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

\[
\partial_{x_1} \cdots \partial_{x_s} := \frac{\partial^s}{\partial x_1 \cdots \partial x_s}.
\]
For any \( \delta > 0 \), define \( \Delta_\delta^i \) to be the difference operator by
\[
\Delta_\delta^i f(x_1, \ldots, x_p) := f(x_1, \ldots, x_{i-1}, x_i + \delta, x_{i+1}, \ldots, x_p) - f(x_1, \ldots, x_{i-1}, x_i, \ldots, x_p),
\]
for \( i \in \{1, 2, \ldots, p\} \). We have
\[
\lim_{\delta \to 0} \frac{\Delta_\delta^1 \cdots \Delta_\delta^p f(x)}{\delta_1 \cdots \delta_p} = \partial_{x_1} \cdots \partial_{x_p} f(x),
\]
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 \( \mathcal{Y} \) for any \( X \). Chen et al. (2004), we define the mixed smoothness ball of order \( m \) of the form
\[
\mathcal{Y}^m := \left\{ f : f(\cdot) \in \mathcal{E}^m(\mathcal{X}), \sup_{\{\|x\|_1 = \epsilon\} \in \mathcal{X}} \sup_{\{\|x\|_1 = \epsilon\} \in \mathcal{X}} \sup_{\{\|x\|_1 = \epsilon\} \in \mathcal{X}} |\Delta_\delta^1 \cdots \Delta_\delta^p f(x) - f(x)| \leq 1 \right\}
\]
and
\[
\mathcal{Y}^m := \left\{ f : f(\cdot) \in \mathcal{E}^m(\mathcal{X}), \sup_{\{\|x\|_1 = \epsilon\} \in \mathcal{X}} \sup_{\{\|x\|_1 = \epsilon\} \in \mathcal{X}} \sup_{\{\|x\|_1 = \epsilon\} \in \mathcal{X}} |\Delta_\delta^1 \cdots \Delta_\delta^p f(x) - f(x)| \leq 1 \right\}
\]

Theorem 1

Let \( f \in \mathcal{Y}^m \) for some \( \epsilon \in [0, 1] \), then \( f \in \mathcal{Y}^m \) and
\[
\nu(f) \leq 2 \left(M_0 \cdot \frac{\ln \frac{1}{\epsilon}}{\epsilon} \right)^m \cdot \mathcal{M}^m
\]
for some universal constant \( \mathcal{M} \) defined by \( \mathcal{M} := M_0 \cdot \left(\frac{1}{\epsilon} \right)^{\frac{1}{2}} \), where the definition of \( M_0 \) is given in (B.3).

Remark 2

In particular, if \( \epsilon = 1 \) in (2.7), we obtain
\[
\mathcal{Y}^m := \left\{ f(\cdot) : f(\cdot) \in \mathcal{E}^m(\mathcal{X}), \sup_{\{\|x\|_1 = \epsilon\} \in \mathcal{X}} \sup_{\{\|x\|_1 = \epsilon\} \in \mathcal{X}} \sup_{\{\|x\|_1 = \epsilon\} \in \mathcal{X}} |\Delta_\delta^1 \cdots \Delta_\delta^p f(x) - f(x)| \leq 1 \right\}
\]
where
\[
\Delta_\delta^1 \cdots \Delta_\delta^p := \frac{\partial^p}{\partial x_1 \cdots \partial x_p}
\]
The functions in the mixed smoothness ball \( \mathcal{Y}^m \) 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., \( \mathcal{E}^m(\mathcal{X}) \). It is worth noting that \( \mathcal{Y}^m \) is much broader than the Sobolev ball of order \( s = m + 2p \); indeed, we have the following inclusion relation:
\[
\mathcal{E}^{m+2p}(\mathcal{X}) \subset \mathcal{Y}^m \subset \mathcal{Y}^{m+1}(\mathcal{X}) \subset \mathcal{Y}^m, \quad \text{for } \epsilon \in [0, 1].
\]
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 \( J = \{0, 1, \ldots, J\} \), 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 \( \mathcal{L}(\cdot) \) denote a nonnegative and strictly convex loss function satisfying \( \mathcal{L}(0) = 0 \) and \( \mathcal{L}(v) \geq 0 \) for all \( v \in \mathbb{R} \). The derivative of \( \mathcal{L}(\cdot) \) exists almost everywhere and non-constant which is denoted by \( \mathcal{L}'(\cdot) \).

where

\[
\beta^* := \arg\min_{\beta} \sum_{d=0}^{J} E[\mathcal{L}(Y^*(d) - \beta_d)].
\]

(3.1)

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

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

However, due to the selection in treatment, the true value \( \beta^* \) 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 \( d \in \mathcal{D} \), \( Y^*(d) \perp D | X \).

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 \( X \) to grow with sample size \( n \). Specifically, we work with triangular array data \( \{(D_{in}, X_{in}, Y_{in}), i = 1, ..., n, n = 1, 2, \ldots\} \) defined on some common probability space \( (\Omega, \mathcal{G}, \mathbb{P}) \). Each \( X_{in} \) is a vector whose dimension \( p_n \) may grow with \( n \), the support of \( X_{in} \) is assumed to be \( \{0, 1\}^p \). For each given \( n \), these vectors are independent across \( i \), but not necessarily identically distributed. The law \( P_n \) of \( \{(D_{in}, X_{in}, Y_{in}), i = 1, ..., n\} \) can change with \( n \), though we do not make explicit use of \( P_n \). Thus, all parameters (including \( p_n \)) that characterize the distribution of \( \{(D_{in}, X_{in}, Y_{in}), i = 1, ..., n\} \) 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 \( \beta^* \) can be identified by the minimizer of the following optimization problem:

\[
\beta^* = \arg\min_{\beta} \sum_{d=0}^{J} E\left[ \mathcal{L} \left( Y - \beta_d, \pi_d^*(X) \right) \right].
\]

(3.2)

where \( \pi_d^*(X) := \mathbb{P}(D = d | X) \) 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 \( \beta^* \) 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

\[
\mathcal{F}_n := \mathcal{F}(\psi, B_n, r_n, p).
\]

Denote \( D_{d} := 1(D = d) \) 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
4. Large sample properties of estimators

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

We first introduce sufficient conditions for the convergence rates of our ANN estimators \((\tilde{g}_d^{(m)})_{d=0}^m\) for the unknown PS nuisance functions.

Assumption 2. For every \(d \in \mathcal{D} = \{0, 1, \ldots, J\}\) and \(m \geq 1\), we assume \(g_d^{(m)}(\cdot) \in \mathcal{F}_P^m\).
Assumption 3. (i) The dimension of $X_i$ is denoted by $p \in \mathbb{N}$ and the number of hidden units is denoted by $r_n \in \mathbb{N}$. They satisfy
\[
\max \left\{ v_{K,m}^r \cdot r_n^{1+\frac{1}{p}}, \sqrt{r_n \cdot p \cdot \log n} \right\} = o\left(n^{\frac{1}{4}}\right).
\]
(ii) The bound of the hidden-to-output weights, $B_n$, specified in (2.2) satisfies $B_n \leq 2v_{K,m}^r$. 

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 $L^2(dF_{\mathbf{X}})$-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 $\mathcal{F}_p^{m,1:1,\infty}(\mathcal{X})$ belongs to the Barron space $\mathcal{F}_p^{m}$. Assumption 3 (i) can be implied by the following primitive condition:

Assumption 3’. (i) Suppose $g_{d}^* \in \mathcal{F}_p^{m,1:1,\infty}(\mathcal{X})$, $p$ and $r_n$ are allowed to grow to infinity as the sample size $n$ increases, with the rates
\[
p \leq a_n \cdot (\log n)^{\frac{1}{2}} \quad \text{and} \quad C_1 \cdot n^{\frac{2m}{5m-2}} \leq r_n \leq C_2 \cdot (\log n)^{\frac{1}{2}} \cdot n^2,
\]
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 $\pi_d^*$.

Theorem 2. Suppose Assumptions 2 and 3 hold. Then
\[
\| \hat{g}_d - g_d^* \|_{L^2(dF_X)} = O_p\left( \max \left\{ v_{K,m}^r \cdot r_n^{1+\frac{1}{p}}, \sqrt{r_n \cdot p \cdot \log n} \right\} \right) = o_p\left(n^{-1/4}\right),
\]
and
\[
\| \hat{\pi}_d - \pi_d^* \|_{L^2(dF_X)} = O_p\left( \max \left\{ v_{K,m}^r \cdot r_n^{1+\frac{1}{p}}, \sqrt{r_n \cdot p \cdot \log n} \right\} \right) = o_p\left(n^{-1/4}\right),
\]
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 $n^{-2s/(2d+1)}$ (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, in the target function class considered in Bauer and Kohler (2019) and Schmidt-Hieber (2020) differs 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 $\mathcal{F}_d(x, \beta_d^*) := \mathbb{E}[\mathcal{F}_d(Y_i(d) - \beta_d^*)|X_i = x]$, $\mathcal{L}_d(x, \beta_d^*) := \mathcal{F}_d(x, \beta_d^*)/\pi_d^*(x)$, and $\mathcal{g}(\theta, \epsilon_n) := (1-\epsilon_n) \cdot \theta + \epsilon_n \cdot (u_{\theta} + g_{\theta})$ be a local alternative value around $g_{\theta} \in \theta_n$. The directional derivative of $\mathcal{L}_d(D_{d}, X; g_{\theta})$ is given by
\[
\frac{\partial}{\partial g_{\theta}} \mathcal{L}_d(D_{d}, X; g_{\theta}) = \lim_{u \to 0} \frac{\mathcal{L}_d(D_{d} + u \cdot g_{\theta}; g_{\theta}) - \mathcal{L}_d(D_{d}; g_{\theta})}{u} = \{D_{d} - L(g_{\theta}(X))\} u(X), \quad u \in L^2(dF_X).
\]

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

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

Assumption 5. (Approximation error) We assume the following conditions hold:
used loss functions such as large numbers, establishing stochastic equicontinuity and weak convergence, see Chen et al. (2008). Again, they are satisfied by widely.

Assumption 4 (iii) is a smoothness condition for approximating the functions and stochastic equicontinuity in neural network space, which is needed for establishing Lemma in the supplemental material. Such a

Consequently, \( V^{-1/2} \cdot \sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, I_{(J+1)\times(J+1)}) \), where \( I_{(J+1)\times(J+1)} \) is the \((J+1) \times (J+1)\) identity matrix, \( V = H^{-1}E[S S^\top]H^{-1} \), \( H = \text{Diag}(H_0, \ldots, H_I) \) and \( S = (S_0, \ldots, S_I) \). 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 \( \hat{\beta} \) 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\( w_0^*(x) := f_0(x)/f_0(D_0|x,d) \), \( z_0^*(x) = E[Y|X = x, D = d] \), and \( E(z_d, \varepsilon_n) := (1 - \varepsilon_n) \cdot z_d + \varepsilon_n \cdot (w_0^* + z_d^*) \) be a local alternative value around \( z_d \in \mathcal{D}_n \).

**Assumption 7.** For every \( d \in \mathcal{D} = \{0, 1, \ldots, J\} \) and \( m \geq 1 \), we assume \( z_d^*(\cdot) \in \mathcal{D}_m^* \).

**Assumption 8.** (i) We assume
\[
\max \left\{ \frac{v_{z_d,m} \cdot r_n}{\sqrt{r_n \cdot p \cdot \log n}}, \sqrt{\frac{r_n \cdot p \cdot \log n}{n}} \right\} = o\left(n^{-\frac{1}{2}}\right).
\]
(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) \( \mathbb{P}(D_d = 1) \in (0, 1) \) and \( \pi_d^*(X) \in (0, 1) \); (ii) There exists a constant \( \varepsilon \) such that
\[
\mathbb{E}\left[\left\{w_0^*(X)\right\}^2\right] = \mathbb{E}\left[\left\{f_0(X)/f_0(D_0|x,d)\right\}^2\right] < \infty.
\]

**Assumption 10.** (Approximation error) We assume the following conditions hold:
\[
\sup_{\omega \in \mathcal{B} \setminus \{0\}} \|\text{Proj}_{\mathcal{B} \setminus \{0\}}(z_d, \varepsilon_n) - \bar{z}(z_d, \varepsilon_n)\|_{\mathcal{L}^2(\mathcal{D}_0)} = O\left(\frac{\varepsilon_n^2}{\lambda^2}\right), \text{ and}
\]
\[
\sup_{\omega \in \mathcal{B} \setminus \{0\}} \frac{1}{n} \sum_{i=1}^{n} (z_d, \varepsilon_n)(X_i) - \text{Proj}_{\mathcal{B} \setminus \{0\}}(z_d, \varepsilon_n)(X_i) = O_p(\varepsilon_n^2),
\]
where \( \text{Proj}_{\mathcal{B} \setminus \{0\}}(z_d, \varepsilon_n) \) denotes the \( L^2(\mathcal{D}_0) \)-projection of \( \mathbb{P}(z_d, \varepsilon_n) \) on the ANN space \( \mathcal{D}_n \).

**Assumption 11.** \( \sup_{\omega \in \mathcal{B}} \mathbb{E}[\left\{Y^*(d)\right\}^2|X = x] < \infty \) 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 \( d \in \{0, 1, \ldots, J\} \), we have \( \beta_d^\text{OR} \sim \beta_d^\ast \) and
\[
\sqrt{n} (\hat{\beta}_d - \beta_d^\ast) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_d^\text{OR} \left(Y, D_d, X_i; \beta_d^\ast\right) + o_P(1),
\]
where
\[
S_d^\text{OR} = S_d^\ast \left(Y, D_d, X_i; \beta_d^\ast\right) = \frac{D_d}{\pi_d^\ast(X_i)} Y_i - \left(\frac{D_d}{\pi_d^\ast(X_i)} - \pi_d^\ast(X_i)\right) \cdot z_d^*(X) - E[z_d^*(X)].
\]
Consequently,
\[
\left\{V_d^\text{OR}\right\}^{-1/2} \cdot \sqrt{n} \left(\hat{\beta}_d - \beta_d^\ast\right) \overset{d}{\rightarrow} \mathcal{N}(0, I_{(J+1)\times(J+1)}),
\]
where \( I_{(J+1)\times(J+1)} \) is the \((J + 1) \times (J + 1)\) identity matrix, \( V_d^\text{OR} = E[S_d^\text{OR} \cdot (S_d^\ast)^\top] \), and \( S_d^\text{OR} = (S_d^\text{OR}, \ldots, S_d^\text{OR})^\top \). The proof of Theorem 4 is provided in Supplement E.

Note that \( 1/\pi_d^\ast(X) = w_0^*(X)/\mathbb{P}(D_d = 1) \), with Assumptions 9 and 11, we have that the asymptotic variance is finite, which implies the proposed estimator \( \hat{\beta}_d^\text{OR} \) is \( \sqrt{n} \)-consistent. Moreover, the ANN-OR estimator \( \hat{\beta}_d^\text{OR} \) has the same asymptotic variance as the ANN-IPW estimator \( \hat{\beta} \) when \( \mathcal{D}(\nu) = \nu^2 \) 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 \( \beta^\ast \). 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 \( \{a_d, \ldots, a_{Jd}\} \) be i.i.d. positive random weights that are independent of data satisfying \( \mathbb{E}[a_d] = 1 \) and \( \text{Var}(a_d) = 1 \), where \( d \in \{0, 1, \ldots, J\} \). The weighted bootstrap estimator of the inverse logistic PS \( g_d^\ast \) is defined by satisfying

\[
L^B_{d,n}(g_d) \geq \sup_{g \in S_d} L^B_{d,n}(g_d) - O(\epsilon^2),
\]

where \( L^B_{d,n}(g_d) := n^{-1} \sum_{i=1}^n a_d i \{\mathbb{E}[\mathbb{I} \{Y \leq \beta\} | D_{di}, X_i; g_d(\cdot)] \} \) is the bootstrapped empirical criterion, and \( \epsilon = o(n^{-1/2}) \). Let \( \hat{g}_d^B := L(\hat{g}_d^B) \). Then the weighted bootstrap IPW estimator of \( \beta_d^\ast \) is given by

\[
\hat{\beta}_d^B = \arg\min_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^n a_d \mathbb{E}_i \mathbb{I} \{Y \leq \beta\}, \quad d \in \{0, 1, \ldots, J\}.
\]

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

\[
L^B_{d,n}(z_d) \geq \sup_{z \in \mathbb{R}} L^B_{d,n}(z_d) - O(\epsilon^2),
\]

where \( I^B_{d,n}(z_d) := n^{-1} \sum_{i=1}^n a_d i \{\mathbb{E}[\mathbb{I} \{Y \leq \beta\} | D_{di}, X_i; z_d(\cdot)] \} \). Then the weighted bootstrap OR estimator of \( \mathbb{E}[Y^*(d)] \) is given by

\[
\hat{\beta}_d^B = \frac{1}{n} \sum_{i=1}^n a_d \mathbb{E}_i \mathbb{I} \{Y \leq \beta\}, \quad d \in \{0, 1, \ldots, J\}.
\]

Let \( \tilde{\beta}^B := (\hat{\beta}_0^B, \ldots, \hat{\beta}_J^B)^\top \) and \( \hat{\beta}^{OR,B} := (\hat{\beta}_0^{OR,B}, \ldots, \hat{\beta}_J^{OR,B})^\top \). The following theorem justifies the validation of the proposed bootstrap inference.

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

\[
V^{-1/2} \cdot \sqrt{n} (\hat{\beta}^B - \beta^B) \xrightarrow{d} \mathcal{N}(0, I_{(J+1) \times (J+1)}).
\]

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

\[
\left\{ \mathbb{E}[Y^*(d)] \right\}^{-1/2} \cdot \sqrt{n} (\hat{\beta}^{OR,B} - \beta^{OR,B}) \xrightarrow{d} \mathcal{N}(0, I_{(J+1) \times (J+1)}).
\]

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 \( \beta_d^\ast \) can be obtained by solving the estimated efficient score function (Tsiatis, 2007). For example, when the loss function \( \mathcal{L}(v) = v^2 \) corresponding to ATE, the EIF of \( \beta_d^\ast = \mathbb{E}[Y^*(d)] \) is

\[
\frac{D_{di}}{\mathbb{E}_{\pi_d^\ast}(X_i) Y_i} - \left\{ \frac{D_{di}}{\mathbb{E}_{\pi_d^\ast}(X_i)} - 1 \right\} \mathbb{E}[Y | D_{di} = 1, X_i] = \beta_d^\ast.
\]

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

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

\[
\sum_{i=1}^n \left\{ \frac{D_{di}}{\mathbb{E}_{\pi_d}(X_i)} (\tau - \mathbb{E}[Y \leq \beta]) - \left\{ \frac{D_{di}}{\mathbb{E}_{\pi_d}(X_i)} - 1 \right\} (\tau - \mathbb{E}[\mathbb{I} \{Y \leq \beta\} | D_{di} = 1, X_i = x]) \right\} = 0,
\]

where \( \tilde{\pi}_d(x) \) and \( \mathbb{E}[\mathbb{I} \{Y \leq \beta\} | D_{di} = 1, X_i = x] \) are estimates of \( \pi_d(x) \) and \( \mathbb{E}[\mathbb{I} \{Y \leq \beta\} | D_{di} = 1, X_i = x] \), 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 \( \mathbb{E}[\mathbb{I} \{Y \leq \beta\} | D_{di} = 1, X_i = x] \), as it intertwines with the unknown quantile TE parameter \( \beta \) 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

$$\hat{\beta}_d := \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathcal{J}(Y_i(d) - \beta_i)]$$

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

- $\mathcal{J}(\psi) = \psi^2$, then $\hat{\beta}_{0,d} = \mathbb{E}[Y_i(d) | D_i = d]$ is the average treatment effects on the treated.
- $\mathcal{J}(\psi) = \psi(1 - \psi \leq 0)$, then $\hat{\beta}_{r,d} = \mathbb{E}[Y_i(d) | \psi(d)]$ is the $r$th quantile of $Y_i(d)$ conditioned on the treated group $\{D_i = d\}$.

Under Assumption 1, using the property of conditional expectation, the parameter of interest $\hat{\beta}_d$ is identified by

$$\hat{\beta}_d := \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathcal{J}(Y_i(d) - \beta_i)]$$

$$= \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathcal{J}(Y_i(d) - \beta_i) | \mathcal{J}(Y_i(d) - \beta_i) | X_i]$$

$$= \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\hat{\beta}_d(X_i) \cdot \mathcal{J}(Y_i(d) - \beta_i) | \mathcal{J}(Y_i(d) - \beta_i) | X_i\right]$$

$$= \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\hat{\beta}_d(X_i) | \mathcal{J}(Y_i(d) - \beta_i) | X_i\right]$$

where $p_d = \mathbb{P}(D_i = d)$. The estimator of $\hat{\beta}_d$ is obtained by minimizing the empirical analogue of the above equation:

$$\hat{\beta}_d = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} D_{i,d} \hat{\beta}_d(X_i) \mathcal{J}(Y_i(d) - \beta_i) / \sum_{i=1}^{n} D_{i,d}$$

where $\hat{\beta}_d$ is the ANN estimator of $\beta_d$. The estimator of $\beta_{0,d}$ for $d \in \{0, 1, ..., J\}$ can be defined as

$$\hat{\beta}_{0,d} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} D_{i,d} \hat{\beta}_d(X_i) \mathcal{J}(Y_i(d) - \beta_i)$$

Similar to the proof of Theorem 3 we obtain the following result for $\hat{\beta}_d$.

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

$$\sqrt{n} \left( \hat{\beta}_{d,d} - \beta_{d,d} \right) = H_{d,d}' \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_{d,d}(X_i, D_{i,d}, Y_i, \beta_{d,d}) + o_P(1),$$

where $H_{d,d} = -\partial_{\beta} \mathbb{E}[\mathcal{J}(Y_i(d) - \beta_i)]$ and

$$S_{d,d}(X_i, D_{i,d}, \beta_{d,d}) := D_{i,d} \mathcal{J}(Y_i(d) - \beta_i) - \left\{ \frac{D_{i,d}}{\hat{\beta}_d(X_i)} \mathcal{J}(Y_i(d) - \beta_i) \right\} \mathbb{E}_{\beta} \left( Y_i, \beta_{d,d} \right).$$

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 $p = 5$.

<table>
<thead>
<tr>
<th>n</th>
<th>IPW</th>
<th>OR</th>
<th>DR</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bias</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0008</td>
<td>0.0010</td>
<td>0.0012</td>
<td>0.0026</td>
</tr>
<tr>
<td></td>
<td>0.0758</td>
<td>0.0713</td>
<td>0.0726</td>
<td>0.0668</td>
</tr>
<tr>
<td></td>
<td>0.0701</td>
<td>0.0701</td>
<td>0.0715</td>
<td>0.0701</td>
</tr>
<tr>
<td></td>
<td>0.9275</td>
<td>0.9425</td>
<td>0.9425</td>
<td>0.9325</td>
</tr>
<tr>
<td></td>
<td>0.9375</td>
<td>0.9475</td>
<td>0.9475</td>
<td>0.9600</td>
</tr>
<tr>
<td></td>
<td>0.0007</td>
<td>0.0009</td>
<td>0.0010</td>
<td>0.0007</td>
</tr>
<tr>
<td></td>
<td>0.0524</td>
<td>0.0499</td>
<td>0.0502</td>
<td>0.0523</td>
</tr>
<tr>
<td></td>
<td>0.0488</td>
<td>0.0488</td>
<td>0.0370</td>
<td>0.0493</td>
</tr>
<tr>
<td></td>
<td>0.9375</td>
<td>0.9500</td>
<td>0.9475</td>
<td>0.9475</td>
</tr>
<tr>
<td></td>
<td>0.9475</td>
<td>0.9500</td>
<td>0.9475</td>
<td>0.9600</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>0.0334</td>
<td>0.0312</td>
<td>0.0305</td>
<td>0.0322</td>
</tr>
<tr>
<td></td>
<td>0.0399</td>
<td>0.0309</td>
<td>0.0307</td>
<td>0.0306</td>
</tr>
<tr>
<td></td>
<td>0.0331</td>
<td>0.0308</td>
<td>0.0305</td>
<td>0.0305</td>
</tr>
<tr>
<td></td>
<td>0.9475</td>
<td>0.9575</td>
<td>0.9525</td>
<td>0.9525</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>emp_sd</th>
<th>est_sd</th>
<th>cover_rate</th>
<th>cover_rate_boot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0584</td>
<td>0.0594</td>
<td>0.0523</td>
<td>0.0562</td>
</tr>
<tr>
<td></td>
<td>0.0830</td>
<td>0.0844</td>
<td>0.0789</td>
<td>0.0715</td>
</tr>
<tr>
<td></td>
<td>0.0511</td>
<td>0.0483</td>
<td>0.0715</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8750</td>
<td>0.6700</td>
<td>0.6400</td>
<td>0.9250</td>
</tr>
<tr>
<td></td>
<td>0.0901</td>
<td>0.0922</td>
<td>0.0821</td>
<td>0.0853</td>
</tr>
<tr>
<td></td>
<td>0.0888</td>
<td>0.0905</td>
<td>0.0950</td>
<td>0.9375</td>
</tr>
<tr>
<td></td>
<td>0.0578</td>
<td>0.0585</td>
<td>0.0495</td>
<td>0.0425</td>
</tr>
<tr>
<td></td>
<td>0.0675</td>
<td>0.0620</td>
<td>0.0634</td>
<td>0.0563</td>
</tr>
<tr>
<td></td>
<td>0.0652</td>
<td>0.0637</td>
<td>0.0371</td>
<td>0.0370</td>
</tr>
<tr>
<td></td>
<td>0.8175</td>
<td>0.8100</td>
<td>0.6250</td>
<td>0.6500</td>
</tr>
<tr>
<td></td>
<td>0.0679</td>
<td>0.0665</td>
<td>0.0612</td>
<td>0.0645</td>
</tr>
<tr>
<td></td>
<td>0.8275</td>
<td>0.8400</td>
<td>0.8400</td>
<td>0.8575</td>
</tr>
<tr>
<td></td>
<td>0.0558</td>
<td>0.0545</td>
<td>0.0407</td>
<td>0.0225</td>
</tr>
<tr>
<td></td>
<td>0.0390</td>
<td>0.0362</td>
<td>0.0376</td>
<td>0.0335</td>
</tr>
<tr>
<td></td>
<td>0.0413</td>
<td>0.0244</td>
<td>0.0258</td>
<td>0.0307</td>
</tr>
<tr>
<td></td>
<td>0.0402</td>
<td>0.0355</td>
<td>0.0365</td>
<td>0.0337</td>
</tr>
<tr>
<td></td>
<td>0.7350</td>
<td>0.7225</td>
<td>0.7850</td>
<td>0.8725</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>emp_sd</th>
<th>est_sd</th>
<th>cover_rate</th>
<th>cover_rate_boot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0584</td>
<td>0.0594</td>
<td>0.0523</td>
<td>0.0562</td>
</tr>
<tr>
<td></td>
<td>0.0830</td>
<td>0.0844</td>
<td>0.0789</td>
<td>0.0715</td>
</tr>
<tr>
<td></td>
<td>0.0511</td>
<td>0.0483</td>
<td>0.0715</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8750</td>
<td>0.6700</td>
<td>0.6400</td>
<td>0.9250</td>
</tr>
<tr>
<td></td>
<td>0.0901</td>
<td>0.0922</td>
<td>0.0821</td>
<td>0.0853</td>
</tr>
<tr>
<td></td>
<td>0.0888</td>
<td>0.0905</td>
<td>0.0950</td>
<td>0.9375</td>
</tr>
<tr>
<td></td>
<td>0.0578</td>
<td>0.0585</td>
<td>0.0495</td>
<td>0.0425</td>
</tr>
<tr>
<td></td>
<td>0.0675</td>
<td>0.0620</td>
<td>0.0634</td>
<td>0.0563</td>
</tr>
<tr>
<td></td>
<td>0.0652</td>
<td>0.0637</td>
<td>0.0371</td>
<td>0.0370</td>
</tr>
<tr>
<td></td>
<td>0.8175</td>
<td>0.8100</td>
<td>0.6250</td>
<td>0.6500</td>
</tr>
<tr>
<td></td>
<td>0.0679</td>
<td>0.0665</td>
<td>0.0612</td>
<td>0.0645</td>
</tr>
<tr>
<td></td>
<td>0.8275</td>
<td>0.8400</td>
<td>0.8400</td>
<td>0.8575</td>
</tr>
<tr>
<td></td>
<td>0.0558</td>
<td>0.0545</td>
<td>0.0407</td>
<td>0.0225</td>
</tr>
<tr>
<td></td>
<td>0.0390</td>
<td>0.0362</td>
<td>0.0376</td>
<td>0.0335</td>
</tr>
<tr>
<td></td>
<td>0.0413</td>
<td>0.0244</td>
<td>0.0258</td>
<td>0.0307</td>
</tr>
<tr>
<td></td>
<td>0.0402</td>
<td>0.0355</td>
<td>0.0365</td>
<td>0.0337</td>
</tr>
<tr>
<td></td>
<td>0.7350</td>
<td>0.7225</td>
<td>0.7850</td>
<td>0.8725</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>cover_rate</th>
<th>cover_rate_boot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.9250</td>
<td>0.9325</td>
</tr>
<tr>
<td>2000</td>
<td>0.9500</td>
<td>0.9500</td>
</tr>
<tr>
<td>5000</td>
<td>0.9500</td>
<td>0.9500</td>
</tr>
</tbody>
</table>
Table 2
The summary statistics of the estimated ATEs for Model 1 with $p = 10$.

<table>
<thead>
<tr>
<th></th>
<th>IPW</th>
<th></th>
<th>OR</th>
<th></th>
<th>DR</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ANN</td>
<td>GLM</td>
<td>GAM</td>
<td>RF</td>
<td>GBM</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0010</td>
<td>0.0635</td>
<td>0.0655</td>
<td>0.0574</td>
<td>0.0536</td>
<td>0.0010</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.0855</td>
<td>0.0900</td>
<td>0.1127</td>
<td>0.0887</td>
<td>0.0915</td>
<td>0.0859</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.0789</td>
<td>0.0943</td>
<td>0.0982</td>
<td>0.0547</td>
<td>0.0512</td>
<td>0.0784</td>
</tr>
<tr>
<td>cover_rate</td>
<td>0.9250</td>
<td>0.9075</td>
<td>0.8675</td>
<td>0.6725</td>
<td>0.6500</td>
<td>0.9275</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.9350</td>
<td>0.8950</td>
<td>0.8825</td>
<td>0.8925</td>
<td>0.8950</td>
<td>0.9375</td>
</tr>
<tr>
<td>$n = 2000$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0034</td>
<td>0.0659</td>
<td>0.0661</td>
<td>0.0584</td>
<td>0.0462</td>
<td>0.0002</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.0597</td>
<td>0.0660</td>
<td>0.0692</td>
<td>0.0634</td>
<td>0.0629</td>
<td>0.0657</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.0532</td>
<td>0.0668</td>
<td>0.0667</td>
<td>0.0397</td>
<td>0.0382</td>
<td>0.0539</td>
</tr>
<tr>
<td>cover_rate</td>
<td>0.9250</td>
<td>0.8275</td>
<td>0.8225</td>
<td>0.6050</td>
<td>0.6400</td>
<td>0.9250</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.9475</td>
<td>0.8250</td>
<td>0.8250</td>
<td>0.8125</td>
<td>0.8375</td>
<td>0.9425</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0015</td>
<td>0.0679</td>
<td>0.0670</td>
<td>0.0565</td>
<td>0.0362</td>
<td>0.0012</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.0363</td>
<td>0.0439</td>
<td>0.0439</td>
<td>0.0416</td>
<td>0.0419</td>
<td>0.0357</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.0322</td>
<td>0.0422</td>
<td>0.0418</td>
<td>0.0258</td>
<td>0.0267</td>
<td>0.0319</td>
</tr>
<tr>
<td>cover_rate</td>
<td>0.9275</td>
<td>0.6100</td>
<td>0.5975</td>
<td>0.4225</td>
<td>0.6150</td>
<td>0.9250</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.9368</td>
<td>0.0441</td>
<td>0.0443</td>
<td>0.0420</td>
<td>0.0423</td>
<td>0.0349</td>
</tr>
<tr>
<td></td>
<td>0.9475</td>
<td>0.6550</td>
<td>0.6475</td>
<td>0.6575</td>
<td>0.7875</td>
<td>0.9450</td>
</tr>
</tbody>
</table>
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):

\[
\begin{align*}
\logit\{E(D_i|X_i)\} & = 0.5(X_{1i}'X_{2i} - 0.7 \sin((X_{1i} + X_{d})) (X_{2i} - 0.2) - 0.1), \\
E(Y_i(1)|X_i) & = E(Y_i|X_i, D_i = 1) = 0.3(X_{1i} - 0.9)^2 + 0.1(X_{2i} - 0.5)^2 - 0.6X_{1i}'X_{3i} + \sin(-1.7(X_{1i}' + X_{3i}') - 1.1) + X_{3i}X_{2i} + 1, \\
E(Y_i(0)|X_i) & = E(Y_i|X_i, D_i = 0) = 0.64(X_{1i} - 0.9)^2 + 0.16(X_{2i} + 0.2)^2 - 0.6X_{1i}'X_{3i} + \sin(-1.7(X_{1i}' + X_{3i}') - 1.1) + X_{3i}X_{2i} - 1;
\end{align*}
\]

where \(X_{qi} = \sum_{j=p-q}^{q/5} X_{ij} \) for \(1 \leq j \leq 5.1 \leq i \leq n\), and \(Y_i(d) = E(Y_i|X_i) + \epsilon_i, d = \{0, 1\}, \epsilon_i \sim N(0, 1)\) for \(1 \leq i \leq n\).

We use the confounders from \(X_i = 2(F(Z_i) - 0.5)\), where \(Z_i = (Z_{i1}, ..., Z_{ip})' \sim N(0, \Sigma)\), \(\Sigma = \{\sigma_{kk}\}, \sigma_{kk} = 0.2^{k-k}\) for \(1 \leq k, \ k' \leq p\), and \(F(.)\) is the cumulative distribution function of the standard normal for \(1 \leq i \leq n, 1 \leq j \leq p\). Let \(\text{cp}=1\). We partition the confounders into 5 subgroups, and \(X_{qi} = \) the average of the \(p/5\) confounders in the \(j\)th subgroup for \(j = 1, ..., 5\), 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 \(X_{qi} = 2(F(Z_{qi}|\sigma_i) - 0.5)\), where \(\sigma_i\) is the standard deviation of \(Z_i\), and \(Z_i = (Z_{i1}, ..., Z_{ip})'\) are generated from two designs.

- Design 1 (factor model): \(Z_{qi} = F_{i1}'L_i + \eta_{qi}\), where \(F_{i1} \sim N(d, \Sigma_i)\), \(\Sigma_i = \{\sigma_{kk}\}, \sigma_{kk} = 0.2^{k-k}\) for \(1 \leq k, \ k' \leq 10\), \(L_i\) is a constant vector kept fixed for each realization and is generated from \(L_i \sim N(d, \Sigma_i)\) and \(\eta_{qi} \sim N(0, 1)\). Let \(\text{cp}=7\).
- Design 2 (multivariate normal): \(Z_{qi} \sim N(d, \Sigma)\), \(\Sigma = \{\sigma_{kk}\}, \sigma_{kk} = 0.2^{a(k-k)}\) for \(1 \leq k, \ k' \leq p\), where \(m(x) = [x/10]\), and \([a]\) denotes the smallest integer no less than \(a\). Let \(\text{cp}=4\).

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.

| n=1000 | Q1 | | | | | | Q2 | | | | | | Q3 | | | | | |
| | ANN | GLM | GAM | RF | GBM | DNN | | ANN | GLM | GAM | RF | GBM | DNN | | ANN | GLM | GAM | RF | GBM | DNN | |
| 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 | 0.0545 | 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 | 0.0542 | 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.

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ANN</td>
<td>GLM</td>
<td>GAM</td>
</tr>
<tr>
<td>n=1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0019</td>
<td>0.0717</td>
<td>0.0750</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.1585</td>
<td>0.1345</td>
<td>0.1634</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.2530</td>
<td>0.1444</td>
<td>0.4770</td>
</tr>
<tr>
<td>cover_rate</td>
<td>0.9675</td>
<td>0.9200</td>
<td>0.9590</td>
</tr>
<tr>
<td>est_sd_boot</td>
<td>0.1623</td>
<td>0.1345</td>
<td>0.1552</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.9325</td>
<td>0.9150</td>
<td>0.9125</td>
</tr>
<tr>
<td>n=2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0030</td>
<td>0.0663</td>
<td>0.0678</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.1070</td>
<td>0.1040</td>
<td>0.1083</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.1057</td>
<td>0.1021</td>
<td>0.1149</td>
</tr>
<tr>
<td>cover_rate</td>
<td>0.9425</td>
<td>0.8775</td>
<td>0.8850</td>
</tr>
<tr>
<td>est_sd_boot</td>
<td>0.1072</td>
<td>0.1043</td>
<td>0.1097</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.9450</td>
<td>0.8800</td>
<td>0.8000</td>
</tr>
<tr>
<td>n=5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0065</td>
<td>0.0774</td>
<td>0.0764</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.0657</td>
<td>0.0682</td>
<td>0.0680</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.0629</td>
<td>0.0644</td>
<td>0.0638</td>
</tr>
<tr>
<td>cover_rate</td>
<td>0.9275</td>
<td>0.7675</td>
<td>0.7725</td>
</tr>
<tr>
<td>est_sd_boot</td>
<td>0.0655</td>
<td>0.0686</td>
<td>0.0678</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.9325</td>
<td>0.7825</td>
<td>0.7875</td>
</tr>
</tbody>
</table>
GLM is expected to have the best performance. However, we can see from Tables 5 that 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. Moreover, 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, 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.
The summary statistics of the estimated QTEs and QTTs by the IPW method for Model 1 with $p = 100$ and $n = 2000$.

<table>
<thead>
<tr>
<th>Design 1</th>
<th>Design 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATE</td>
<td>ATT</td>
</tr>
<tr>
<td>IPW</td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0104</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.0983</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.0512</td>
</tr>
<tr>
<td>cover_rate</td>
<td>0.8225</td>
</tr>
<tr>
<td>est_sd_boot</td>
<td>0.0694</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.8650</td>
</tr>
<tr>
<td>OR</td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0097</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.0892</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.0512</td>
</tr>
<tr>
<td>cover_rate</td>
<td>0.8350</td>
</tr>
<tr>
<td>est_sd_boot</td>
<td>0.0625</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.8775</td>
</tr>
<tr>
<td>DR</td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0101</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.0894</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.0521</td>
</tr>
<tr>
<td>cover_rate</td>
<td>0.8375</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Design 1</th>
<th>Design 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTE</td>
<td>QTT</td>
</tr>
<tr>
<td>ANN</td>
<td>ANN-PCA</td>
</tr>
<tr>
<td>Q1</td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0157</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.1379</td>
</tr>
<tr>
<td>est_sd_boot</td>
<td>0.1305</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.9505</td>
</tr>
<tr>
<td>Q2</td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0131</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.1256</td>
</tr>
<tr>
<td>est_sd_boot</td>
<td>0.1104</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.9225</td>
</tr>
<tr>
<td>Q3</td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.0265</td>
</tr>
<tr>
<td>emp_sd</td>
<td>0.1259</td>
</tr>
<tr>
<td>est_sd_boot</td>
<td>0.1213</td>
</tr>
<tr>
<td>cover_rate_boot</td>
<td>0.9175</td>
</tr>
</tbody>
</table>

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 AN 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
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 \( \frac{\bar{x}_s - \bar{x}_n}{\sqrt{s^2_s/n_s + s^2_n/n_n}} \) for continuous variables, and \( \frac{(p_n - p_s)}{\sqrt{pq/n_n + pq/n_s}} \) for categorical variables, where \( \bar{x} \), \( s^2 \) and \( p \) denote sample mean, sample variance and sample proportion, and the subscripts \( n \) and \( s \) refer to nonsmokers and smokers respectively, and \( p \) and \( q \) 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 means 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 BMI increases.

### Table 7

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Non-smoker (( N_n=3288 ))</th>
<th>Smoker (( N_s=3359 ))</th>
<th>Std. Dif.</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender 1 = Male</td>
<td>1404 (41.8%)</td>
<td>2019 (61.4%)</td>
<td>− 15.99</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Gender 0 = Female</td>
<td>1955 (58.2%)</td>
<td>1269 (38.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Mean(SD)</td>
<td>48.97 (19)</td>
<td>51.73 (17.5%)</td>
<td>− 6.14</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Age 0 = No</td>
<td>1370 (40.79%)</td>
<td>1421 (43.22%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital 1 = Yes</td>
<td>1989 (59.21%)</td>
<td>1867 (56.78%)</td>
<td>2.01</td>
<td>0.0446</td>
</tr>
<tr>
<td>Marital 0 = No</td>
<td>1370 (40.79%)</td>
<td>1421 (43.22%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education 1 = College or above</td>
<td>1626 (48.41%)</td>
<td>1297 (39.45%)</td>
<td>7.36</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Education 0 = Less than college</td>
<td>1733 (51.59%)</td>
<td>1991 (60.55%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family PIR Mean(SD)</td>
<td>2.79 (1.63)</td>
<td>2.57 (1.6)</td>
<td>5.62</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Alcohol 1 = Yes</td>
<td>1897 (56.48%)</td>
<td>2708 (82.36%)</td>
<td>− 22.87</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Alcohol 0 = No</td>
<td>1462 (43.52%)</td>
<td>580 (17.64%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHSVIG 1 = Yes</td>
<td>1102 (32.81%)</td>
<td>908 (27.62%)</td>
<td>4.61</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>PHSVIG 0 = No</td>
<td>2257 (67.19%)</td>
<td>2380 (72.38%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHSMOD 1 = Yes</td>
<td>1491 (44.39%)</td>
<td>1376 (41.85%)</td>
<td>2.09</td>
<td>0.0366</td>
</tr>
<tr>
<td>PHSMOD 0 = No</td>
<td>1868 (55.61%)</td>
<td>1912 (58.15%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBP Mean(SD)</td>
<td>126.42 (21.04)</td>
<td>126.63 (19.98)</td>
<td>− 0.43</td>
<td>0.6684</td>
</tr>
<tr>
<td>DBP Mean(SD)</td>
<td>72.1 (13.56)</td>
<td>71.61 (14.1)</td>
<td>1.44</td>
<td>0.15</td>
</tr>
</tbody>
</table>

### Table 8

The estimates and standard errors of ATE and QTE.

<table>
<thead>
<tr>
<th></th>
<th>ATE</th>
<th>OR</th>
<th>QTE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IPW</td>
<td>OR</td>
<td>Q1</td>
</tr>
<tr>
<td>estimate</td>
<td>− 0.224</td>
<td>− 0.241</td>
<td>− 0.400</td>
</tr>
<tr>
<td>est_sd</td>
<td>0.154</td>
<td>0.154</td>
<td>0.157</td>
</tr>
<tr>
<td>z-value</td>
<td>− 1.454</td>
<td>− 1.564</td>
<td>− 2.547</td>
</tr>
<tr>
<td>p-value</td>
<td>0.073</td>
<td>0.058</td>
<td>0.005</td>
</tr>
<tr>
<td>est_sd_boot</td>
<td>0.162</td>
<td>0.149</td>
<td>0.156</td>
</tr>
<tr>
<td>z-value_boot</td>
<td>− 1.383</td>
<td>− 1.617</td>
<td>− 2.564</td>
</tr>
<tr>
<td>p-value_boot</td>
<td>0.083</td>
<td>0.053</td>
<td>0.005</td>
</tr>
</tbody>
</table>

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.
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 $\mathcal{X} = [0, 1]^p$ and $n_0 \in \mathbb{N}$, Lorentz (1966, Theorem 8) states that for any $n_0$-times continuously differentiable function $f(\cdot): \mathcal{X} \to \mathbb{R}$, i.e. $\sup_{t \in \mathcal{X}} |D^p f(x)| \leq 1$, there exist polynomials $P_{n_1 \cdots n_p}(x)$, of degree $n_i$ in $x_i$, such that

$$\sup_{x \in \mathcal{X}} |f(x) - P_{n_1 \cdots n_p}(x)| \leq C_p \cdot \sum_{i=1}^p \frac{1}{n_i^p},$$

where $C_p$ is a constant depending on $p$.

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

$$u_i(x) = 1, \quad u_2(x) = (1, x_1)^\top, \quad u_{p+1}(x) = (1, x_1, \ldots, x_p)^\top, \quad u_{p+2}(x) = (1, x_1, \ldots, x_p, x_1^2)^\top, \quad \ldots$$

To ensure all degrees of $(x_1, \ldots, x_p)$ get up to some order $n_0 \in \mathbb{N}$, i.e. $\min \{n_1, \ldots, n_p\} \geq n_0$, we require $K = (n_0 + 1)^p$. Therefore, for any function $f(\cdot)$ satisfying $\sup_{t \in \mathcal{X}} |D^p f(x)| \leq 1$, the approximation rate based on the polynomial sieve $\{u_k(x)\}$ is

$$\inf_{k \in \mathbb{N}} \sup_{x \in \mathcal{X}} |f(x) - h_k(x)| \leq C_p \cdot p \cdot K^{\frac{1}{2p}}.$$  

Note that for any function in the mixed smoothness ball defined in Section 2.1, i.e. $f(\cdot) \in \mathcal{W}^{m_1, \ldots, m_p}([0, 1]^p)$ for $\epsilon \in (0, 1)$, we only have $\sup_{t \in \mathcal{X}} |D^p f(x)| \leq 1$. In light of the compactness of $\mathcal{X}$, the $L^2(dF_x)$-approximation error based on the polynomial sieve $\{u_k(x)\}$ is $C_p \cdot p \cdot K^{\frac{1}{2p}}$, which severely suffers from the curse of dimensionality.

Appendix B. Proof of Theorem 1

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

$$\tilde{f}(t_1, t_2, \ldots, t_p) = \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} f(x_1, x_2, \ldots, x_p) e^{-it_1 x_1} \cdots e^{-it_p x_p} dx_1 dx_2 \cdots dx_p$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} f(x_1, x_2, \ldots, x_p) e^{-it_1 x_1 - it_2 x_2} \cdots e^{-it_p x_p} dx_1 dx_2 \cdots dx_p$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} f(x_1, x_2, \ldots, x_p) e^{-i(\frac{\pi}{n} t_1 x_1)} \cdots e^{-i(\frac{\pi}{n} t_p x_p)} dx_1 dx_2 \cdots dx_p$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} f\left(x_1 - \frac{\pi}{n} t_1, x_2, \ldots, x_p\right) e^{-i(\frac{\pi}{n} t_1 x_1)} \cdots e^{-i(\frac{\pi}{n} t_p x_p)} dx_1 dx_2 \cdots dx_p.$$

Then
Then for a vector of nonnegative integers \( \mathbf{s} \), for \( \mathbf{i} \in \mathbb{N} \) and a constant \( \epsilon \in (0, 1) \):

\[
\widehat{f}(t_1, t_2, \ldots, t_p) = \frac{1}{n} \sum_{i=1}^{p} (\Delta_{n_i}^e f(x)) \cdot f(x) \cdot \frac{1}{(n_i)}^{n_i+1} e^{-\epsilon n_i x_1} \ldots e^{-\epsilon n_i x_p} dx_{1} \ldots dx_{p}
\]

\[
\triangleq \frac{1}{n} \sum_{i=1}^{p} (\Delta_{n_i}^e f(x)) \cdot f(x) \cdot \frac{1}{(n_i)}^{n_i+1} e^{-\epsilon n_i x_1} \ldots e^{-\epsilon n_i x_p} dx_{1} \ldots dx_{p}
\]

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

\[
\widehat{f}(t_1, t_2, \ldots, t_p) = \frac{1}{n} \sum_{i=1}^{p} (\Delta_{n_i}^e f(x)) \cdot f(x) \cdot \frac{1}{(n_i)}^{n_i+1} e^{-\epsilon n_i x_1} \ldots e^{-\epsilon n_i x_p} dx_{1} \ldots dx_{p}
\]

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

\[
\widehat{f}(t_1, t_2, \ldots, t_p)
\]
Note that
\[
\lim_{t_1,\ldots,t_p \to \infty} \left( \Delta^\gamma_{t_1} \right)^{n_1+1} \cdots \left( \Delta^\gamma_{t_p} \right)^{n_p+1} f(x) = \partial_t^{n_1+1} \cdots \partial_t^{n_p+1} f(x),
\]
provided that the limit exists.

For any \( f(\cdot) \in \mathcal{Y}^{m+\infty}([0,1]^p) \), by definition we have
\[
\sup_{\{u | [u] \leq \omega \}} \sup_{\{v \in [0,1]^p | t_i > 0 \}} \left| \left( \frac{x}{t_1} \right)^{n_1+1} \cdots \left( \frac{x}{t_p} \right)^{n_p+1} \partial_x^n D^n f(x) \right| \leq 1.
\]
Then we can find a large enough constant \( M_0 > 0 \) such that \( \min_{\{u \in [0,1]^p | t_i \geq M_0 \}} \geq M_0 \), we have
\[
\sup_{\{u | [u] \leq \omega \}} \sup_{\{v \in [0,1]^p | x \}} \left| \left( \frac{x}{t_1} \right)^{n_1+1} \cdots \left( \frac{x}{t_p} \right)^{n_p+1} f(x) \right| \leq 2,
\]
namely,
\[
M_0 := \inf \left\{ C \in \mathbb{R} : \sup_{\{u | [u] \leq \omega \}} \sup_{\{v \in [0,1]^p | x \}} \left| \left( \frac{x}{t_1} \right)^{n_1+1} \cdots \left( \frac{x}{t_p} \right)^{n_p+1} f(x) \right| \leq 2 \}.
\]

Hence, by (B.1), for any \( f(\cdot) \in \mathcal{Y}^{m+\infty}([0,1]^p) \), we have that for any \( k \in \{1,\ldots,p\} \):
\[
\sup_{t_1,\ldots,t_p} \left| f(t_1,\ldots,t_k,\ldots,t_p) \right| I(t_i \geq M_0) \leq \left( \frac{2}{\pi^2} \right)^{m+2p} \frac{1}{|t_1|^{n_1+1+\epsilon} \cdots |t_p|^{n_p+1+\epsilon}}, \text{ where } \sum_{j=1}^p r_j \leq m.
\]
We emphasize that (B.4) holds for an arbitrarily fixed \( k \in \{1,\ldots,p\} \) and for all \( \{r_{k+1},\ldots,r_p\} : \sum_{j=k+1}^p r_j \leq m \).

We next find the bound for \( \nu_{t,m} \). Without loss of generality, we assume the function \( f(t_1,\ldots,t_p) \) is symmetric in \( t = (t_1,\ldots,t_p) \). Note that
\[ v_{f,m} = \int_{\mathbb{R}^p} \left| t_1 \right|^m \cdot \left| \bar{f}(t) \right| dt \]

\[ = \left( \int_{[0, M_0]} + \int_{[M_0, \infty]} \right) \left( \int_{[0, M_0]} + \int_{(M_0, \infty)} \right) \left\{ \sum_{k=1}^m |t_k| \right\}^m \cdot \bar{f}(t) dt_1 \cdots dt_p \]

\[ = \sum_{j=0}^p \left( \sum_{i=0}^m \left( a_1, \ldots, a_p \right) \left( \sum_{i=0}^p \left( \frac{p}{i} \right) \right) \right) \times \int_{[0, M_0]} \cdots \int_{[0, M_0]} \sum_{j=0}^m |t_k|^{m_j} \cdot \bar{f}(t) dt_1 \cdots dt_p \]

\[ \leq \sum_{j=0}^p \left( \sum_{i=0}^m \left( a_1, \ldots, a_p \right) \left( \sum_{i=0}^p \left( \frac{p}{i} \right) \right) \right) \cdot M_0^{a_i} \cdots M_0^{a_p} \times \int_{[0, M_0]} \cdots \int_{[0, M_0]} |t_k|^{m_k+1} \cdot \sup_{t_1, \ldots, t_p} \left| \bar{f}(t_1, \ldots, t_i, \ldots, t_p) \right| dt_1 \cdots dt_p. \]

For every \( i \in \{0, 1, \ldots, p\} \) and every \( (a_1, \ldots, a_p) \) in the summand, since \( a_1 + \cdots + a_p \leq m \), by applying (B.4) we have

\[ \sup_{t_j \geq \alpha_j} \left| \bar{f}(t_1, \ldots, t_i, \ldots, t_p) \right| \leq \left( \frac{\pi}{2} \right)^{m+2p} \cdot \frac{2}{(2\pi)^p} \cdot \frac{1}{|t_k|^{m_k+1+\epsilon}}. \]

Then we have

\[ v_{f,m} \leq \sum_{j=0}^p \left( \sum_{i=0}^m \left( a_1, \ldots, a_p \right) \left( \sum_{i=0}^p \left( \frac{p}{i} \right) \right) \right) \cdot M_0^{a_i} \cdots M_0^{a_p} \cdot \left( \frac{\pi}{2} \right)^{m+2p} \cdot \frac{2}{(2\pi)^p} \cdot |t_k|^{m_k+1+\epsilon} \]

\[ \leq \sum_{j=0}^p \left( \sum_{i=0}^m \left( a_1, \ldots, a_p \right) \left( \sum_{i=0}^p \left( \frac{p}{i} \right) \right) \right) \cdot M_0^{a_i} \cdots M_0^{a_p} \cdot \left( \frac{\pi}{2} \right)^{m+2p} \cdot \frac{2}{(2\pi)^p} \cdot \left( \frac{M_0}{\epsilon} \right)^p \]

\[ = \sum_{j=0}^p \left( \sum_{i=0}^m \left( a_1, \ldots, a_p \right) \left( \sum_{i=0}^p \left( \frac{p}{i} \right) \right) \right) \cdot M_0^{a_i} \cdots M_0^{a_p} \cdot \left( \frac{\pi}{2} \right)^{m+2p} \cdot \frac{2}{(2\pi)^p} \cdot \left( \frac{M_0}{\epsilon} \right)^p \]

\[ \leq \sum_{j=0}^p \left( \sum_{i=0}^m \left( a_1, \ldots, a_p \right) \left( \sum_{i=0}^p \left( \frac{p}{i} \right) \right) \right) \cdot M_0^{a_i} \cdots M_0^{a_p} \cdot \left( \frac{\pi}{2} \right)^{m+2p} \cdot \frac{2}{(2\pi)^p} \cdot \frac{M_0^{a_i+1+\epsilon}}{\epsilon} \]

\[ = \frac{\pi^{2p}}{2^{p-1}} \cdot M_0 \cdot \left( \frac{\pi}{2} \right)^{m+2p} \cdot \frac{2}{(2\pi)^p} \cdot \left( \frac{M_0}{\epsilon} \right)^p \]

\[ \leq 2 \cdot \left( \frac{M_0}{\epsilon} \right)^m \cdot \left( \frac{1}{2\epsilon} + \frac{1}{2} \right)^p \]

For some universal large constant \( M \) defined by

\[ M := M_0 \cdot \left( \frac{\pi}{2} \right) \cdot \left( \frac{1}{2\epsilon} + \frac{1}{2} \right). \]
Appendix C. Proof of Theorem 3

We first show $\hat{\beta}_d \overset{p}{\to} \beta_d$ for any $d \in \{0, 1, ..., J\}$. The details of the proof are given in the on-line supplemental materials. Next, we establish the asymptotic normality for $\sqrt{n}(\hat{\beta}_d - \beta_d)$. Since the loss function $\mathcal{L}(\cdot)$ may not be smooth (e.g. $\mathcal{L}(\nu) = \nu(\tau - \mathbb{I}(\nu \leq 0))$ 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

$$G_{d, n}(\beta, \tilde{\pi}_d) := \frac{1}{n} \sum_{i=1}^{n} \frac{D_{\hat{\alpha}}}{\tilde{\pi}(X_i)} \mathcal{L}(Y_i - \beta).$$

(C.1)

By definition

$$\hat{\beta}_d = \arg\min_{\beta \in \Theta} G_{d, n}(\beta, \tilde{\pi}_d) = \arg\min_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \frac{D_{\hat{\alpha}}}{\tilde{\pi}(X_i)} \mathcal{L}(Y_i - \beta).$$

(C.2)

then

$$\hat{\beta}_d = \arg\min_{\beta \in \Theta} \left\{ G_{d, n}(\beta, \tilde{\pi}_d) - G_{d, n}(\beta_d^*, \tilde{\pi}_d) \right\} = \arg\min_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \frac{D_{\hat{\alpha}}}{\tilde{\pi}(X_i)} \left\{ \mathcal{L}(Y_i - \beta) - \mathcal{L}(Y_i - \beta_d^*) \right\}$$

$$= \arg\min_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \frac{D_{\hat{\alpha}}}{\tilde{\pi}(X_i)} \left[ - \mathcal{L}(Y_i - \beta_d^*) (\beta - \beta_d^*) + \left\{ \mathcal{L}(Y_i - \beta) - \mathcal{L}(Y_i - \beta_d^*) \mathcal{L}(Y_i - \beta_d^*) \right\} \right]$$

By using change of variables and defining the following functions:

$$\tilde{u}_d := \sqrt{n}(\hat{\beta}_d - \beta_d^*), \quad u := \sqrt{n}(\beta - \beta_d^*),$$

$$R_d(Y, u) := \mathcal{L}(Y_i - \left\{ \beta_d^* + \frac{u}{\sqrt{n}} \right\}) - \mathcal{L}(Y_i - \beta_d^*) + \mathcal{L}(Y_i - \beta_d^*) \cdot \frac{u}{\sqrt{n}},$$

$$Q_{d, n}(u, \tilde{\pi}_d) := \frac{1}{n} \sum_{i=1}^{n} \frac{D_{\hat{\alpha}}}{\tilde{\pi}(X_i)} \left[ - \mathcal{L}(Y_i - \beta_d^*) \cdot \frac{u}{\sqrt{n}} + R_d(Y_i, u) \right] = n \cdot [G_{d, n}(\beta, \tilde{\pi}_d) - G_{d, n}(\beta_d^*, \tilde{\pi}_d)].$$

Then we get

$$\tilde{u}_d = \arg\min_u Q_{d, n}(u, \tilde{\pi}_d).$$

Next, we define the following quadratic function

$$\tilde{Q}_{d, n}(u) := \frac{u}{\sqrt{n}} \sum_{i=1}^{n} \left[ - \frac{D_{\hat{\alpha}}}{\tilde{\pi}(X_i)} \mathcal{L}(Y_i - \beta_d^*) + \left( \frac{D_{\hat{\alpha}}}{\tilde{\pi}(X_i)} - 1 \right) \mathcal{L}(X_i, \beta_d^*) \right]$$

$$- \partial_\beta \mathbb{E}[\mathcal{L}(Y_i^* (d) - \beta_d^*)] \cdot \frac{u^2}{2},$$

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

$$\tilde{u}_d := \arg\min_u \tilde{Q}_{d, n}(u).$$

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

$$\tilde{u}_d = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} H_d^{-1} \cdot S_d(Y, D_{\hat{\alpha}}, X_i; \beta_d^*).$$

where

$$S_d(Y, D_{\hat{\alpha}}, X_i; \beta_d^*) := \frac{D_{\hat{\alpha}}}{\tilde{\pi}(X_i)} \mathcal{L}(Y_i - \beta_d^*) - \left( \frac{D_{\hat{\alpha}}}{\tilde{\pi}(X_i)} - 1 \right) \mathcal{L}(X_i, \beta_d^*)$$

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

---

X. Chen et al.  
Journal of Econometrics 238 (2024) 105555
The desired result can be obtained via the following steps:

- Step I: showing $\xi_{dn}(u, \tilde{x}_d) := Q_{dn}(u) - Q_{dn}(u, \tilde{x}_d) = o_p(1)$ for every fixed $u$;
- Step II: showing $|\tilde{u}_d - u_d| = o_p(1)$;
- Step III: obtaining the desired result: $\sqrt{n}(\beta_d - \beta_d^*) = \tilde{u}_d + \{\tilde{u}_d - u_d\} = n^{-1/2} \sum_{i=1}^{n} H_d^{-1} \cdot S_d(Y_i, D_i, X_i; \beta_d^*) + o_p(1)$.

Note that both objective functions $Q_{dn}(u)$ and $Q_{dn}(u, \tilde{x}_d)$ 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