DeepMed: Semiparametric Causal Mediation Analysis with Debiased Deep Learning

Siqi Xu
Department of Statistics and Actuarial Sciences
University of Hong Kong
Hong Kong SAR, China
sqxu@hku.hk

Lin Liu*
Institute of Natural Sciences, School of Mathematical Sciences,
and SJTU-Yale Joint Center for Biostatistics and Data Science
Shanghai, China
linliu@sjtu.edu.cn

Zhonghua Liu†
Department of Statistics and Actuarial Sciences
University of Hong Kong
Hong Kong SAR, China
zhhliu@hku.hk

Abstract

Causal mediation analysis can unpack the black box of causality and is therefore a powerful tool for disentangling causal pathways in biomedical and social sciences, and also for evaluating machine learning fairness. To reduce bias for estimating Natural Direct and Indirect Effects in mediation analysis, we propose a new method called DeepMed that uses deep neural networks (DNNs) to cross-fit the infinite-dimensional nuisance functions in the efficient influence functions. We obtain novel theoretical results that our DeepMed method (1) can achieve semiparametric efficiency bound without imposing sparsity constraints on the DNN architecture and (2) can adapt to certain low-dimensional structures of the nuisance functions, significantly advancing the existing literature on DNN-based semiparametric causal inference. Extensive synthetic experiments are conducted to support our novel theoretical findings. As a proof of concept, we apply DeepMed to analyze two real datasets on machine learning fairness and reach conclusions consistent with previous findings. The Appendix and replication files can be found in the supplementary materials.

1 Introduction

Tremendous progress has been made in this decade on deploying deep neural networks (DNNs) in real-world data science problems (Krizhevsky et al., 2012; Wolf et al., 2019; Jumper et al., 2021). Causal inference is no exception. In semiparametric causal inference, several recent works (Chen et al., 2020; Chernozhukov et al., 2020; Farrell et al., 2021) initiated the investigation of statistical

*Co-corresponding author
†Co-corresponding author

Preprint. Under review.

properties of causal effect estimators when the nuisance functions (the outcome regressions and propensity scores) are estimated by DNNs. However, there are several important limitations in the recent literature that need to be addressed before the theoretical results can be used to guide practice:

(1) Most recent works focus on total (treatment) effect (Chen et al., 2020; Farrell et al., 2021). In many settings, however, more intricate causal parameters are often of greater interests. In biomedical and social sciences, one is often interested in "mediation analysis" to decompose the total effect into direct and indirect effect to unpack the underlying black-box causal mechanism (Baron and Kenny, 1986). More recently, mediation analysis also percolated into machine learning fairness. For instance, in the context of predicting the recidivism risk, Nabi and Shpitser (2018) argued that, for a "fair" algorithm, sensitive features such as race should have no direct effect on the predicted recidivism risk. If such direct effects can be accurately estimated, one can detect the potential unfairness of a machine learning algorithm. We will revisit such applications in Section 5 and Appendix F.

(2) Statistical properties of DNN-based causal estimators in recent works mostly follow from several (recent) results on the convergence rates of DNN-based nonparametric regression estimators (Chen and White, 1999; Suzuki, 2019; Schmidt-Hieber, 2020; Tsuji and Suzuki, 2021), with the limitation of relying on *sparse* DNN architectures. The theoretical properties are in turn evaluated by relatively simple synthetic experiments not designed to generate nearly *infinite-dimensional* nuisance functions, a setting considered by almost all the above related works.

The above limitations raise the tantalizing question whether the available statistical guarantees for DNN-based causal inference have practical relevance. In this work, we plan to partially fill these gaps by developing a new method called DeepMed for semiparametric mediation analysis with DNNs. We focus on the *Natural Direct/Indirect Effects* (NDE/NIE) (Robins and Greenland, 1992; Pearl, 2001) (defined in Section 2.1). The DeepMed estimators leverage the "multiply-robust" property of the efficient influence function (EIF) of NDE/NIE (Tchetgen Tchetgen and Shpitser, 2012; Farbmacher et al., 2020) (see Theorem 1 in Section 2.2), together with the flexibility and superior predictive power of DNNs (see Section 3.1 and Algorithm 1). In particular, we also make the following novel contributions to deepen our understanding of DNN-based semiparametric causal inference:

- On the theoretical side, we obtain novel results that our DeepMed method can achieve semiparametric efficiency bound without imposing sparsity constraints on the DNN architecture and can adapt to certain low-dimensional structures of the nuisance functions (see Section 3.2), thus significantly advancing the existing literature on DNN-based semiparametric causal inference. *Non-sparse* DNN architecture is more commonly employed in practice, and the low-dimensional structures of nuisance functions can help avoid curse-of-dimensionality. These two points, taken together, make our theoretical results the state-of-the-art over those of Farrell et al. (2021).
- On the empirical side, in Section 4, we designed sophisticated synthetic experiments to simulate nearly *infinite-dimensional* functions, which are much more complex than those in previous related works (Chen et al., 2020; Farrell et al., 2021; Adcock and Dexter, 2021). We emphasize that these novel experiments could be of independent interest beyond causal mediation analysis or even causal inference; see Remark 9 for an extended discussion. As a proof of concept, in Section 5 and Appendix F, we also apply DeepMed to re-analyze two real-world datasets on algorithmic fairness and reach similar conclusions to related works. The smaller validation losses of DeepMed suggest smaller bias compared with competing methods.
- Finally, a user-friendly R package will be freely available after the review process. Making such resources accessible to the community helps enhance reproducibility, a highly recognized problem in all scientific disciplines, including machine learning (Pineau et al., 2021).

2 Definition, identification, and estimation of NDE and NIE

2.1 Definition of NDE and NIE

Throughout this paper, we denote Y as the primary outcome of interest, D as a binary treatment variable, M as the mediator on the causal pathway from D to Y, and $X \in [0,1]^p$ as baseline covariates including all potential confounders. We denote the observed data vector as $O \equiv (X, D, M, Y)$. Let

$M(d)$ denote the potential outcome for the mediator when setting $D = d$ and $Y(d, m)$ be the potential outcome of Y under $D = d$ and $M = m$, where $d \in \{0, 1\}$ and m is in the support of M. We define the average total (treatment) effect as $\tau_{tot} := \mathsf{E}[Y(1, M(1)) - Y(0, M(0))]$, the average NDE of the treatment D on the outcome Y when the mediator takes the natural potential outcome when $D = d$ as $\tau_{\mathsf{NDE}}(d) := \mathsf{E}[Y(1, M(d)) - Y(0, M(d))]$, and the average NIE of the treatment D on the outcome Y via the mediator M as $\tau_{\mathsf{NIE}}(d) := \mathsf{E}[Y(d, M(1)) - Y(d, M(0))]$. We have the trivial decomposition $\tau_{tot} \equiv \tau_{\mathsf{NDE}}(d) + \tau_{\mathsf{NIE}}(d')$ for $d \neq d'$. In causal mediation analysis, the parameters of interest are $\tau_{\mathsf{NDE}}(d)$ and $\tau_{\mathsf{NIE}}(d)$.

2.2 Semiparametric multiply-robust estimators of NDE/NIE

Estimating $\tau_{\mathsf{NDE}}(d)$ and $\tau_{\mathsf{NIE}}(d)$ can be reduced to estimating $\phi(d, d') := \mathsf{E}[Y(d, M(d'))]$ for $d, d' \in \{0, 1\}$. We make the following standard identification assumptions:

i. **Consistency**: if $D = d$, then $M = M(d)$ for all $d \in \{0, 1\}$; while if $D = d$ and $M = m$, then $Y = Y(d, m)$ for all $d \in \{0, 1\}$ and all m in the support of M,

ii. **Ignorability**: $\{Y(d, m), M(d')\} \perp D|X$ almost surely and $Y(d, m) \perp M|D, X$ almost surely for all $d, d' \in \{0, 1\}$ and all m in the support $\mathcal{M}$ of M,

iii. **Positivity**: The propensity score $a(d|X) \equiv \mathsf{Pr}(D = d|X) \in (c, C)$ for some constants $0 < c \leq C < 1$, almost surely for all $d \in \{0, 1\}$; $f(m|X, d)$, the conditional density (mass) function of $M = m$ (when M is discrete) given X and $D = d$, is strictly bounded between $[\underline{\rho}, \bar{\rho}]$ for some constants $0 < \underline{\rho} \leq \bar{\rho} < \infty$ almost surely for all m in $\mathcal{M}$ and all $d \in \{0, 1\}$.

Under the above assumptions, the causal parameter $\phi(d, d')$ for $d, d' \in \{0, 1\}$ can be identified as either of the following three observed-data functionals:

$$\phi(d, d') \equiv \mathsf{E}\left[\frac{\mathbb{1}\{D = d\} f(M|X, d')Y}{a(d|X) f(M|X, d)}\right] \equiv \mathsf{E}\left[\frac{\mathbb{1}\{D = d'\}}{a(d'|X)} \mu(X, d, M)\right] \equiv \int \mu(X, d, m) f(m|X, d')\, \mathrm{d}m,$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function and $\mu(x, d, m) := \mathsf{E}[Y|X = x, D = d, M = m]$ is the outcome regression model, for which we also make the following standard boundedness assumption:

iv. $\mu(x, d, m)$ is also strictly bounded between $[-R, R]$ for some constant $R > 0$.

Following the convention in the semiparametric causal inference literature, we call a, f, μ "nuisance functions". Tchetgen Tchetgen and Shpitser (2012) derived the EIF of $\phi(d, d')$: $\mathsf{EIF}_{d,d'} \equiv \psi_{d,d'}(O) - \phi(d, d')$, where

$$\begin{aligned}\psi_{d,d'}(O) = &\frac{\mathbb{1}\{D = d\} \cdot f(M|X, d')}{a(d|X) \cdot f(M|X, d)}(Y - \mu(X, d, M)) \\ &+ \left(1 - \frac{\mathbb{1}\{D = d'\}}{a(d'|X)}\right) \int_{m \in \mathcal{M}} \mu(X, d, m) f(m|X, d') \mathrm{d}m + \frac{\mathbb{1}\{D = d'\}}{a(d'|X)} \mu(X, d, M). \qquad (1)\end{aligned}$$

The nuisance functions $\mu(x, d, m)$, $a(d|x)$ and $f(m|x, d)$ appeared in $\psi_{d,d'}(o)$ are unknown and generally high-dimensional. But with a sample $\mathcal{D} \equiv \{O_j\}_{j=1}^N$ of the observed data, based on $\psi_{d,d'}(o)$, one can construct the following generic sample-splitting multiply-robust estimator of $\phi(d, d')$:

$$\widetilde{\phi}(d, d') = \frac{1}{n} \sum_{i \in \mathcal{D}_n} \widetilde{\psi}_{d,d'}(O_i), \qquad (2)$$

where $\mathcal{D}_n \equiv \{O_i\}_{i=1}^n$ is a subset of all N data, and $\widetilde{\psi}_{d,d'}(o)$ replaces the unknown nuisance functions a, f, μ in $\psi_{d,d'}(o)$ by some generic estimators $\widetilde{a}, \widetilde{f}, \widetilde{\mu}$ computed using the remaining $N - n$ *nuisance sample data*, denoted as $\mathcal{D}_\nu$. Cross-fit is then needed to recover the information lost due to sample splitting (Chernozhukov et al., 2018); see Algorithm 1. It is clear from (1) that $\widetilde{\phi}(d, d')$ is a consistent estimator of $\phi(d, d')$ as long as any two of $\widetilde{a}, \widetilde{f}, \widetilde{\mu}$ are consistent estimators of the corresponding true nuisance functions, hence the name "multiply-robust". Throughout this paper, we take $n \asymp N - n$ and assume:

v. Any nuisance function estimators are strictly bounded within the respective lower and upper bounds of a, f, μ.

To further ease notation, we define: for any $d \in \{0,1\}$, $r_{a,d} := \left\{\int \delta_{a,d}(x)^2 \mathrm{d}F(x)\right\}^{1/2}$, $r_{f,d} := \left\{\int \delta_{f,d}(x,m)^2 \mathrm{d}F(x,m|d=0)\right\}^{1/2}$, and $r_{\mu,d} := \left\{\int \delta_{\mu,d}(x,m)^2 \mathrm{d}F(x,m|d=0)\right\}^{1/2}$, where $\delta_{a,d}(x) := \widetilde{a}(d|x) - a(d|x)$, $\delta_{f,d}(x,m) := \widetilde{f}(m|x,d) - f(m|x,d)$ and $\delta_{\mu,d}(x,m) := \widetilde{\mu}(x,d,m) - \mu(x,d,m)$ are point-wise estimation errors of the estimated nuisance functions. In defining the above L_2-estimation errors, we choose to take expectation with respect to (w.r.t.) the law $F(m,x|d=0)$ *only* for convenience, with no loss of generality by Assumptions iii and v.

To show the cross-fit version of $\widetilde{\phi}(d,d')$ is semiparametric efficient for $\phi(d,d')$, we shall demonstrate under what conditions $\sqrt{n}(\widetilde{\phi}(d,d') - \phi(d,d')) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathsf{E}[\mathsf{EIF}^2_{d,d'}])$ (Newey, 1990). The following meta-theorem on the statistical properties of $\widetilde{\phi}(d,d')$ is a key step towards this objective.

Theorem 1. *Denote* $\mathrm{Bias}(\widetilde{\phi}(d,d')) := \mathsf{E}[\widetilde{\phi}(d,d') - \phi(d,d')|\mathcal{D}_\nu]$ *as the bias of* $\widetilde{\phi}(d,d')$ *conditional on the nuisance sample* $\mathcal{D}_\nu$*. Under Assumptions i – v,* $\mathrm{Bias}(\widetilde{\phi}(d,d'))$ *is of second-order:*

$$|\mathrm{Bias}(\widetilde{\phi}(d,d'))| \lesssim \max\left\{r_{a,d}\cdot r_{f,d}, \max_{d''\in\{0,1\}} r_{f,d''}\cdot r_{\mu,d}, r_{a,d}\cdot r_{\mu,d}\right\}. \tag{3}$$

Furthermore, if the RHS of (3) *is* $o(n^{-1/2})$*, then*

$$\sqrt{n}\left(\widetilde{\phi}(d,d') - \phi(d,d')\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi_{d,d'}(O_i) - \phi(d,d')) + o(1) \xrightarrow{d} \mathcal{N}\left(0, \mathsf{E}\left[\mathsf{EIF}^2_{d,d'}\right]\right). \tag{4}$$

For completeness, we prove Theorem 1 in Appendix A.

Remark 2. *The total effect* $\tau_{tot} = \phi(1,1) - \phi(0,0)$ *can be viewed as a special case, for which* $d = d'$ *for* $\phi(d,d')$*. Then* $\mathsf{EIF}_{d,d} \equiv \mathsf{EIF}_d$ *corresponds to the nonparametric EIF of* $\phi(d,d) \equiv \phi(d) \equiv \mathsf{E}[Y(d,M(d))]$*:*

$$\mathsf{EIF}_d = \psi_d(O) - \phi(d) \text{ with } \psi_d(O) = \frac{\mathbb{1}\{D=d\}}{a(d|X)}Y + \left(1 - \frac{\mathbb{1}\{D=d\}}{a(d|X)}\right)\mu(X,d),$$

where $\mu(x,d) := \mathsf{E}[Y|X=x, D=d]$*. Hence all the theoretical results in this paper are applicable for total effect estimation.*

3 Estimation and inference of NDE/NIE using DeepMed

We now introduce DeepMed, a method for mediation analysis with nuisance functions estimated by DNNs. By leveraging the second-order bias property of the multiply-robust estimators of NDE/NIE (Theorem 1), we will derive statistical properties of DeepMed in this section. The nuisance function estimators by DNNs are denoted as $\widehat{a}, \widehat{f}, \widehat{\mu}$.

3.1 Details on DeepMed

First, we introduce the fully-connected feed-forward neural network with the rectified linear units (ReLU) as the activation function for the hidden layer neurons (FNN-ReLU), which will be used to estimate the nuisance functions. Then, we will introduce an estimation procedure using a V-fold cross-fitting with sample-splitting to avoid the Donsker-type empirical-process assumption on the nuisance functions, which, in general, is violated in high-dimensional setup. Finally, we provide the asymptotic statistical properties of the DNN-based estimators of τ_{tot}, $\tau_{\mathsf{NDE}}(d)$ and $\tau_{\mathsf{NIE}}(d)$.

We denote the ReLU activation function as $\sigma(u) := \max(u,0)$ for any $u \in \mathbb{R}$. Given vectors x, b, we denote $\sigma_b(x) := \sigma(x-b)$, with σ acting on the vector $x-b$ component-wise.

Let $\mathcal{F}_{nn}$ denote the class of the FNN-ReLU functions

$$\mathcal{F}_{nn} := \left\{f : \mathbb{R}^p \to \mathbb{R}; f(x) = W^{(L)}\sigma_{b^{(L)}} \circ \cdots \circ W^{(1)}\sigma_{b^{(1)}}(x)\right\},$$

where $\circ$ is the composition operator, L is the number of layers (i.e. depth) of the network, and for $l = 1, \cdots, L$, $W^{(l)}$ is a $K_{l+1} \times K_l$-dimensional weight matrix with K_l being the number of

neurons in the l-th layer (i.e. width) of the network, with $K_1 = p$ and $K_{L+1} = 1$, and $b^{(l)}$ is a K_l-dimensional vector. To avoid notation clutter, we concatenate all the network parameters as $\Theta = (W^{(l)}, b^{(l)}, l = 1, \cdots, L)$ and simply take $K_2 = \cdots = K_L = K$. We also assume Θ to be bounded: $\|\Theta\|_\infty \leq B$ for some universal constant $B > 0$. We may let the dependence on L, K, B explicit by writing $\mathcal{F}_{nn}$ as $\mathcal{F}_{nn}(L, K, B)$.

DeepMed estimates $\tau_{tot}, \tau_{\mathsf{NDE}}(d), \tau_{\mathsf{NIE}}(d)$ by (2), with the nuisance functions a, f, μ estimated using $\mathcal{F}_{nn}$ with the V-fold cross-fitting strategy, summarized in Algorithm 1 below. DeepMed inputs the observed data $\mathcal{D} \equiv \{O_i\}_{i=1}^N$ and outputs the estimated total effect $\widehat{\tau}_{tot}$, NDE $\widehat{\tau}_{\mathsf{NDE}}(d)$ and NIE $\widehat{\tau}_{\mathsf{NIE}}(d)$, together with their variance estimators $\widehat{\sigma}^2_{tot}$, $\widehat{\sigma}^2_{\mathsf{NDE}}(d)$ and $\widehat{\sigma}^2_{\mathsf{NIE}}(d)$.

Algorithm 1 DeepMed with V-fold cross-fitting

1: Choose some integer V (usually $V \in \{2, 3, \cdots, 10\}$)
2: Split the N observations into V subsamples $I_v \subset \{1, \cdots, N\} \equiv [N]$ with equal size $n = N/V$;
3: **for** $v = 1, \cdots, V$: **do**
4: Fit the nuisance functions by DNNs using observations in $[N] \setminus I_v$
5: Compute the nuisance functions in the subsample I_v using the estimated DNNs in step 4
6: Obtain $\{\widehat{\psi}_d(O_i), \widehat{\psi}_{d,d'}(O_i)\}_{i \in I_v}$ for the subsample I_v based on (1), respectively, with the nuisance functions replaced by their estimates in step 5
7: **end for**
8: Estimate average potential outcomes by $\widehat{\phi}(d) := \frac{1}{N}\sum_{i=1}^N \widehat{\psi}_d(O_i)$, $\widehat{\phi}(d, d') := \frac{1}{N}\sum_{i=1}^N \widehat{\psi}_{d,d'}(O_i)$
9: Estimate causal effects by $\widehat{\tau}_{tot}$, $\widehat{\tau}_{\mathsf{NDE}}(d)$ and $\widehat{\tau}_{\mathsf{NIE}}(d)$ with $\widehat{\phi}(d)$ and $\widehat{\phi}(d, d')$
10: Estimate the variances of $\widehat{\tau}_{tot}$, $\widehat{\tau}_{\mathsf{NDE}}(d)$ and $\widehat{\tau}_{\mathsf{NIE}}(d)$ by:
$\widehat{\sigma}^2_{tot} := \frac{1}{N^2}\sum_{i=1}^N (\widehat{\psi}_1(O_i) - \widehat{\psi}_0(O_i))^2 - \frac{1}{N}\widehat{\tau}^2_{tot}$; $\widehat{\sigma}^2_{\mathsf{NDE}}(d) := \frac{1}{N^2}\sum_{i=1}^N (\widehat{\psi}_{1,d}(O_i) - \widehat{\psi}_{0,d}(O_i))^2 - \frac{1}{N}\widehat{\tau}^2_{\mathsf{NDE}}(d)$;
$\widehat{\sigma}^2_{\mathsf{NIE}}(d) := \frac{1}{N^2}\sum_{i=1}^N (\widehat{\psi}_{d,1}(O_i) - \widehat{\psi}_{d,0}(O_i))^2 - \frac{1}{N}\widehat{\tau}^2_{\mathsf{NIE}}(d)$

Output: $\widehat{\tau}_{tot}$, $\widehat{\tau}_{\mathsf{NDE}}(d)$, $\widehat{\tau}_{\mathsf{NIE}}(d)$, $\widehat{\sigma}^2_{tot}$, $\widehat{\sigma}^2_{\mathsf{NDE}}(d)$ and $\widehat{\sigma}^2_{\mathsf{NIE}}(d)$

Remark 3 (Continuous or multi-dimensional mediators)**.** *For binary treatment D and continuous or multi-dimensional M, to avoid nonparametric/high-dimensional conditional density estimation, we can rewrite $\frac{f(m|x,d')}{a(d|x)f(m|x,d)}$ as $\frac{1-a(d|x,m)}{a(d|x,m)(1-a(d|x))}$ by the Bayes' rule and the integral w.r.t. $f(m|x, d')$ in (1) as $\mathsf{E}[\mu(X, d, M)|X = x, D = d']$. Then we can first estimate $\mu(x, d, m)$ by $\widehat{\mu}(x, d, m)$ and in turn estimate $\mathsf{E}[\mu(X, d, M)|X = x, D = d']$ by regressing $\widehat{\mu}(X, d, M)$ against (X, D) using the FNN-ReLU class. We mainly consider binary M to avoid unnecessary complications.*

3.2 Statistical properties of DeepMed: Non-sparse DNN architecture and low-dimensional structures of the nuisance functions

According to Theorem 1, to analyze the statistical properties DeepMed, it is sufficient to control the L_2-estimation errors of nuisance function estimates $\widehat{a}, \widehat{f}, \widehat{\mu}$ fit by DNNs. To ease presentation, we first study the theoretical guarantees on the L_2-estimation error for a generic nuisance function $g : W \in [0, 1]^p \to Z \in \mathbb{R}$, for which we assume:

vi. $Z = g(W) + \xi$, with ξ sub-Gaussian with mean zero and independent of W.

Note that when g corresponds to a, f, μ, (W, Z) corresponds to $(X, \mathbb{1}(D = 1))$, $((X, D), \mathbb{1}(M = 1))$ and $((X, D, M), Y)$, respectively.

We denote the DNN output from the nuisance sample $\mathcal{D}_\nu$ as $\widehat{g}$. For theoretical results, we consider $\widehat{g}$ as the following empirical risk minimizer (ERM):

$$\widehat{g} := \underset{\bar{g} \in \mathcal{F}_{nn}(L,K,B)}{\arg\min} \sum_{i \in \mathcal{D}_\nu} \left(Z_i - \bar{g}(W_i)\right)^2. \tag{5}$$

To avoid model misspecification, one often assumes $g \in \mathcal{G}$, where $\mathcal{G}$ is some *infinite-dimensional* function space. A common choice is $\mathcal{G} = \mathcal{H}_p(\alpha; C)$, the Hölder ball on the input domain $[0, 1]^p$, with smoothness exponent α and radius C. Hölder space is one of the most well-studied function

spaces in semiparametric statistics; see Appendix B for a review. It is well-known that estimating Hölder functions suffers from curse-of-dimensionality (Stone, 1982). One remedy is to consider the following generalized Hölder space, by imposing certain low-dimensional structures of g in terms of the dependence on W:

$$\mathcal{H}_k^\dagger(\alpha; C) := \left\{ g(w) = h(\Gamma w) : h \in \mathcal{H}_k(\alpha; C), \Gamma \in \mathbb{R}^{k \times p} \text{ unknown}, k \leq p \right\}.$$

Remark 4. *The above definition contains $g(w) = h(w_I)$, where $I \subset \{1, \cdots, p\}$, as a special case, in which g is assumed to only depend on some sub-coordinates of w. One can easily generalize the above definition to additive models $g(w) = \sum_{j=1}^p h_j(w_j)$ where $h_j \in \mathcal{H}_{k_j}(\alpha_j; C_j)$, allowing even more modeling flexibility. For the theoretical results, we only consider the above simpler model with no essential loss of generality.*

We then have the following:

Lemma 5. *Under Assumptions iii – vi, if $g \in \mathcal{H}_k^\dagger(\alpha; C)$ for $k \leq p$, for $\widehat{g}$ defined in (5) with $LK \asymp n^{\frac{k}{2(k+2\alpha)}}$, we have $\sup_{g \in \mathcal{H}_k^\dagger(\alpha;C)} \left\{ \mathsf{E}\left[(g(W) - \widehat{g}(W))^2 \right] \right\}^{1/2} \lesssim n^{-\frac{\alpha}{2\alpha+k}} (\log n)^3$.*

Lemma 5, together with Theorem 1, implies the main theoretical result of the paper.

Theorem 6. *Under Assumptions i – vi and the following condition on a, f, μ: $a \in \mathcal{H}_k^\dagger(\alpha_a; C)$, $f \in \mathcal{H}_k^\dagger(\alpha_f; C)$, $\mu \in \mathcal{H}_k^\dagger(\alpha_\mu; C)$, with*

$$\min \left\{ \frac{\alpha_a}{2\alpha_a + k} + \frac{\alpha_f}{2\alpha_f + k}, \frac{\alpha_f}{2\alpha_f + k} + \frac{\alpha_\mu}{2\alpha_\mu + k}, \frac{\alpha_a}{2\alpha_a + k} + \frac{\alpha_\mu}{2\alpha_\mu + k} \right\} > \frac{1}{2} + \epsilon, \tag{6}$$

for $k \leq p$ and some arbitrarily small $\epsilon > 0$, if $\widehat{a}$, $\widehat{f}$, $\widehat{\mu}$ are respectively the ERM (5) from FNN-ReLU classes $\mathcal{F}_{nn}(L_a, K_a, B)$, $\mathcal{F}_{nn}(L_f, K_f, B)$, $\mathcal{F}_{nn}(L_\mu, K_\mu, B)$, of which the product of the depth and width satisfies $L_g K_g \asymp n^{\frac{k}{2(k+2\alpha_g)}}$ for $g \in \{a, f, \mu\}$, then the DeepMed *estimators $\widehat{\tau}_{tot}$, $\widehat{\tau}_{\mathsf{NDE}}(d)$ and $\widehat{\tau}_{\mathsf{NIE}}(d)$ computed by Algorithm 1 are semiparametric efficient:*

$$\widehat{\sigma}_{tot}^{-1}(\widehat{\tau}_{tot} - \tau_{tot}), \widehat{\sigma}_{\mathsf{NDE}}^{-1}(d)(\widehat{\tau}_{\mathsf{NDE}}(d) - \tau_{\mathsf{NDE}}(d)), \widehat{\sigma}_{\mathsf{NIE}}^{-1}(d)(\widehat{\tau}_{\mathsf{NIE}}(d) - \tau_{\mathsf{NIE}}(d)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1), \text{ with}$$

$$N\widehat{\sigma}_{tot}^2 \xrightarrow{p} \mathsf{E}[(\mathsf{EIF}_1 - \mathsf{EIF}_0)^2], \; N\widehat{\sigma}_{\mathsf{NDE}}^2(d) \xrightarrow{p} \mathsf{E}[(\mathsf{EIF}_{1,d} - \mathsf{EIF}_{0,d})^2],$$

$$\text{and } N\widehat{\sigma}_{\mathsf{NIE}}^2(d) \xrightarrow{p} \mathsf{E}[(\mathsf{EIF}_{d,1} - \mathsf{EIF}_{d,0})^2].$$

Remark 7. *To unload notation in the above theorem, consider the special case where the smoothness of all the nuisance functions coincides, i.e. $\alpha_a = \alpha_f = \alpha_\mu = \alpha$. Then Condition (6) reduces to $\alpha > k/2 + \epsilon$ for some arbitrarily small $\epsilon > 0$. For example, if the covariates X have dimension $p = 2$ and no low-dimensional structures are imposed on the nuisance functions (i.e. $k \equiv p$), one needs $\alpha > 1$ to ensure semiparametric efficiency of the* DeepMed *estimators.*

We would like to emphasize that Lemma 5 and Theorem 6 do not constrain the overall network sparsity S, better reflecting how DNNs are usually used in practice. Theorem 6 also improves existing results on total effect estimation with non-sparse DNNs (Farrell et al., 2021, Theorem 1) in terms of (1) weaker smoothness conditions and (2) adapting to certain low-dimensional structures of the nuisance functions. The proof of Lemma 5 follows from a combination of the improved DNN approximate rate obtained in Lu et al. (2021); Jiao et al. (2021) and standard DNN metric entropy bound (Suzuki, 2019). We prove Lemma 5 and Theorem 6 in Appendix B for completeness. One weakness of Lemma 5 and Theorem 6, as well as in other contemporary works (Chen et al., 2020; Farrell et al., 2021), is the lack of algorithmic/training process considerations; see Remark 10 and Appendix D for extended discussions.

Remark 8 (Hyperparameter tuning and explicit input-layer regularization)**.** *Training DNNs in practice involves hyperparameters, including the depth L and width K in Theorem 6 and others like epochs. In the synthetic experiments, we consider the nuisance functions only depending on a k-subset of p-dimensional input. A reasonable heuristic is to add L_1-regularization in the input-layer of the DNN. Then the regularization weight λ is also a hyperparameter. In practice, we simply use cross-validation to select the hyperparameters that minimize the validation loss. We leave the theoretical justification to future works.*

4 Synthetic experiments

In this section and Appendix D, we showcase five synthetic experiments. Since ground truth is rarely known in real data, we believe synthetic experiments play an equally, if not more, important role as real data. Before describing the experimental setups, we garner the following key take-home message:

(a) Compared with the other competing methods, DeepMed exhibits better finite-sample performance in most of our experiments;

(b) Cross-validation for DNN hyperparameter tuning works reasonably well in our experiments;

(c) We find DeepMed with explicit regularization in the input layer improves performance (see Table A2) when the true nuisance functions have certain low-dimensional structures in their dependence on the covariates. Farrell et al. (2021) warned against blind explicit regularization in DNNs for total effect estimation. Our observation does not contradict Farrell et al. (2021) as (1) the purpose of the input-layer regularization is not to control the sparsity of the DNN architecture and (2) we do not further regularize hidden layers;

(d) Experimental setups for Cases 3 to 5 generate nuisance functions that are nearly *infinite-dimensional* and close to the boundary of a Hölder ball with a given smoothness exponent. Thus these synthetic experiments should be better benchmarks than Cases 1 and 2 or settings in other related works such as Farrell et al. (2021). We hope that these novel synthetic experiments may be helpful to researchers even beyond mediation analysis or causal inference. We share the code for generating these functions as a part of the DeepMed package.

We consider a sample with 10,000 i.i.d. observations. The covariates $X = (X_1, ..., X_p)^\top$ are independently drawn from uniform distribution $\text{Uniform}([-1, 1])$. The outcome Y, treatment D and mediator M are generated as follows:

$$D \sim \text{Bernoulli}(\mathfrak{s}(d(X))), M \sim 0.2D + m(X) + \mathcal{N}(0, 1), Y \sim 0.2D + M + y(X) + \mathcal{N}(0, 1),$$

where $\mathfrak{s}(x) := (1 + e^{-x})^{-1}$, and we consider the following three cases to generate the nonlinear functions $d(x), m(x)$ and $y(x)$ in the main text:

• Case 1 (simple functions):

$$d(x) = x_1x_2 + x_3x_4x_5 + \sin x_1, m(x) = 4\sum_{i=1}^{5} \sin 3x_i, y(x) = (x_1 + x_2)^2 + 5 \sin \sum_{i=1}^{5} x_i.$$

• Case 2 (composition of simple functions): we simulate more complex interactions among covariates by composing simple functions as follow:

$$d(x) = d_2 \circ d_1 \circ d_0(x_1, \cdots, x_5), \text{ with } d_0(x_1, \cdots, x_5) = \left(\prod_{i=1}^{2} x_i, \prod_{i=3}^{5} x_i, \prod_{i=1}^{2} \sin x_i, \prod_{i=3}^{5} \sin x_i\right),$$

$$d_1(a_1, \cdots, a_4) = (\sin(a_1 + a_2), \sin a_2, a_3, a_4), \text{and } d_2(b_1, \cdots, b_4) = 0.5 \sin(b_1 + b_2) + 0.5(b_3 + b_4),$$

$$m(x) = m_1 \circ m_0(x_1, \ldots, x_5), \text{with } m_0(x_1, \cdots, x_5) = (\sin x_1, \cdots, \sin x_5), m_1(a_1, \cdots, a_5) = 5 \sin \sum_{i=1}^{5} a_i$$

$$\text{and } y(x) = y_2 \circ y_1 \circ y_0(x_1, \cdots, x_5), \text{ with } y_0(x_1, \cdots, x_5) = \left(\sin \sum_{i=1}^{2} x_i, \sin \sum_{i=3}^{5} x_i, \sin \sum_{i=1}^{5} x_i\right),$$

$$y_1(a_1, a_2, a_3) = (\sin(a_1 + a_2), a_3), \text{and } y_2(b_1, b_2) = 10 \sin(b_1 + b_2).$$

• Case 3 (Hölder functions): we consider more complex nonlinear functions as follows:

$$d(x) = x_1x_2 + x_3x_4x_5 + 0.5\eta(0.2x_1; \alpha), m(x) = \sum_{i=1}^{5} \eta(0.5x_i; \alpha), y(x) = x_1x_2 + 3\eta\left(0.2\sum_{i=1}^{5} x_i; \alpha\right)$$

where $\eta(x; \alpha) = \sum_{j \in J, l \in \mathbb{Z}} 2^{-j(\alpha+0.25)} w_{j,l}(x)$ with $J = \{0, 3, 6, 9, 10, 16\}$ and $w_{j,l}(\cdot)$ is the D6 father wavelet functions dilated at resolution j shifted by l. By construction, $\eta(x; \alpha) \in \mathcal{H}_1(\alpha; B)$ for some known constant $B > 0$ following Härdle et al. (1998, Theorem 9.6). Here we set $\alpha = 1.2$ and

the intrinsic dimension $k = 1$. Thus we expect the DeepMed estimators are semiparametric efficient. It is indeed the case based on the columns corresponding to Case 3 in Table 1, suggesting that DNNs can be adaptive to certain low-dimensional structures.

Remark 9. *The nuisance functions in Cases 3 – 5 (see Appendix D) are less smooth than what have been considered elsewhere, including Farrell et al. (2021), Chen et al. (2020), and even Adcock and Dexter (2021), a paper dedicated to exposing the gap between theoretical approximation rates and DNN practice. These nuisance functions are designed to be near the boundary of a Hölder ball with a given smoothness exponent as we add wavelets at very high resolution in $\eta(x; \alpha)$. This is the assumption under which most of the known statistical properties of DNNs are developed.*

Table 1: The biases, empirical standard errors (SE) and root mean squared errors (RMSE) of the estimated $\tau_{tot}, \tau_{\mathsf{NDE}}(1)$ and $\tau_{\mathsf{NIE}}(1)$, and the coverage probabilities (CP) of their corresponding 95% confidence intervals. $p = 5$ (no irrelevant covariates). The simulation is based on 200 replicates. The full table including $\widehat{\tau}_{\mathsf{NDE}}(0)$ and $\widehat{\tau}_{\mathsf{NIE}}(0)$ can be found in Table A1 in the Appendix.

		Case 1				Case 2				Case 3			
	Method	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
τ_{tot}	DeepMed	-0.001	0.032	0.032	0.945	-0.004	0.032	0.032	0.955	0.008	0.037	0.038	0.920
	Lasso	0.192	0.089	0.212	0.460	-0.304	0.116	0.325	0.215	0.346	0.079	0.355	0.010
	RF	0.067	0.042	0.079	0.775	-0.078	0.056	0.096	0.950	-0.009	0.042	0.043	0.985
	GBM	-0.015	0.036	0.039	0.940	-0.044	0.055	0.070	0.850	0.019	0.041	0.045	0.930
	Oracle	-0.001	0.029	0.029	0.955	-0.003	0.029	0.029	0.925	-0.001	0.032	0.032	0.935
$\tau_{\mathsf{NDE}}(1)$	DeepMed	0.000	0.027	0.027	0.945	-0.007	0.023	0.024	0.955	0.000	0.026	0.026	0.965
	Lasso	0.130	0.043	0.137	0.220	-0.375	0.059	0.380	0.000	0.226	0.064	0.235	0.050
	RF	0.048	0.029	0.056	0.700	-0.188	0.044	0.193	0.005	0.030	0.038	0.048	0.980
	GBM	-0.040	0.031	0.051	0.770	-0.164	0.046	0.170	0.040	0.011	0.042	0.043	0.920
	Oracle	0.000	0.022	0.022	0.945	-0.002	0.020	0.020	0.985	0.001	0.022	0.022	0.955
$\tau_{\mathsf{NIE}}(1)$	DeepMed	-0.001	0.025	0.025	0.960	0.005	0.029	0.029	0.915	0.008	0.031	0.032	0.905
	Lasso	0.058	0.077	0.096	0.875	0.069	0.094	0.117	0.905	0.120	0.045	0.128	0.220
	RF	0.066	0.037	0.076	0.665	0.108	0.059	0.123	0.860	-0.045	0.038	0.059	0.765
	GBM	0.023	0.031	0.039	0.890	0.120	0.064	0.136	0.485	-0.001	0.037	0.037	0.935
	Oracle	-0.001	0.020	0.020	0.975	0.000	0.021	0.021	0.930	-0.002	0.022	0.022	0.920

In all the above cases, $\tau_{tot} = 0.4$ and $\tau_{\mathsf{NDE}}(d) = \tau_{\mathsf{NIE}}(d) = 0.2$ for $d \in \{0, 1\}$. We also consider the cases where the total number of covariates $p = 20$ and 100 but only the first five covariates are relevant to Y, M and D. All simulation results are based on 200 replicates. The sigmoid function is used in the final layer when the response variable is binary. For comparison, we also use the Lasso, random forest (RF) and gradient boosted machine (GBM) to estimate the nuisance functions, and use the true nuisance functions (Oracle) as the benchmark. The Lasso is implemented using the R package "hdm" with a data-driven penalty. The DNN, RF and GBM are implemented using the R packages "keras", "randomForest" and "gbm", respectively. We adopt a 3-fold cross-validation to choose the hyperparameters for DNNs (depth L, width K, L_1-regularization parameter λ and epochs), RF (number of trees and maximum number of nodes) and GBM (numbers of trees and depth). We use a completely independent sample for the hyperparameter selection. In this paper, we only use one extra dataset to conduct the cross-validation for hyperparameter selection, so our simulation results are conditional on this extra dataset. We use the cross-entropy loss for the binary response and the mean-squared loss for the continuous response. We fix the batch-size as 100 and the other hyperparameters for the other methods are set to the default values in their R packages. See Appendix D for more details.

We compare the performances of different methods in terms of the biases, empirical standard errors (SE) and root mean squared errors (RMSE) of the estimates as well as the coverage probabilities (CP) of their 95% confidence intervals. When $p = k = 5$ (all covariates are relevant or no low-dimensional structures), DeepMed has smaller bias and RMSE than the other competing methods, and is only slightly worse than Oracle. Lasso has the largest bias and poor CP as expected since it does not capture the nonlinearity of the nuisance functions. RF and GBM also have substantial biases, especially in Case 2 with compositions of simple functions. Overall, DeepMed performs better than the competing methods (Table 1). From the empirical distributions, we can also see that they are nearly unbiased and normally distributed in Cases 1-3 (Figures A1-A3). When $p = 20$ or 100 but only the first five covariates are relevant ($k = 5$), L_1-regularization in the input-layer drastically improves the

performance of DeepMed (Table A2). DeepMed with L_1-regularization in the input-layer also has smaller bias and RMSE than the other competing methods (Tables A3 and A4).

As expected, we find that more precise nuisance function estimates (i.e., smaller validation loss) generally lead to more precise causal effect estimates (i.e. smaller bias and RMSE). The validation losses of nuisance function estimates from DeepMed are generally much smaller than those using Lasso, RF and GBM (Tables A5-A7).

Remark 10. *Due to space limitations, we display the first three experiments in the main text and defer Cases 4 and 5 to Appendix D. In Cases 4 and 5,* DeepMed *fails to be semiparametric efficient; see an extended discussion in Appendix D. We conjecture this is due to the implicit regularization effect of the training process as* adam *(Kingma and Ba, 2015) is used to train the DNNs to estimate the nuisance parameters, instead of actually solving the ERM (5). Such phenomenon have not been documented in the literature on DNN-based causal inference before and exemplify the value of our novel synthetic experiments. We believe this is an important research direction for theory to fully capture the empirical performance of DNN-based causal inference methods such as* DeepMed.

5 Real data analysis on fairness

As a proof of concept, we use DeepMed and other competing methods to re-analyze the COMPAS algorithm (Dressel and Farid, 2018). In particular, we are interested in the NDE of race D on the recidivism risk (or the COMPAS score) Y with the number of prior convictions as the mediator M. For race, we mainly focus on the Caucasians population ($D = 0$) and the African-Americans population ($D = 1$), and exclude the individuals of other ethnicity groups. The COMPAS score (Y) is ordinal, ranging from 1 to 10 (1: lowest risk; 10: highest risk). We also include the demographic information (age and gender) as covariates X.

All the methods find significant positive NDE of race on the COMPAS score at α-level 0.005 (Table 2; all p-values $< 10^{-7}$), consistent with previous findings (Nabi and Shpitser, 2018). Thus the COMPAS algorithm tends to assign higher recidivism risks to African-Americans than to Caucasians, even when they have the same number of prior convictions. The validation losses of nuisance function estimates by DeepMed are smaller than the other competing methods (Table A8), possibly suggesting smaller biases of the corresponding NDE/NIE estimators.

We want to emphasize that research in machine learning fairness should be held accountable (Bao et al., 2021). Our data analysis is merely a proof-of-concept that DeepMed works in practice and the conclusion from our data analysis should not be treated as definitive. Finally, we defer the comments on potential issues of unmeasured confounding to Appendix E and another real data analysis to Appendix F.

Table 2: Results for real data application to COMPAS algorithm fairness.

Method	Effect	Estimate	SE	Method	Effect	Estimate	SE
DeepMed	τ_{tot}	1.136	0.069	RF	τ_{tot}	1.083	0.111
	$\tau_{\mathsf{NDE}}(1)$	0.564	0.068		$\tau_{\mathsf{NDE}}(1)$	0.589	0.070
	$\tau_{\mathsf{NDE}}(0)$	0.524	0.062		$\tau_{\mathsf{NDE}}(0)$	0.569	0.103
	$\tau_{\mathsf{NIE}}(1)$	0.612	0.042		$\tau_{\mathsf{NIE}}(1)$	0.514	0.049
	$\tau_{\mathsf{NIE}}(0)$	0.572	0.051		$\tau_{\mathsf{NIE}}(0)$	0.494	0.065
Lasso	τ_{tot}	1.150	0.068	GBM	τ_{tot}	1.180	0.068
	$\tau_{\mathsf{NDE}}(1)$	0.575	0.063		$\tau_{\mathsf{NDE}}(1)$	0.550	0.063
	$\tau_{\mathsf{NDE}}(0)$	0.587	0.062		$\tau_{\mathsf{NDE}}(0)$	0.526	0.061
	$\tau_{\mathsf{NIE}}(1)$	0.563	0.032		$\tau_{\mathsf{NIE}}(1)$	0.654	0.041
	$\tau_{\mathsf{NIE}}(0)$	0.575	0.040		$\tau_{\mathsf{NIE}}(0)$	0.630	0.044

6 Conclusion and Discussion

In this paper, we proposed DeepMed for semiparametric mediation analysis with DNNs. We established the state-of-the-art statistical properties for DNN-based causal effect estimation that

can (1) circumvent sparse DNN architectures and (2) leverage certain low-dimensional structures of the nuisance functions. These results significantly advance our current understanding of DNN-based causal inference including mediation analysis.

Evaluated by our extensive synthetic experiments, DeepMed exhibits improved finite-sample performance over the other competing machine learning methods. But as mentioned in Remark 10, an important future direction is to incorporate the training process while investigating the statistical properties to have a deeper theoretical understanding of DNN-based causal inference. It is also of future research interests to enable DeepMed to handle unmeasured confounding.

Finally, we warn readers that all causal inference methods, including DeepMed, may have negative societal impact if they are used without carefully checking their working assumptions.

References

Ben Adcock and Nick Dexter. The gap between theory and practice in function approximation with deep neural networks. *SIAM Journal on Mathematics of Data Science*, 3(2):624–655, 2021.

Michelle Bao, Angela Zhou, Samantha A Zottola, Brian Brubach, Sarah Desmarais, Aaron Seth Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.

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

Xiaohong Chen, Ying Liu, Shujie Ma, and Zheng Zhang. Efficient estimation of general treatment effects using neural networks with a diverging number of confounders. *arXiv preprint arXiv:2009.07055*, 2020.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of Riesz representers. *arXiv preprint arXiv:2101.00009*, 2020.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

Helmut Farbmacher, Martin Huber, Henrika Langen, and Martin Spindler. Causal mediation analysis with double machine learning. *arXiv preprint arXiv:2002.12710*, 2020.

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

Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 1998.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximately low-dimensional manifolds. *arXiv preprint arXiv:2104.06708*, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Whitney K Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2): 99–135, 1990.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*, pages 411–420. Morgan Kaufman, 2001.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22, 2021.

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

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

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics*, 40(3): 1816–1845, 2012.

Kazuma Tsuji and Taiji Suzuki. Estimation error analysis of deep learning on the regression problem on the variable exponent Besov space. *Electronic Journal of Statistics*, 15(1):1869–1908, 2021.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

A The bias of generic sample-splitting multiply-robust estimators of NDE/NIE

In this section, we prove Theorem 1, which is a consequence of the Proposition below.

Proposition 11. *Conditional on the nuisance sample data $\mathcal{D}_\nu$, the bias of $\widetilde{\phi}(d,d')$ as an estimator of $\phi(d,d')$ is of the following second-order form:*

$$
\begin{aligned}
&\mathsf{E}\left[\widetilde{\phi}(d,d') - \phi(d,d')|\mathcal{D}_\nu\right] \\
&= \mathsf{E}_X\left[\int_{m\in\mathcal{M}} \left(1 - \frac{a(d'|X)}{\widetilde{a}(d'|X)}\right)\left(\frac{\widetilde{f}(m|X,d')}{f(m|X,d')} - 1\right)\widetilde{\mu}(X,d,m) f(m|X,d')\mathrm{d}m\right] \\
&+ \mathsf{E}_X\left[\int_{m\in\mathcal{M}} \left(1 - \frac{f(m|X,d)}{\widetilde{f}(m|X,d)}\frac{\widetilde{f}(m|X,d')}{f(m|X,d')}\right)\left(\widetilde{\mu}(X,d,m) - \mu(X,d,m)\right) f(m|X,d')\mathrm{d}m\right] \\
&+ \mathsf{E}_X\left[\int_{m\in\mathcal{M}} \left(1 - \frac{a(d|X)}{\widetilde{a}(d|X)}\right)\frac{\widetilde{f}(m|X,d')}{\widetilde{f}(m|X,d)}\left(\widetilde{\mu}(X,d,m) - \mu(X,d,m)\right) f(m|X,d)\mathrm{d}m\right].
\end{aligned}
$$

Consequently, one obtains the following upper bound of the bias:

$$
\begin{aligned}
\mathrm{Bias}(\widetilde{\phi}(d,d')) &\equiv \left|\mathsf{E}\left[\widetilde{\phi}(d,d') - \phi(d,d')|\mathcal{D}_\nu\right]\right| \\
&\lesssim r_{a,d}\cdot r_{f,d} + \max_{d''\in\{0,1\}} r_{f,d''}\cdot r_{\mu,d} + r_{a,d}\cdot r_{\mu,d}.
\end{aligned} \tag{7}
$$

Proof. The first statement on the bias follows directly from sample-splitting and the form of the EIF $\psi_{d,d'}(o) - \phi(d,d')$. The second statement is obtained by the application of triangle inequality and Cauchy-Schwarz inequality. □

It is worth noting that the upper bound in (7) by Cauchy-Schwarz inequality is by no means the only analysis strategy. For instance, one could also upper bound the bias by Hölder inequality if convergence rates of DNN-based nuisance function estimators are available in general L_p-norms beyond $p = 2$. Theorem 1 is a generalization of the results in Robins et al. (2008); Chernozhukov et al. (2018) to mediation analysis.

B Hölder functions, their ERM DNN-based estimators and statistical properties

As in the main text, we denote $\mathcal{H}_p(\alpha; C)$ as the Hölder balls of functions from $\mathbb{R}^p$ to $\mathbb{R}$, with smoothness exponent α and radii C (Triebel, 2010; Giné and Nickl, 2016), formally defined below:

$$
\mathcal{H}_p(\alpha;C) := \begin{cases} \left\{ g:[0,1]^p\to\mathbb{R};\ \begin{array}{c} \max_{m\in\mathbb{Z}^p_{\geq 0},|m|_1<\lfloor\alpha\rfloor} \|\partial^m g\|_\infty \leq C \\ \text{and} \max_{m\in\mathbb{Z}^p_{\geq 0},|m|_1=\lfloor\alpha\rfloor} \sup_{w,w'\in[0,1]^p, w\neq w'} \frac{|\partial^m g(w) - \partial^m g(w')|}{\|w-w'\|_\infty^{\alpha-\lfloor\alpha\rfloor}} \leq C \end{array} \right\} & \alpha \geq 1, \\ \left\{ g:[0,1]^p\to\mathbb{R};\ \sup_{w,w'\in[0,1]^p, w\neq w'} \frac{|g(w)-g(w')|}{\|w-w'\|_\infty^\alpha} \leq C \right\} & 0<\alpha<1. \end{cases}
$$

It is well-known (Stone, 1982) that the minimax optimal convergence rate of estimating $g \in \mathcal{H}_p(\alpha; C)$ in L_2-norm is $n^{-\frac{\alpha}{2\alpha+p}}$, suffering from curse-of-dimensionality. As mentioned in the main text, one possibility is to consider the function space $\mathcal{H}_k^\dagger(\alpha; C)$ by assuming that the nuisance functions only depend on the covariates $w \in \mathbb{R}^p$ via a k-dimensional linear subspace Γw, where $\Gamma \in \mathbb{R}^{k\times p}$ and is unknown.

There exist many estimators attaining the optimal rate: e.g. wavelet projection estimators, kernel estimators, etc. In particular, sparse DNN-based estimators have also shown to attain the optimal rate up to a log-factor (Schmidt-Hieber, 2020; Suzuki, 2019). However, since sparse DNNs are computationally demanding to search over $\mathcal{F}_{nn}$, we prefer results that avoid such sparsity constraints. To this end, it is easy to show the following by adapting the proof of Theorem 1.1 of Lu et al. (2021):

Lemma 12. *Given $g \in \mathcal{H}_k^\dagger(\alpha; C)$, for large enough depth and width $L, K \in \mathbb{Z}_{>0}$ and some known constant $B > 0$, there exists $\widetilde{g} \in \mathcal{F}_{nn}(L, K, B)$ such that*

$$\|g - \widetilde{g}\|_\infty \lesssim \left(\frac{L}{\log L} \frac{K}{\log K} \right)^{-2\alpha/k}.$$

The proof is straightforward by simply taking the parameter of the input layer to be $W^{(1)} = (\Gamma, -\Gamma) \in \mathbb{R}^{2k \times p}$ and $b^{(1)} = 0$ and the second layer parameters chosen appropriately such that the input becomes Γx before the ReLU activation function. The rest of the proof then follows directly by applying Theorem 1.1 of Lu et al. (2021).

Next, we invoke the metric entropy bound of $\mathcal{F}_{nn}(L, K, B)$ established by Lemma 3 of Suzuki (2019):

Lemma 13 (Metric entropy bound of $\mathcal{F}_{nn}(L, K, B)$)**.** *Denote the covering number (van der Vaart and Wellner, 1996) of $\mathcal{F}_{nn}(L, K, B)$ w.r.t. L_∞-norm as $N(\epsilon, \mathcal{F}_{nn}(L, K, B), \|\cdot\|_\infty)$. Then for any $\epsilon > 0$, for large enough $L, K \in \mathbb{Z}_{>0}$ and $B > 0$, we have*

$$\log N(\epsilon, \mathcal{F}_{nn}(L, K, B), \|\cdot\|_\infty) \lesssim (LK)^2 \log \left(\frac{LK}{\epsilon} \right).$$

Combining the above two lemmas, we are now ready to prove Lemma 5.

Proof of Lemma 5. Following Lemma 3.2 of Jiao et al. (2021) or standard M-estimation and empirical process theory (van der Vaart and Wellner, 1996), under sub-Gaussian assumption of the noise ξ, for the ERM estimator $\widehat{g}$ given in (5), we have

$$\begin{aligned}
\sup_{g \in \mathcal{H}_k^\dagger(\alpha; C)} \mathsf{E}\left[(\widehat{g}(W) - g(W))^2 \right] &\lesssim \frac{(\log n)^2 \log N(1/n, \mathcal{F}_{nn}(L, K, B), \|\cdot\|_\infty)}{n} + \inf_{\widetilde{g} \in \mathcal{F}_{nn}(L, K, B)} \|\widetilde{g} - g\|_\infty^2 \\
&\lesssim \frac{(\log n)^3 (LK)^2 \log(LK)}{n} + \left(\frac{L}{\log L} \frac{K}{\log K} \right)^{-4\alpha/k},
\end{aligned}$$

where the second inequality follows from Lemma 12 and 13.

Finally, with a simple bias-variance trade-off argument, we can choose $LK \asymp n^{\frac{k}{2(k+2\alpha)}}$ to obtain the desired rate. □

Before proceeding, we make the following remark regarding the optimality of the results in Theorem 6.

Remark 14. *We believe the conditions in Theorem 6 are not tight. For example, when the nuisance functions belong to certain Hölder balls, the sufficient and necessary Hölder-type condition for the existence of semiparametric efficient estimator of τ_{tot} is $\frac{\alpha_a + \alpha_\mu}{2k} > \frac{1}{4}$. The infinite-order U-statistic estimator of Mukherjee et al. (2017) is the only known semiparametric efficient estimator under the above minimal Hölder-type condition. Yet, their estimators require delicate regularity properties of the estimated nuisance functions, which are difficult to verify for DNNs. It is an interesting open theoretical problem how to achieve semiparametric efficiency under minimal Hölder-type condition even simply for τ_{tot}, when the nuisance functions are estimated by DNNs.*

It is possible to generalize Hölder balls in several directions: e.g. assuming different smoothness exponents in different dimensions of the input (Suzuki, 2019) or composing Hölder functions hierarchically to mimick the composition structure of DNNs (Schmidt-Hieber, 2020) (e.g. Case 5 in Appendix D).

Proof of Theorem 6. Theorem 6 is an immediate consequence of Lemma 5 and Theorem 1. □

C Supplementary Tables and Figures

In this section, we collect tables and figures that are related to Cases 1 – 3 of the simulated experiments and real data analysis of the COMPAS dataset, including Table A1 to Table A8 and Figure A1 to Figure A3.

Table A1: The biases, empirical standard errors (SE) and root mean squared errors (RMSE) of the estimated average total effects, NDE, and NIE, and the coverage probabilities (CP) of their corresponding 95% confidence intervals. $p = 5$ (no irrelevant covariates). The simulation is based on 200 replicates.

		Case 1				Case 2				Case 3			
	Method	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
τ_{tot}	DeepMed	-0.001	0.032	0.032	0.945	-0.004	0.032	0.032	0.955	0.008	0.037	0.038	0.920
	Lasso	0.192	0.089	0.212	0.460	-0.304	0.116	0.325	0.215	0.346	0.079	0.355	0.010
	RF	0.067	0.042	0.079	0.775	-0.078	0.056	0.096	0.950	-0.009	0.042	0.043	0.985
	GBM	-0.015	0.036	0.039	0.940	-0.044	0.055	0.070	0.850	0.019	0.041	0.045	0.930
	Oracle	-0.001	0.029	0.029	0.955	-0.003	0.029	0.029	0.925	-0.001	0.032	0.032	0.935
$\tau_{\mathsf{NDE}}(1)$	DeepMed	0.000	0.027	0.027	0.945	-0.007	0.023	0.024	0.955	0.000	0.026	0.026	0.965
	Lasso	0.130	0.043	0.137	0.220	-0.375	0.059	0.380	0.000	0.226	0.064	0.235	0.050
	RF	0.048	0.029	0.056	0.700	-0.188	0.044	0.193	0.005	0.030	0.038	0.048	0.980
	GBM	-0.040	0.031	0.051	0.770	-0.164	0.046	0.170	0.040	0.011	0.042	0.043	0.920
	Oracle	0.000	0.022	0.022	0.945	-0.002	0.020	0.020	0.985	0.001	0.022	0.022	0.955
$\tau_{\mathsf{NDE}}(0)$	DeepMed	0.000	0.025	0.025	0.940	-0.009	0.026	0.028	0.915	-0.001	0.027	0.027	0.940
	Lasso	0.134	0.043	0.141	0.170	-0.373	0.058	0.377	0.000	0.227	0.064	0.236	0.045
	RF	0.001	0.030	0.030	0.970	-0.186	0.047	0.192	0.020	0.036	0.036	0.051	0.950
	GBM	-0.037	0.030	0.048	0.800	-0.164	0.049	0.171	0.055	0.020	0.044	0.048	0.920
	Oracle	0.000	0.022	0.022	0.955	-0.002	0.019	0.019	0.985	0.001	0.022	0.022	0.950
$\tau_{\mathsf{NIE}}(1)$	DeepMed	-0.001	0.025	0.025	0.960	0.005	0.029	0.029	0.915	0.008	0.031	0.032	0.905
	Lasso	0.058	0.077	0.096	0.875	0.069	0.094	0.117	0.905	0.120	0.045	0.128	0.220
	RF	0.066	0.037	0.076	0.665	0.108	0.059	0.123	0.860	-0.045	0.038	0.059	0.765
	GBM	0.023	0.031	0.039	0.890	0.120	0.064	0.136	0.485	-0.001	0.037	0.037	0.935
	Oracle	-0.001	0.020	0.020	0.975	0.000	0.021	0.021	0.930	-0.002	0.022	0.022	0.920
$\tau_{\mathsf{NIE}}(0)$	DeepMed	-0.001	0.028	0.028	0.940	0.003	0.027	0.027	0.930	0.008	0.029	0.030	0.910
	Lasso	0.062	0.078	0.100	0.870	0.071	0.095	0.119	0.905	0.120	0.045	0.128	0.220
	RF	0.019	0.037	0.042	0.935	0.110	0.058	0.124	0.835	-0.038	0.036	0.052	0.845
	GBM	0.025	0.031	0.040	0.910	0.120	0.059	0.134	0.495	0.008	0.037	0.038	0.935
	Oracle	-0.001	0.019	0.019	0.980	0.000	0.022	0.022	0.920	-0.001	0.023	0.023	0.930

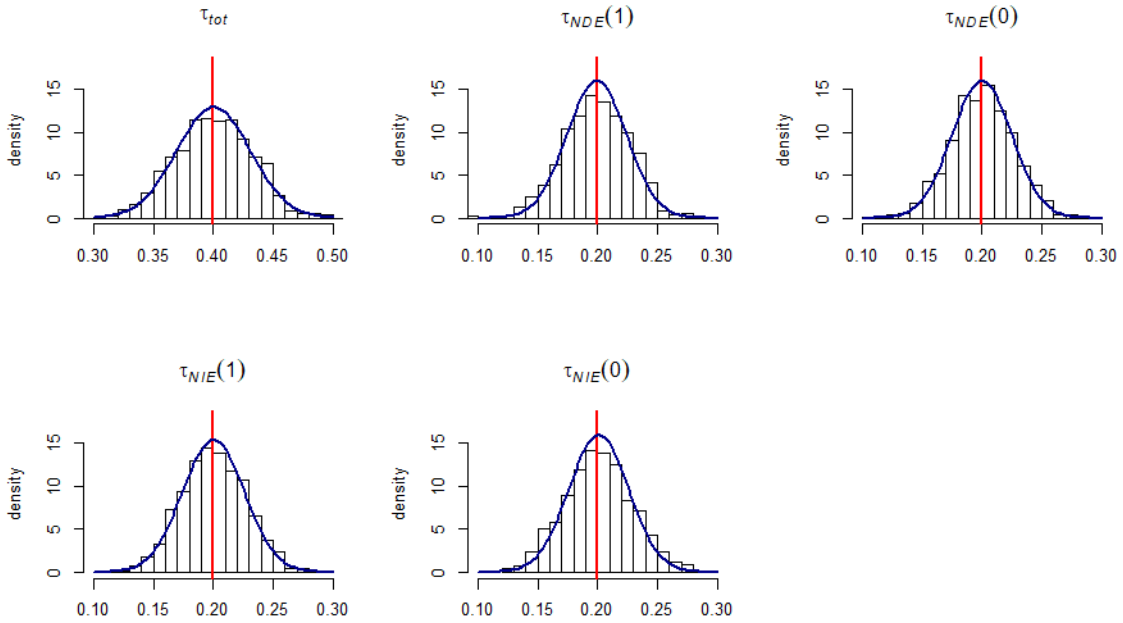

Figure A1: (Case 1) The empirical distributions of the estimated total effects, NIE and NDE by DeepMed. The results are based on 1000 simulation replicates and the number of covariates $p = 5$. The red vertical lines indicate the true effects. The blue curves represent the normal density with the means at the true effects and the estimated standard errors.

Table A2: The biases, empirical standard errors (SE) and root mean squared errors (RMSE) of the total effects, NDE and NIE, and the coverage probabilities (CP) of their corresponding 95% confidence intervals. There exist irrelevant covariates in this setup ($p = 100$). The simulation is based on 200 replicates.

		Input-layer L_1 regularization				No regularization			
		Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
Case 1	τ_{tot}	0.008	0.046	0.047	0.955	0.155	0.086	0.177	0.540
	$\tau_{\mathsf{NDE}}(1)$	0.009	0.043	0.044	0.920	0.061	0.044	0.075	0.745
	$\tau_{\mathsf{NDE}}(0)$	0.006	0.040	0.040	0.920	0.002	0.049	0.049	0.920
	$\tau_{\mathsf{NIE}}(1)$	0.002	0.050	0.050	0.930	0.152	0.090	0.177	0.550
	$\tau_{\mathsf{NIE}}(0)$	-0.001	0.044	0.044	0.945	0.094	0.081	0.124	0.800
Case 2	τ_{tot}	-0.018	0.033	0.038	0.895	-0.033	0.035	0.048	0.875
	$\tau_{\mathsf{NDE}}(1)$	-0.025	0.036	0.044	0.855	-0.266	0.045	0.270	0.000
	$\tau_{\mathsf{NDE}}(0)$	-0.018	0.036	0.040	0.855	-0.306	0.051	0.310	0.000
	$\tau_{\mathsf{NIE}}(1)$	0.000	0.036	0.036	0.950	0.273	0.060	0.280	0.000
	$\tau_{\mathsf{NIE}}(0)$	0.007	0.037	0.038	0.920	0.233	0.055	0.239	0.005
Case 3	τ_{tot}	0.019	0.043	0.047	0.915	0.224	0.055	0.231	0.015
	$\tau_{\mathsf{NDE}}(1)$	0.013	0.035	0.037	0.920	0.075	0.051	0.091	0.600
	$\tau_{\mathsf{NDE}}(0)$	0.016	0.033	0.037	0.925	0.089	0.054	0.104	0.515
	$\tau_{\mathsf{NIE}}(1)$	0.003	0.038	0.038	0.940	0.135	0.064	0.149	0.340
	$\tau_{\mathsf{NIE}}(0)$	0.006	0.036	0.036	0.960	0.149	0.051	0.157	0.155

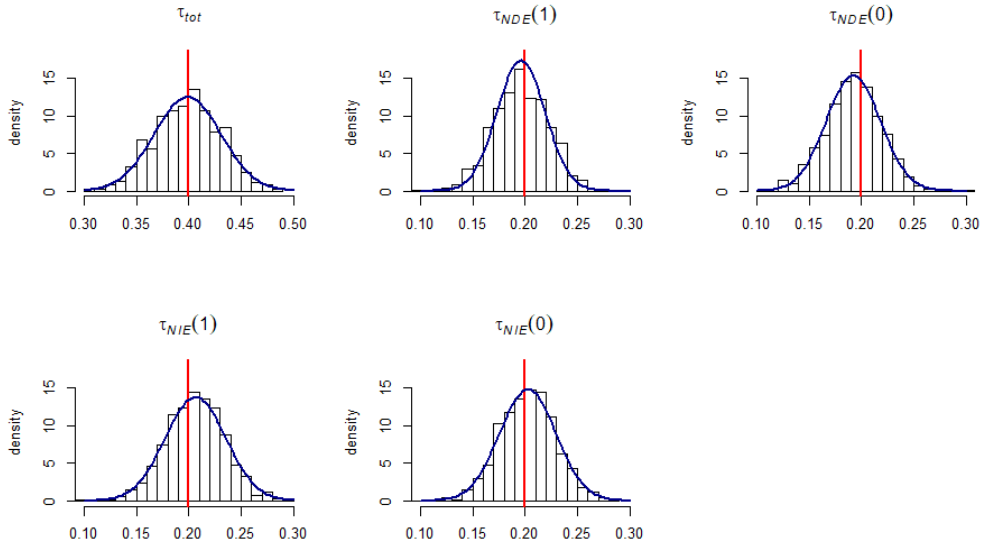

Figure A2: (Case 2) The empirical distributions of the estimated total effects, NIE and NDE by DeepMed. The results are based on 1000 simulation replicates and the number of covariates $p = 5$. The red vertical lines indicate the true effects. The blue curves represent the normal density with the means at the true effects and the estimated standard errors.

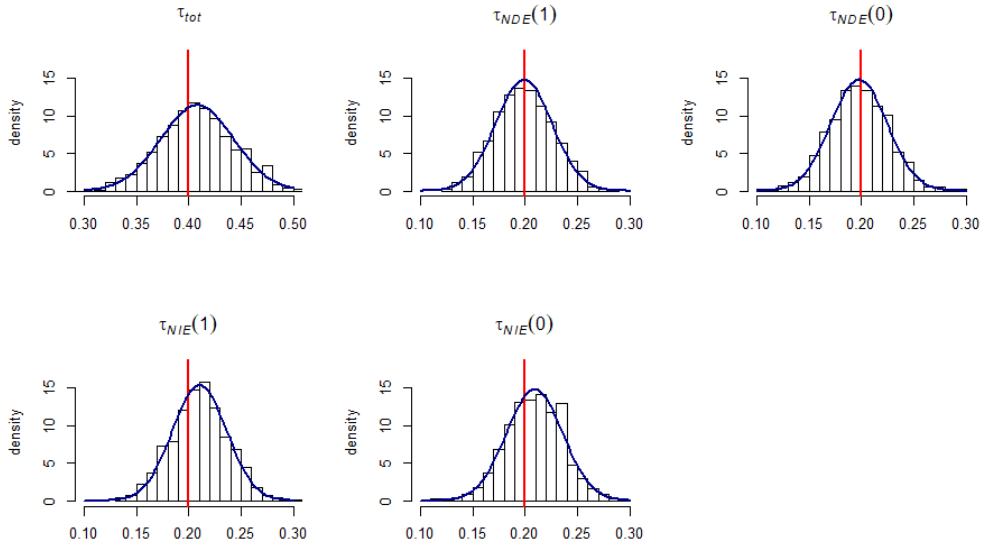

Figure A3: (Case 3) The empirical distributions of the estimated total effects, NIE and NDE by DeepMed. The results are based on 1000 simulation replicates and the number of covariates $p = 5$. The red vertical lines indicate the true effects. The blue curves represent the normal density with the means at the true effects and the estimated standard errors.

D Additional information on synthetic experiments

In all the synthetic experiments, we adopt a 3-fold cross-validation to choose the hyperparameters of DNN, RF and GBM over a grid of candidate values. For DNN, we fix the batch-size as 100, and choose depth L from 1 to 3, width K from 10 to 500, L_1-regularization parameter λ from 0 to 0.4, and epochs from 100 to 500. For RF, we choose the number of trees from 1 to 20, and maximum number of nodes from 10 to 1000. For GBM, we choose the number of trees from 1 to 20, and depth from 10 to 1000. The other hyperparameters are set to the default values in the R packages.

Case 4 (Hölder functions): we repeat the simulation in Case 3 but set $\alpha = 0.6$ to further decrease the smoothness of the Hölder functions. In particular, $\alpha = 0.6$ is close to the limit ($\alpha = 0.5 + \epsilon$ for arbitrarily small $\epsilon > 0$) for DeepMed estimators to be semiparametric efficient based on Theorem 6. Thus we can examine whether surpassing this limit for nuisance function estimates computed by ERM (5) without considering the DNN training process actually translates to practical success of DeepMed. Unfortunately, the results in Table A9 show otherwise. In general, DeepMed still has superior performance than the other competing methods for $p = 5, 20, 100$. However, even at $p = 5$, the biases of the DeepMed estimators are close to their standard errors. As a result, their CPs undercover the true causal parameters (though the CP is not very far from 95%). However, based on Lemma 5, one should be able to estimate all the nuisance functions at rate $O(n^{-1/4})$ as $\alpha > 0.5$ if the nuisance function estimates are ERM (5), which should in turn leads to the semiparametric efficiency of DeepMed NDE/NIE estimators and valid inference.

There are several possible explanations for the DeepMed estimators failing to be semiparametric efficient: (1) it is entirely possible that gradient-based training algorithm such as adam (used in our paper) or stochastic gradient descent (SGD) could find nuisance function estimators from the FNN-ReLU class that are close to the ERM (5) but we just failed to do so in our implementation; or (2) it is a manifestation of a certain "low-frequency-bias" (Rahaman et al., 2019; Hu et al., 2020; Xu et al., 2020) of DNNs trained by gradient-based algorithms or the computational hardness of learning DNNs by gradient-based algorithms (Goel et al., 2020; Chen et al., 2022), which predicts that DNNs trained by off-the-shelf algorithms bias towards functions with lower complexity. We conjecture that (2) is indeed the reason. It suggests that in future works, to fully establish the practically relevant statistical guarantees of DNN-based nonparametric regression or DNN-based causal inference method such as DeepMed, it is important to take the effect of training algorithms into consideration.

Remark 15. *As we have discussed, most established convergence rates of DNNs in the nonparametric statistics literature do not take the potential implicit regularization effect of the training algorithms into account. In this remark, we suggest several possible directions to explore in future works. The so-called Barron space (E et al., 2021) was recently shown to be a natural function space describing the class of neural network functions trained with SGD. If the nuisance functions lie in a Barron space, then the rate of convergence is dimension-independent (E et al., 2021; Chen et al., 2021), which seems to be consistent with the observation that DNNs overcome curse-of-dimensionality in practice. However, as shown in E et al. (2021), the complexity of Barron functions is extremely small, thus casting doubt on if the theoretical claim under Barron spaces is "too good to be true" in fields such as biomedical and social sciences or algorithmic fairness, in which model misspecification bias might have catastrophic consequences. Recent results (Siegel and Xu, 2021) trying to generalize Barron spaces to model more complex functions might be a useful direction to consider in problems related to semiparametric causal inference.*

Case 5 (composition Hölder functions): In the last setting, we consider composition Hölder functions by composing $\eta(x;\alpha) \in \mathcal{H}_1(\alpha; B)$ hierarchically for some constant $B > 0$ as follows:

$$
\begin{aligned}
d(x) &= 0.2\eta\left(\sum_{i=1}^{3} x_i; \alpha\right) + 0.2\eta\left(\sum_{i=1}^{3} \eta(x_i;\alpha); \alpha\right), \\
m(x) &= 0.5\eta\left(\sum_{i=1}^{3} x_i; \alpha\right) + 0.2\eta\left(\sum_{i=1}^{3} \eta(x_i;\alpha); \alpha\right), \\
y(x) &= 0.2\eta\left(\sum_{i=1}^{3} x_i; \alpha\right) + 0.5\eta\left(\sum_{i=1}^{3} \eta(x_i;\alpha); \alpha\right),
\end{aligned}
$$

where we set $\alpha = 1.5$. We choose the "depth" of compositions as 2 for simplicity. The nuisance functions a, f, μ in Case 5 correspond to the composition Hölder functions studied in the seminal work by Schmidt-Hieber (2020). For such function spaces, Schmidt-Hieber (2020) showed that linear estimators cannot achieve minimax optimal estimation rate, yet nonlinear DNN-based estimators can. As shown in Table A10, the DeepMed estimators do exhibit superior performance compared with the other competing methods but they are far from being semiparametric efficient. When $\alpha = 1.5$, at least based on Schmidt-Hieber (2020), the ERM-based DNN regression estimators should converge to

the true function in L_2-norm at a rate faster than $n^{-1/4}$. Then using the meta theorem (Theorem 1), the DeepMed estimators should have been semiparametric efficient. But as in Table A10, our empirical results suggest otherwise. This is another instance that suggests the necessity of developing more refined theoretical properties of DeepMed.

Remark 16. DeepMed *has the option of estimating nuisance functions by other types of machine learning methods, such as those mentioned in Section 4. It is important to develop statistical methodology that can help practitioners decide which method one should use, especially when different methods output qualitatively different results. This is an important research question to pursue as mediation analysis is often followed with critical decision making.*

E A comment on the potential issue of unmeasured confounding in applications related to algorithmic fairness

Finally, we briefly comment on the potential issue of unmeasured confounding in applications related to algorithmic fairness. It is definitely possible to have unmeasured treatment-mediator confounding in real data analysis. But unmeasured treatment-outcome and mediator-outcome confounding may not be huge issues in mediation analysis related to algorithmic fairness because in most cases we have access to all the features used to fit the prediction map, which is the outcome Y in our notation. However, they could be violated when some of the features used to fit the model are concealed to protect data privacy. We did not consider these issues in the real data application but we admit that the final results might be biased due to these unmeasured confounding biases. But the same caveat also applies to most of the other works using mediation analysis in algorithmic fairness.

For fields such as epidemiology or social sciences, more often than not, we do not have the luxury of having access to all important features for the mediator, exposure, and outcome. Thus in general, it is important to incorporate instrumental variables (Frölich and Huber, 2017), valid proxies (Dukes et al., 2021) or other identification strategies (Sun and Ye, 2020) into DeepMed to handle unmeasured confounding.

F Real data analysis (continued)

This section is a continuation of Section 5 of the main text.

In this section, we apply DeepMed to a second dataset and study whether gender has direct effect on personal annual income not mediated by occupation. We use the Adult dataset (https://archive.ics.uci.edu/ml/datasets/adult) from the 1994 Census database in U.S., which includes 48,842 individuals (Kohavi, 1996). We set $D = 1$ for male and $D = 0$ for female. Occupation (M) is a categorical variable containing 14 general types of occupations. The personal annual income is a binary variable, with $Y = 1$ (or $Y = 0$) indicating that an individual makes more (or less) than \$50,000 annually. We also include age, race, education level and employment status as covariates. After removing observations with missing values, the remaining sample size is 45,997.

All the methods find significant NDE of gender on personal annual income (see Table A11). This positive NDE suggests that males tend to have higher income than females, and this cannot be explained by the indirect effect through occupation. In this dataset, the DeepMed estimators again have smaller validation errors than the other competing methods, possibly suggesting smaller biases of the corresponding NDE/NIE estimators (see Table A12).

References for Appendix

Sitan Chen, Adam R Klivans, and Raghu Meka. Learning deep ReLU networks is fixed-parameter tractable. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 696–707. IEEE, 2022.

Ziang Chen, Jianfeng Lu, and Yulong Lu. On the representation of solutions to elliptic PDEs in Barron spaces. *arXiv preprint arXiv:2106.07539*, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21 (1):C1–C68, 2018.

Oliver Dukes, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Proximal mediation analysis. *arXiv preprint arXiv:2109.11904*, 2021.

Weinan E, Chao Ma, and Lei Wu. The Barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, pages 1–38, 2021.

Markus Frölich and Martin Huber. Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1645–1666, 2017.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pages 3587–3596. PMLR, 2020.

Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. *arXiv preprint arXiv:2006.14599*, 2020.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximately low-dimensional manifolds. *arXiv preprint arXiv:2104.06708*, 2021.

Ron Kohavi. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

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

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

James Robins, Lingling Li, Eric Tchetgen Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

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

Jonathan W Siegel and Jinchao Xu. Sharp lower bounds on the approximation rate of shallow neural networks. *arXiv preprint arXiv:2106.14997*, 2021.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4): 1040–1053, 1982.

BaoLuo Sun and Ting Ye. Semiparametric causal mediation analysis with unmeasured mediator-outcome confounding. *arXiv preprint arXiv:2012.06762*, 2020.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Hans Triebel. *Theory of Function Spaces*. Modern Birkhäuser Classics. Springer Basel, 2010.

Aad W van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media, 1996.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

Table A3: The biases, empirical standard errors (SE) and root mean squared errors (RMSE) of the estimated total effects, NDE, and NIE, and the coverage probabilities (CP) of their corresponding 95% confidence intervals. There exist irrelevant covariates in this setup ($p = 20$). The simulation is based on 200 replicates.

		Case 1				Case 2				Case 3			
	Method	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
τ_{tot}	DeepMed	0.001	0.041	0.041	0.950	-0.019	0.030	0.036	0.920	0.010	0.036	0.037	0.945
	Lasso	0.191	0.091	0.212	0.510	-0.318	0.108	0.336	0.155	0.334	0.077	0.343	0.010
	RF	0.035	0.049	0.060	0.955	-0.187	0.076	0.202	0.475	0.100	0.052	0.113	0.545
	GBM	-0.020	0.039	0.044	0.910	-0.139	0.063	0.153	0.445	0.050	0.050	0.071	0.800
	Oracle	-0.002	0.032	0.032	0.925	-0.003	0.029	0.029	0.925	-0.003	0.030	0.030	0.950
$\tau_{\mathsf{NDE}}(1)$	DeepMed	0.001	0.033	0.033	0.915	-0.021	0.026	0.033	0.875	0.003	0.028	0.028	0.955
	Lasso	0.129	0.048	0.138	0.205	-0.378	0.060	0.383	0.000	0.216	0.062	0.225	0.065
	RF	0.033	0.037	0.050	0.890	-0.213	0.043	0.217	0.000	0.088	0.047	0.100	0.545
	GBM	-0.054	0.036	0.065	0.675	-0.228	0.049	0.233	0.005	0.038	0.049	0.062	0.905
	Oracle	-0.002	0.023	0.023	0.925	-0.002	0.020	0.020	0.985	-0.002	0.020	0.020	0.970
$\tau_{\mathsf{NDE}}(0)$	DeepMed	0.010	0.037	0.038	0.900	-0.019	0.028	0.034	0.890	-0.001	0.029	0.029	0.945
	Lasso	0.133	0.048	0.141	0.185	-0.376	0.060	0.381	0.000	0.216	0.064	0.225	0.075
	RF	0.007	0.038	0.039	0.955	-0.212	0.044	0.217	0.000	0.081	0.046	0.093	0.605
	GBM	-0.054	0.038	0.066	0.675	-0.233	0.051	0.239	0.000	0.044	0.051	0.067	0.860
	Oracle	-0.002	0.023	0.023	0.930	-0.002	0.019	0.019	0.985	-0.002	0.020	0.020	0.960
$\tau_{\mathsf{NIE}}(1)$	DeepMed	-0.009	0.032	0.033	0.915	0.000	0.028	0.028	0.955	0.011	0.035	0.037	0.885
	Lasso	0.059	0.078	0.098	0.890	0.058	0.093	0.110	0.920	0.118	0.042	0.125	0.270
	RF	0.028	0.040	0.049	0.965	0.025	0.079	0.083	0.965	0.019	0.033	0.038	0.895
	GBM	0.034	0.035	0.049	0.870	0.093	0.075	0.119	0.755	0.006	0.041	0.041	0.925
	Oracle	0.000	0.021	0.021	0.950	0.000	0.021	0.021	0.930	0.000	0.023	0.023	0.935
$\tau_{\mathsf{NIE}}(0)$	DeepMed	0.002	0.037	0.037	0.915	0.002	0.027	0.027	0.940	0.008	0.031	0.032	0.935
	Lasso	0.062	0.079	0.100	0.870	0.060	0.094	0.112	0.915	0.119	0.042	0.126	0.220
	RF	0.002	0.039	0.039	0.980	0.025	0.078	0.082	0.965	0.012	0.032	0.034	0.920
	GBM	0.033	0.035	0.048	0.855	0.088	0.075	0.116	0.795	0.012	0.040	0.042	0.935
	Oracle	0.000	0.021	0.021	0.940	0.000	0.022	0.022	0.920	-0.001	0.023	0.023	0.925

Table A4: The biases, empirical standard errors (SE) and root mean squared errors (RMSE) of the estimated total effects, NDE, and NIE, and the coverage probabilities (CP) of their corresponding 95% confidence intervals. There exist irrelevant covariates in this setup ($p = 100$). The simulation is based on 200 replicates.

		Case 1				Case 2				Case 3			
	Method	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
τ_{tot}	DeepMed	0.008	0.046	0.047	0.955	-0.018	0.033	0.038	0.895	0.019	0.043	0.047	0.915
	Lasso	0.195	0.093	0.216	0.440	-0.306	0.107	0.324	0.185	0.337	0.079	0.346	0.015
	RF	0.187	0.052	0.194	0.095	-0.205	0.095	0.226	0.560	0.206	0.057	0.214	0.040
	GBM	-0.016	0.035	0.038	0.965	-0.155	0.072	0.171	0.440	0.082	0.053	0.098	0.620
	Oracle	0.000	0.030	0.030	0.945	0.000	0.030	0.030	0.960	0.001	0.031	0.031	0.935
$\tau_{\mathsf{NDE}}(1)$	DeepMed	0.009	0.043	0.044	0.920	-0.025	0.036	0.044	0.855	0.013	0.035	0.037	0.920
	Lasso	0.129	0.050	0.138	0.275	-0.369	0.055	0.373	0.000	0.215	0.067	0.225	0.095
	RF	0.104	0.040	0.111	0.305	-0.216	0.048	0.221	0.000	0.155	0.054	0.164	0.150
	GBM	-0.022	0.038	0.044	0.925	-0.235	0.055	0.241	0.010	0.051	0.056	0.076	0.790
	Oracle	0.001	0.022	0.022	0.950	0.000	0.021	0.021	0.935	0.002	0.022	0.022	0.930
$\tau_{\mathsf{NDE}}(0)$	DeepMed	0.006	0.040	0.040	0.920	-0.018	0.036	0.040	0.855	0.016	0.033	0.037	0.925
	Lasso	0.132	0.050	0.141	0.270	-0.368	0.056	0.372	0.000	0.215	0.067	0.225	0.085
	RF	0.063	0.038	0.074	0.690	-0.213	0.047	0.218	0.005	0.139	0.056	0.150	0.255
	GBM	-0.032	0.037	0.049	0.870	-0.241	0.054	0.247	0.005	0.059	0.057	0.082	0.765
	Oracle	0.001	0.022	0.022	0.940	0.000	0.021	0.021	0.935	0.002	0.022	0.022	0.940
$\tau_{\mathsf{NIE}}(1)$	DeepMed	0.002	0.050	0.050	0.930	0.000	0.036	0.036	0.950	0.003	0.038	0.038	0.940
	Lasso	0.063	0.080	0.102	0.885	0.062	0.089	0.108	0.915	0.122	0.044	0.130	0.200
	RF	0.124	0.050	0.134	0.335	0.008	0.099	0.099	0.945	0.068	0.036	0.077	0.395
	GBM	0.016	0.040	0.043	0.905	0.086	0.084	0.120	0.835	0.023	0.036	0.043	0.905
	Oracle	-0.001	0.021	0.021	0.930	0.000	0.021	0.021	0.950	-0.001	0.021	0.021	0.950
$\tau_{\mathsf{NIE}}(0)$	DeepMed	-0.001	0.044	0.044	0.945	0.007	0.037	0.038	0.920	0.006	0.036	0.036	0.960
	Lasso	0.066	0.081	0.104	0.870	0.063	0.090	0.110	0.915	0.123	0.044	0.131	0.215
	RF	0.084	0.047	0.096	0.645	0.011	0.098	0.099	0.945	0.051	0.034	0.061	0.615
	GBM	0.007	0.038	0.039	0.940	0.080	0.082	0.115	0.845	0.031	0.034	0.046	0.890
	Oracle	-0.001	0.021	0.021	0.935	0.000	0.021	0.021	0.935	-0.001	0.021	0.021	0.960

Table A5: The validation loss of the nuisance functions. The cross-entropy loss is used for fitting $a(d|x, m)$ and $a(d|x)$, and mean squared loss is used for fitting the other nuisance functions. There exist no irrelevant covariates in this setup.

		$a(1\|x, m)$	$a(1\|x)$	$\mu(x, 1, m)$	$\mathsf{E}_0(\mu_1)^*$	$\mu(x, 1)$	$\mu(x, 0, m)$	$\mathsf{E}_1(\mu_0)^*$	$\mu(x, 0)$
Case 1	DeepMed	0.646	0.647	1.151	1.353	2.290	1.172	1.275	2.304
	Lasso	0.660	0.660	5.677	14.889	20.725	5.634	15.099	20.705
	RF	0.662	0.664	3.284	6.322	6.189	3.777	5.618	6.909
	GBM	0.651	0.651	2.344	3.003	3.290	2.383	2.819	3.370
	Oracle	0.639	0.642	1.057	1.031	2.100	1.063	1.033	2.108
Case 2	DeepMed	0.680	0.681	1.309	1.434	2.318	1.305	1.213	2.311
	Lasso	0.694	0.694	8.285	23.037	31.239	8.275	23.117	31.278
	RF	0.694	0.697	5.046	19.915	16.393	4.924	19.291	16.568
	GBM	0.688	0.689	4.587	11.961	8.089	4.441	11.592	7.918
	Oracle	0.670	0.676	1.055	1.037	2.109	1.061	1.039	2.116
Case 3	DeepMed	0.647	0.649	1.35	1.388	2.572	1.376	1.36	2.568
	Lasso	0.662	0.664	9.31	4.612	13.935	8.945	4.584	13.35
	RF	0.657	0.664	4.966	2.728	5.388	5.101	2.647	5.425
	GBM	0.648	0.649	4.087	3.136	4.162	4.131	2.946	4.172
	Oracle	0.637	0.643	1.031	1.019	2.11	1.033	1.02	2.111

$^*\mathsf{E}_0(\mu_1) = \mathsf{E}[\mu(X, D = 1, M)|X, D = 0]$ and $\mathsf{E}_1(\mu_0) = \mathsf{E}[\mu(X, D = 0, M)|X, D = 1]$.

Table A6: The validation loss of the nuisance functions. The cross-entropy loss is used for fitting $a(d|x,m)$ and $a(d|x)$, and mean squared loss is used for fitting the other nuisance functions. There exist irrelevant covariates in this setup ($p = 20$).

		$a(1\|x,m)$	$a(1\|x)$	$\mu(x,1,m)$	$\mathsf{E}_0(\mu_1)^*$	$\mu(1,x)$	$\mu(x,0,m)$	$\mathsf{E}_1(\mu_0)^*$	$\mu(0,x)$
Case 1	DeepMed	0.657	0.659	1.316	2.943	3.241	1.454	2.709	3.319
	Lasso	0.666	0.667	5.583	14.84	20.663	5.541	15.25	20.65
	RF	0.662	0.661	4.449	8.407	8.078	4.958	7.619	8.628
	GBM	0.657	0.658	2.987	3.24	3.976	3.11	3.265	3.992
	Oracle	0.64	0.645	1.014	1.019	2.097	1.009	1.02	2.087
Case 2	DeepMed	0.692	0.691	1.42	1.822	2.432	1.543	1.793	2.533
	Lasso	0.694	0.694	7.98	22.882	31.023	8.003	23.34	30.977
	RF	0.692	0.692	5.708	25.493	22.644	5.618	26.092	21.803
	GBM	0.693	0.693	5.176	17.787	12.627	5.187	18.722	12.227
	Oracle	0.672	0.677	1.014	1.017	2.095	1.01	1.018	2.083
Case 3	DeepMed	0.653	0.652	1.448	2.456	3.214	1.504	2.512	3.172
	Lasso	0.655	0.657	9.581	4.161	13.785	8.699	4.663	13.162
	RF	0.649	0.651	6.626	2.824	6.42	6.789	2.924	6.591
	GBM	0.646	0.647	6.344	3.372	5.14	6.49	3.619	5.426
	Oracle	0.635	0.638	1.021	1.034	2.059	1.025	1.037	2.061

$^*\mathsf{E}_0(\mu_1) = \mathsf{E}[\mu(X, D = 1, M)|X, D = 0]$ and $\mathsf{E}_1(\mu_0) = \mathsf{E}[\mu(X, D = 0, M)|X, D = 1]$.

Table A7: The validation loss of the nuisance functions. The cross-entropy loss is used for fitting $a(d|x,m)$ and $a(d|x)$, and mean squared loss is used for fitting the other nuisance functions. There exist irrelevant covariates in this setup ($p = 100$).

		$a(1\|x,m)$	$a(1\|x)$	$\mu(x,1,m)$	$\mathsf{E}_0(\mu_1)^*$	$\mu(x,1)$	$\mu(x,0,m)$	$\mathsf{E}_1(\mu_0)^*$	$\mu(x,0)$
Case 1	DeepMed	0.672	0.672	1.769	4.154	4.071	1.692	4.723	4.138
	Lasso	0.667	0.667	5.438	14.828	20.607	5.478	15.592	20.607
	RF	0.667	0.668	4.964	9.123	8.836	5.487	8.794	9.530
	GBM	0.664	0.663	3.266	3.659	4.112	3.469	3.684	4.147
	Oracle	0.648	0.652	1.001	1.025	2.069	1.001	1.027	2.066
Case 2	DeepMed	0.695	0.694	1.81	2.583	2.909	1.926	2.615	2.936
	Lasso	0.695	0.695	8.006	22.765	31.272	8.005	23.63	31.239
	RF	0.692	0.692	5.708	25.493	22.644	5.618	26.092	21.803
	GBM	0.688	0.689	4.587	11.961	8.089	4.441	11.592	7.918
	Oracle	0.677	0.68	1.003	1.022	2.067	1.003	1.023	2.065
Case 3	DeepMed	0.679	0.677	2.416	4.563	4.643	2.500	5.098	4.621
	Lasso	0.660	0.662	9.784	4.858	14.364	8.523	4.360	13.000
	RF	0.661	0.663	8.124	3.102	7.808	7.973	2.970	7.586
	GBM	0.653	0.653	7.906	3.320	6.205	7.252	3.101	6.159
	Oracle	0.635	0.641	1.011	1.039	2.144	1.014	1.042	2.144

$^*\mathsf{E}_0(\mu_1) = \mathsf{E}[\mu(X, D = 1, M)|X, D = 0]$ and $\mathsf{E}_1(\mu_0) = \mathsf{E}[\mu(X, D = 0, M)|X, D = 1]$.

Table A8: The validation losses of nuisance functions in real data application to the COMPAS algorithm fairness.

	DeepMed	Lasso	RF	GBM
$a(1\|x,m)$	0.622	0.626	0.638	0.625
$a(1\|x)$	0.648	0.650	0.699	0.650
$\mu(x,1,m)$	4.924	5.480	5.436	5.064
$\mathsf{E}[\mu(X, D = 1, M)\|X = x, D = 0]$	1.636	0.993	0.832	1.579
$\mu(x,1)$	7.265	7.392	7.393	7.378
$\mu(x,0,m)$	3.710	4.108	4.266	3.928
$\mathsf{E}[\mu(X, D = 0, M)\|X = x, D = 1]$	7.582	2.443	0.993	2.012
$\mu(x,0)$	5.197	5.299	5.414	5.269

Table A9: Simulation Case 4 ($\alpha = 0.6$): The biases, empirical standard errors (SE) and root mean squared errors (RMSE) of the estimated average treatment effects and coverage probabilities (CP) of 95% confidence intervals. The simulation is based on 200 replicates.

		$p = 5$				$p = 20$				$p = 100$			
	Method	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
τ_{tot}	DeepMed	0.027	0.047	0.054	0.880	0.035	0.047	0.059	0.880	0.062	0.051	0.080	0.740
	Lasso	0.353	0.087	0.364	0.010	0.335	0.082	0.345	0.020	0.343	0.087	0.354	0.020
	RF	0.004	0.051	0.051	0.995	0.080	0.058	0.099	0.780	0.202	0.065	0.212	0.095
	GBM	0.020	0.050	0.054	0.930	0.055	0.056	0.078	0.830	0.088	0.061	0.107	0.650
	Oracle	-0.002	0.031	0.031	0.955	-0.003	0.031	0.031	0.950	0.002	0.031	0.031	0.930
$\tau_{\mathsf{NDE}}(1)$	DeepMed	0.004	0.038	0.038	0.925	-0.006	0.041	0.041	0.940	0.003	0.047	0.047	0.945
	Lasso	0.227	0.071	0.238	0.095	0.211	0.066	0.221	0.120	0.217	0.073	0.229	0.115
	RF	0.022	0.045	0.050	0.980	0.087	0.054	0.102	0.690	0.175	0.060	0.185	0.140
	GBM	0.017	0.051	0.054	0.905	0.044	0.057	0.072	0.845	0.073	0.062	0.096	0.750
	Oracle	0.000	0.022	0.022	0.965	-0.003	0.022	0.022	0.920	0.002	0.022	0.022	0.940
$\tau_{\mathsf{NDE}}(0)$	DeepMed	0.005	0.034	0.034	0.945	0.005	0.038	0.038	0.960	0.014	0.047	0.049	0.925
	Lasso	0.229	0.069	0.239	0.090	0.213	0.066	0.223	0.115	0.217	0.073	0.229	0.125
	RF	0.032	0.047	0.057	0.960	0.081	0.051	0.096	0.730	0.158	0.061	0.169	0.210
	GBM	0.023	0.050	0.055	0.930	0.049	0.056	0.074	0.865	0.075	0.061	0.097	0.715
	Oracle	0.000	0.022	0.022	0.965	-0.003	0.023	0.023	0.925	0.002	0.021	0.021	0.950
$\tau_{\mathsf{NIE}}(1)$	DeepMed	0.021	0.039	0.044	0.865	0.030	0.042	0.052	0.865	0.047	0.044	0.064	0.830
	Lasso	0.124	0.051	0.134	0.295	0.123	0.048	0.132	0.280	0.126	0.049	0.135	0.260
	RF	-0.029	0.040	0.049	0.895	-0.002	0.039	0.039	0.935	0.044	0.038	0.058	0.760
	GBM	-0.003	0.042	0.042	0.935	0.006	0.045	0.045	0.905	0.013	0.039	0.041	0.935
	Oracle	-0.002	0.024	0.024	0.940	0.000	0.023	0.023	0.930	-0.001	0.023	0.023	0.950
$\tau_{\mathsf{NIE}}(0)$	DeepMed	0.023	0.040	0.046	0.875	0.041	0.041	0.058	0.815	0.059	0.047	0.075	0.730
	Lasso	0.126	0.052	0.136	0.270	0.124	0.048	0.133	0.250	0.126	0.049	0.135	0.250
	RF	-0.019	0.042	0.046	0.935	-0.007	0.035	0.036	0.955	0.027	0.036	0.045	0.885
	GBM	0.004	0.042	0.042	0.920	0.012	0.041	0.043	0.930	0.015	0.039	0.042	0.930
	Oracle	-0.001	0.024	0.024	0.945	0.000	0.023	0.023	0.925	-0.001	0.023	0.023	0.955

Table A10: Simulation Case 5: The biases, empirical standard errors (SE) and root mean squared errors (RMSE) of the estimated average treatment effects and coverage probabilities (CP) of 95% confidence intervals. The simulation is based on 200 replicates.

	Method	Bias	SE	RMSE	CP
τ_{tot}	DeepMed	0.225	0.032	0.227	0.000
	Lasso	0.369	0.032	0.370	0.000
	RF	0.277	0.035	0.279	0.000
	GBM	0.368	0.032	0.369	0.000
	Oracle	-0.003	0.028	0.028	0.960
$\tau_{\mathsf{NDE}}(1)$	DeepMed	0.040	0.023	0.046	0.570
	Lasso	0.113	0.023	0.115	0.005
	RF	0.087	0.029	0.092	0.170
	GBM	0.114	0.024	0.116	0.005
	Oracle	-0.001	0.020	0.020	0.965
$\tau_{\mathsf{NDE}}(0)$	DeepMed	0.044	0.023	0.050	0.470
	Lasso	0.112	0.023	0.114	0.005
	RF	0.083	0.029	0.088	0.200
	GBM	0.114	0.024	0.116	0.005
	Oracle	-0.002	0.020	0.020	0.960
$\tau_{\mathsf{NIE}}(1)$	DeepMed	0.181	0.027	0.183	0.000
	Lasso	0.257	0.026	0.258	0.000
	RF	0.193	0.033	0.196	0.000
	GBM	0.255	0.027	0.256	0.000
	Oracle	-0.001	0.021	0.021	0.955
$\tau_{\mathsf{NIE}}(0)$	DeepMed	0.184	0.028	0.186	0.000
	Lasso	0.256	0.027	0.257	0.000
	RF	0.190	0.034	0.193	0.000
	GBM	0.255	0.027	0.256	0.000
	Oracle	-0.002	0.021	0.021	0.940

Table A11: Real data application to income fairness. The estimated NDE/NIE of gender (D) on income (Y) with occupation (M) as the mediator.

Method	Effect	Estimate	SE	P value
DeepMed	τ_{tot}	0.155	0.004	$< 10^{-16}$
	$\tau_{\mathsf{NDE}}(1)$	0.161	0.007	$< 10^{-16}$
	$\tau_{\mathsf{NDE}}(0)$	0.148	0.004	$< 10^{-16}$
	$\tau_{\mathsf{NIE}}(1)$	0.007	0.002	0.003
	$\tau_{\mathsf{NIE}}(0)$	-0.005	0.005	0.343
Lasso	τ_{tot}	0.171	0.004	$< 10^{-16}$
	$\tau_{\mathsf{NDE}}(1)$	0.165	0.006	$< 10^{-16}$
	$\tau_{\mathsf{NDE}}(0)$	0.155	0.004	$< 10^{-16}$
	$\tau_{\mathsf{NIE}}(1)$	0.016	0.002	3×10^{-11}
	$\tau_{\mathsf{NIE}}(0)$	0.006	0.004	0.160
RF	τ_{tot}	0.092	0.005	$< 10^{-16}$
	$\tau_{\mathsf{NDE}}(1)$	0.153	0.003	$< 10^{-16}$
	$\tau_{\mathsf{NDE}}(0)$	0.114	0.006	$< 10^{-16}$
	$\tau_{\mathsf{NIE}}(1)$	-0.022	0.003	5×10^{-12}
	$\tau_{\mathsf{NIE}}(0)$	-0.060	0.003	$< 10^{-16}$
GBM	τ_{tot}	0.157	0.004	$< 10^{-16}$
	$\tau_{\mathsf{NDE}}(1)$	0.152	0.006	$< 10^{-16}$
	$\tau_{\mathsf{NDE}}(0)$	0.146	0.004	$< 10^{-16}$
	$\tau_{\mathsf{NIE}}(1)$	0.011	0.002	5×10^{-6}
	$\tau_{\mathsf{NIE}}(0)$	0.005	0.004	0.247

Table A12: The validation losses of nuisance functions in real data application to income fairness.

	DeepMed	Lasso	RF	GBM
$a(1\|x,m)$	0.501	0.516	0.560	0.502
$a(1\|x)$	0.600	0.612	0.631	0.600
$\mu(x,1,m)$	0.465	0.493	0.681	0.467
$\mathsf{E}[\mu(X,D=1,M)\|X=x,D=0]$	0.010	0.011	0.024	0.007
$\mu(x,1)$	0.479	0.510	1.040	0.480
$\mu(x,0,m)$	0.285	0.300	0.478	0.287
$\mathsf{E}[\mu(X,D=0,M)\|X=x,D=1]$	0.003	0.005	0.002	0.002
$\mu(x,0)$	0.288	0.306	0.711	0.291