Econometrica Supplementary Material

SUPPLEMENT TO "DEEP NEURAL NETWORKS FOR ESTIMATION AND INFERENCE"

(*Econometrica*, Vol. 89, No. 1, January 2021, 181–213)

MAX H. FARRELL
Booth School of Business, University of Chicago

TENGYUAN LIANG
Booth School of Business, University of Chicago

SANJOG MISRA
Booth School of Business, University of Chicago

This supplement to "Deep Neural Networks for Estimation and Inference" contains results from a simulation study of the finite sample properties of deep neural networks and their use in semiparametric causal inference. The code (in Python/Tensorflow) used for the simulation exercise is available.

S.1. SET UP

WE STUDY inference on the average treatment effect, $\tau = \mathbb{E}[Y(0) - Y(1)]$, under different data generating processes (DGPs). In all cases, the data generating process obeys the ignorability and unconfoundedness assumptions stated in the main text. In each DGP, we take $n = 10{,}000$ i.i.d. samples and use 1000 replications. For either $\dim(\boldsymbol{X}) = d = 20$ or 100, $\boldsymbol{X}$ includes a constant term and d independent uniform random variables, $\mathcal{U}(0, 1)$. Treatment assignment is Bernoulli with probability $p(\boldsymbol{x})$, where $p(\boldsymbol{x})$ is the propensity score. We consider both (i) randomized treatments with $p(\boldsymbol{x}) = 0.5$ and (ii) observational data with $p(\boldsymbol{x}) = (1 + \exp(-\boldsymbol{\alpha}_p'\boldsymbol{x}))^{-1}$, where $\alpha_{p,1} = 0.09$ and the remainder are drawn once as $\mathcal{U}(-0.55, 0.55)$, and then fixed for the replications. For $d = 100$, we maintain $\|\boldsymbol{\alpha}_p\|_0 = 20$ for the simplicity. These generate propensities with an approximate range of approximately $(0.30, 0.75)$ and mean roughly 0.5.

Given covariates and treatment assignment, the outcomes are generated according to

$$y_i = \mu_0(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)t_i + \varepsilon_i, \qquad \mu_0(\boldsymbol{x}) = \boldsymbol{\alpha}_\mu'\boldsymbol{x} + \boldsymbol{\beta}_\mu'\varphi(\boldsymbol{x}), \qquad \tau(x_i) = \boldsymbol{\alpha}_\tau'\boldsymbol{x} + \boldsymbol{\beta}_\tau'\varphi(\boldsymbol{x}),$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$ and $\varphi(\boldsymbol{x})$ are second-degree polynomials including pairwise interactions. For $\mu_0(\boldsymbol{x})$ and $\tau(\boldsymbol{x})$, we consider two cases: linear and nonlinear models. In both cases, the intercepts are $\alpha_{\mu,1} = 0.09$ and $\alpha_{\tau,1} = -0.05$ and slopes are drawn (once) as $\alpha_{\mu,k} \sim \mathcal{N}(0.3, 0.7)$ and $\alpha_{\tau,k} \sim \mathcal{U}(0.1, 0.22)$, $k = 2, \ldots, d + 1$. The linear models set $\boldsymbol{\beta}_\mu = \boldsymbol{\beta}_\tau = \boldsymbol{0}$ while the nonlinear models take $\beta_{\mu,k} \sim \mathcal{N}(0.01, 0.3)$ and $\beta_{\tau,k} \sim \mathcal{U}(-0.05, 0.06)$. Altogether this yields eight designs: $d = 20$ or 100, $p(\boldsymbol{x})$ constant or not, and outcome models linear or nonlinear.

For each DGP, we consider a variety of network architectures, all of which are multilayer perceptrons (MLPs) with the rectified linear unit (ReLU) activation function, following the theoretical results in the main paper. These architectures are variants of those used in the empirical application, reported in the main paper. All networks vary in their depth and width, which are listed in Table S.I.

Max H. Farrell: max.farrell@chicagobooth.edu
Tengyuan Liang: tengyuan.liang@chicagobooth.edu
Sanjog Misra: sanjog.misra@chicagobooth.edu

 https://doi.org/10.3982/ECTA16901

TABLE S.I

MONTE CARLO ARCHITECTURES EXPLORED

Architecture	Structure
1	{20, 15, 5}
2	{60, 30, 20}
3	{80, 80, 80}
4	{20, 15, 10, 5}
5	{60, 30, 20, 10}
6	{80, 80, 80, 80}
7	{20, 15, 15, 10, 10, 5}
8	{60, 30, 20, 20, 10, 5}
9	{80, 80, 80, 80, 80, 80}

S.2. RESULTS

Tables S.II and S.III show the results for all eight DGPs. Table S.II shows DGPs with constant propensity score, that is, randomized experiments, while Table S.III shows results mimicking observational data. Overall, the results reported show excellent performance of deep learning based semiparametric inference. The bias is minimal (and in most cases is likely dominated by monte carlo error) and the coverage is quite accurate, while the interval length is under control. Notice that most architectures yield similar results with no architecture dominating the others. Further, the coverage and interval length are fairly similar with the more complex architecture not exhibiting any systematic patterns of length inflation. None of these results employ regularization. Our own preliminary exploration of dropout and other forms of regularization found expected departures from

TABLE S.II

SIMULATIONS RESULTS—CONSTANT PROPENSITY SCORE

		20 Covariates			100 Covariates		
Model	Architecture	Bias	IL	Coverage	Bias	IL	Coverage
Linear	1	0.00027	0.079	0.947	0.00067	0.080	0.946
	2	−0.00032	0.079	0.951	0.00012	0.080	0.958
	3	−0.00025	0.079	0.955	−0.00167	0.080	0.939
	4	−0.00068	0.079	0.949	0.00038	0.080	0.949
	5	0.00008	0.079	0.945	−0.00219	0.080	0.929
	6	0.00007	0.079	0.955	−0.00010	0.080	0.946
	7	0.00128	0.079	0.952	−0.00041	0.080	0.944
	8	0.00108	0.079	0.949	−0.00088	0.080	0.941
	9	0.00021	0.078	0.948	−0.00080	0.081	0.953
Nonlinear	1	0.00087	0.081	0.946	−0.00067	0.163	0.940
	2	0.00015	0.079	0.954	0.00093	0.153	0.927
	3	−0.00072	0.079	0.940	0.00245	0.148	0.926
	4	0.00101	0.080	0.945	−0.00087	0.165	0.956
	5	0.00027	0.079	0.935	−0.00190	0.154	0.923
	6	−0.00025	0.079	0.929	−0.00117	0.146	0.902
	7	−0.00052	0.080	0.947	0.00091	0.165	0.941
	8	0.00077	0.079	0.938	0.00201	0.153	0.927
	9	−0.00013	0.079	0.940	0.00049	0.154	0.936

TABLE S.III

SIMULATIONS RESULTS—NONCONSTANT PROPENSITY SCORE

		20 Covariates			100 Covariates		
Model	Architecture	Bias	IL	Coverage	Bias	IL	Coverage
Linear	1	−0.00202	0.080	0.948	0.0009	0.081	0.955
	2	0.00011	0.079	0.946	0.0007	0.081	0.945
	3	−0.00130	0.079	0.964	−0.0001	0.081	0.937
	4	−0.00106	0.079	0.945	0.0002	0.081	0.933
	5	−0.00083	0.079	0.951	−0.0004	0.081	0.944
	6	−0.00068	0.079	0.955	0.0001	0.081	0.924
	7	−0.00119	0.079	0.953	−0.0001	0.081	0.942
	8	−0.00056	0.079	0.952	−0.0008	0.081	0.939
	9	−0.00096	0.079	0.948	−0.0007	0.081	0.952
Nonlinear	1	−0.00076	0.081	0.946	−0.00279	0.164	0.937
	2	−0.00122	0.080	0.939	0.00020	0.155	0.941
	3	−0.00074	0.080	0.926	−0.00080	0.148	0.914
	4	−0.00171	0.081	0.940	−0.00184	0.166	0.938
	5	−0.00135	0.080	0.952	−0.00103	0.154	0.912
	6	−0.00075	0.080	0.950	−0.00174	0.147	0.905
	7	−0.00153	0.081	0.928	−0.00377	0.165	0.929
	8	0.00082	0.080	0.953	0.00031	0.154	0.919
	9	−0.00127	0.080	0.931	−0.00094	0.156	0.917

nonregularized models. In most, but not all, cases the coverage remained accurate, but with increased bias and interval length compared to Tables S.II and S.III. The results preach caution when applying regularization in applications.

Co-editor Ulrich K. Müller handled this manuscript.

Manuscript received 13 December, 2018; final version accepted 6 August, 2020; available online 3 September, 2020.