Interpretable Deep Causal Learning for Moderation Effects

Alberto Caron [1 2] **Gianluca Baio** [1] **Ioanna Manolopoulou** [1]

Abstract

In this extended abstract paper, we address the problem of interpretability and targeted regularization in causal machine learning models. In particular, we focus on the problem of estimating *individual causal/treatment effects* under observed confounders, which can be controlled for and moderate the effect of the treatment on the outcome of interest. Black-box ML models adjusted for the causal setting perform generally well in this task, but they lack interpretable output identifying the main drivers of treatment heterogeneity and their functional relationship. We propose a novel deep counterfactual learning architecture for estimating individual treatment effects that can simultaneously: i) convey targeted regularization on, and produce quantify uncertainty around the quantity of interest (i.e., the *Conditional Average Treatment Effect*); ii) disentangle baseline *prognostic* and *moderating* effects of the covariates and output interpretable score functions describing their relationship with the outcome. Finally, we demonstrate the use of the method via a simple simulated experiment[1].

1. Introduction

In the past years, there has been a growing interest towards applying ML methods for causal inference. Disciplines such as precision medicine and socio-economic sciences inevitably call for highly personalized decision making when designing and deploying policies. Although in these fields exploration of policies in the real world through randomized experiments is costly, in order to answer counterfactual questions such as "what would have happened if individual i undertook medical treatment A instead of treatment B"

one can rely on observational data, provided that the confounding factors can be controlled for. Black-box causal ML models proposed in many recent contributions perform remarkably well in the task of estimating individual counterfactual outcomes, but significantly lack interpretability, which is a key component in the design of personalized treatment rules. This is because they jointly model the outcome dependency on the covariates and on the treatment variable. Knowledge of what are the main moderating factors of a treatment can unequivocally lead to overall better policy design, as moderation effects can be leveraged to achieve higher cumulative utility when deploying the policy (e.g., by avoiding treating patient with uncertain or borderline response, better treatment allocation on budget/resources constraints, ...). Another main issue of existing causal ML models, related to the one of interpretability, is carefully designed regularization (Nie & Wager, 2020; Hahn et al., 2020; Caron et al., 2022b). Large observational studies generally include measurements on a high number of pre-treatment covariates, and disentangling prognostic[2] and moderating effects allows the application of targeted regularization on both, that avoids incurring in unintended finite sample bias and large variance (see (Hahn et al., 2020) for a detailed discussion on *Regularization Induced Confounding* bias). This is useful in many scenarios where treatment effect is believed to be a sparser and relatively less complex function of the covariates compared to the baseline prognostic effect, so it necessitates carefully tailored regularization.

1.1. Related Work

Among the most influential and recent contributions on ML regression-based techniques for individualized treatment effects learning, we particularly emphasize the work of (Johansson et al., 2016; Shalit et al., 2017; Yao et al., 2018) on deep learning models, (Alaa & van der Schaar, 2017; 2018) on Gaussian Processes, (Hahn et al., 2020; Caron et al., 2022b) on Bayesian Additive Regression Trees, and finally the literature on the more general class of Meta-Learners models (Künzel et al., 2017; Nie et al., 2020). We refer the reader to (Caron et al., 2022a) for a detailed review of the above methods.

---

[1]Department of Statistical Science, University College London, London, UK [2]The Alan Turing Institute, London, UK. Correspondence to: Alberto Caron <alberto.caron.19@ucl.ac.uk>.

[1]Code for full reproducibility can be found at https://github.com/albicaron/ICNN.

[2]Prognostic effect is defined as the baseline effect of the covariates on the outcome, in absence of treatment.

$$\mathbf{X} = f_X(\varepsilon_X)$$
$$A = f_A(\mathbf{X}, \varepsilon_A)$$
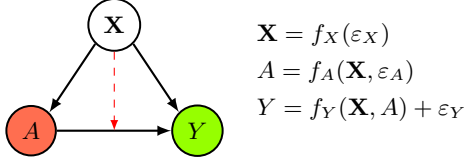$$Y = f_Y(\mathbf{X}, A) + \varepsilon_Y$$

Figure 1. Causal DAG and set of structural equations describing a setting that satisfies the *backdoor criterion*. The underlying assumption is that conditioning on the confounders $\mathbf{X}$ is sufficient to identify the causal effect $A \rightarrow Y$. Models generally assume mean-zero additive error term for the outcome equation. The red arrow in the DAG represent the moderating effect of $\mathbf{X}$ in the $A \rightarrow Y$ relationship.

In particular we build on top of contributions by (Nie et al., 2020; Hahn et al., 2020; Caron et al., 2022b), that have previously addressed the two issues of targeted regularization in causal ML. Our work proposes a new deep architecture that can separate baseline prognostic and treatment effects, and, by borrowing ideas from recent work on Neural Additive Models (NAMs) (Agarwal et al., 2021), a deep learning version of Generalized Additive Models, can output interpretable score functions describing the impact of each covariate in terms of their prognostic and treatment effects.

2. Problem Framework

In this section we briefly introduce the main notation setup for causal effects identification and estimation under observed confounders scenarios, by utilizing the framework of Structural Causal Models (SCMs) and *do*-calculus (Pearl, 2009). We assume we have access to data of observational nature described by the tuple $\mathbb{D}_i = \{\mathbf{X}_i, A_i, Y_i\} \sim p(\cdot)$, with $i \in \{1, ..., N\}$, where $\mathbf{X}_i \in \mathcal{X}$ is a set of covariates, $A_i \in \mathcal{A}$ a binary manipulative variable, and $Y_i \in \mathbb{R}$ is the outcome. We assume then that the causal relationships between the three variables are fully described by the SCM depicted in Figure 1, both in the forms of causal DAG and set of structural equations. Notice in particular that we assume, in line with most of the literature, zero-mean additive error structure for the outcome equation. The ultimate goal is to identify and estimate the Conditional Average Treatment Effects (CATE), defined as the effect of intervening on the manipulative variable A_i, by setting equal to some value a (or $do(A_i = a)$ in the *do*-calculus notation), on the outcome Y_i, conditional on covariates $\mathbf{X}_i$ (i.e., conditional on patient's characteristics, ...). In the case of binary A_i, CATE is defined as:

$$\texttt{CATE:} \quad \tau(\boldsymbol{x}_i) = \mathbb{E}[Y_i \mid do(A_i = 1), \boldsymbol{X}_i = \boldsymbol{x}]$$
$$- \mathbb{E}[Y_i \mid do(A_i = 0), \boldsymbol{X}_i = \boldsymbol{x}]. \quad (1)$$

In order to identify the quantity in (1) we make two standard assumptions. The first assumption is that there are

no unobserved confounders — or equivalently in Pearl's terminology, that $\boldsymbol{X}_i$ satisfies the *backdoor criterion*. The second assumption is *common support*, which states that there is no deterministic selection into either of the treatment arms conditional on the covariates, or equivalently that $p(A_i = 1|\boldsymbol{X}_i = \boldsymbol{x}) \in (0, 1)$, $\forall i$. The latter guarantees that we could theoretically observe data points with $\boldsymbol{X}_i = \boldsymbol{x}$ in each of the two arms of A. Under these two assumptions, we can identify CATE $\tau(\boldsymbol{x}_i)$ in terms of observed quantities only, replacing the *do*-operator in (1) with the factual A_i, by conditioning on $\boldsymbol{X}_i$:

$$\mathbb{E}[Y_i|do(A_i = a), \boldsymbol{X}_i = \boldsymbol{x}] = \mathbb{E}[Y_i|A_i = a, \boldsymbol{X}_i = \boldsymbol{x}].$$

Once CATE is identified as above, there are different ways in which it can be estimated in practice. We will briefly describe few of them in the next section.

3. Targeted CATE estimation

Very early works in the literature on CATE estimation proposed fitting a single model $\hat{f}_Y(\boldsymbol{X}_i, A_i)$ (S-Learners). The main drawback of S-Learners is that they are unable to account for any group-specific distributional difference, which becomes more relevant the stronger the selection bias is. Most of the subsequent contributions instead suggested splitting the sample into treatment subgroups and fit separate, arm-specific models $\hat{f}_{Y_a}(\boldsymbol{x}_i)$ (T-Learners). While T-Learners are able to account for distributional variation attributable to A_i, they are less sample efficient, prone to CATE overfitting and to regularization induced confounding bias (Künzel et al., 2017; Hahn et al., 2020; Caron et al., 2022b). Finally some of the most recent additions to the literature (Hahn et al., 2020; Nie et al., 2020; Caron et al., 2022b) proposed using (Robinson, 1988) re-parametrization of the outcome function, which reads:

$$\texttt{Robinson:} \quad Y_i = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)A_i + \varepsilon_i, \quad (2)$$

where $\mu(\boldsymbol{x}_i) = \mathbb{E}[Y_i \mid do(A_i = 0), \boldsymbol{X}_i = \boldsymbol{x}]$ is the prognostic effect function and $\tau(\boldsymbol{x}_i)$ is the CATE function as defined in (1). The distinctive trait of Robinson's parametrization is that the outcome function explicitly includes the function of interest, i.e. CATE $\tau(\boldsymbol{x}_i)$, while in the usual S- or T-Learner (and subsequent variations of these) parametrizations CATE is implicitly obtained post-estimation as $\hat{\tau}(\boldsymbol{x}_i) = \hat{f}_1(\boldsymbol{x}_i) - \hat{f}_0(\boldsymbol{x}_i)$. This means that (2) is able to differentiate between the baseline prognostic effect $\mu(\boldsymbol{x}_i)$ (in absence of treatment) and moderating effects embedded in the CATE function $\tau(\boldsymbol{x}_i)$, of the covariates. As a consequence, by utilizing (2), one can convey different degree of regularization when estimating the two functions. This is particularly useful as CATE is usually believed to display simpler patterns than $\mu(\boldsymbol{x}_i)$; so by estimating it separately, one is able to apply stronger targeted regularization.
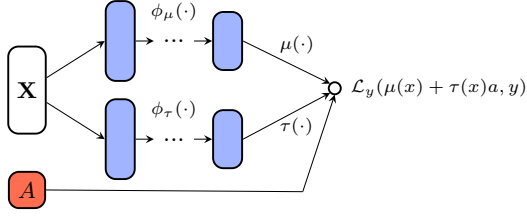
Figure 2. Intuitive TCNN structure. The deep architecture is modelled through a sample efficient, tailored loss function based on Robinson's parametrization.

3.1. Interpretable Causal Neural Networks

Following (Robinson, 1988), and the more recent work by (Nie et al., 2020; Hahn et al., 2020; Caron et al., 2022b), we propose a very simple deep learning architecture for interpretable and targeted CATE estimation, based on Robinson parametrization. The architecture is made of two separable neural nets blocks that respectively learn the prognostic function $\mu(\boldsymbol{x}_i)$ and the CATE function $\tau(\boldsymbol{x}_i)$, but are "reconnected" at the end of the pipeline to minimize a single loss function, unlike T-Learners which instead minimize separate loss functions on $f_1(\cdot)$ and $f_0(\cdot)$. Our target loss function to minimize is generally defined as follows:

$$\min_{\mu(\cdot),\tau(\cdot)} \mathcal{L}_y\big(\mu(\boldsymbol{x}) + \tau(\boldsymbol{x})a, y\big) \ , \qquad (3)$$

where $\mathcal{L}_y(\cdot)$ can be any standard loss function (e.g., MSE, negative log-likelihood,...). Through its separable block structure, the model allows the design of different NN architectures for learning $\mu(\cdot)$ and $\tau(\cdot)$ while preserving sample efficiency (i.e., avoiding sample splitting as in T-Learners), and to produce uncertainty measures around CATE $\tau(\cdot)$ directly. Thus, if $\tau(\cdot)$ is believed to display simple moderating patterns as a function of $\boldsymbol{X}_i$, a shallower NN structure with less hidden layers and units, and more aggressive regularization (e.g., higher regularization rate or dropout probabilities), can be specified, while retaining higher level of complexity in the $\mu(\cdot)$ block. We generally refer to this model as Targeted Causal Neural Network (TCNN) for simplicity from now onwards. Figure 2 provide a simple visual representation. While in this work we focus on binary intervention variables A_i for simplicity, TCNN can be easily extended to multi-category A_i by adding extra blocks to the structure in Figure 2.

In addition to the separable structure, and in order to guarantee higher level of interpretability on prognostic and moderating factors, we also propose using a recently developed neural network version of Generalized Additive Models (GAMs), named Neural Additive Models (NAMs) (Agarwal et al., 2021), as the two $\mu(\cdot)$ and $\tau(\cdot)$ NN building blocks of TCNN. We refer to this particular version of TCNN as Interpretable Causal Neural Network (ICNN). Contrary to normal NNs, which fully "connect" inputs to every nodes in the first hidden layer, NAMs "connect" each single input to its own NN structure and thus outputs input-specific *score functions*, that fully describe the predicted relationship between each input and the outcome. NAM's score functions have an intuitive interpretation as *Shapley values* (Shapley, 1953): how much of an impact each input has on the final predicted outcome. The NAM architecture in ICNN allows us to identify and estimate the Shapley values for each covariates as prognostic $\mu(\cdot)$ and/or moderating $\tau(\cdot)$ effects, and quantify uncertainty around them as well. Naturally, the downside of NAMs is that they might miss out on interaction terms among the covariates. These could possibly be constructed and added manually as additional inputs, although this is not particularly convenient nor computationally ideal.

We conclude the section by highlighting similarities and differences between TCNN (and ICNN) and other popular methods employing Robinson's parametrization. Differently than R-Learner (Nie et al., 2020), TCNN is not a multi-step plug-in (and cross-validated) estimator and does not envisage the use of propensity score. Instead, similarly to Bayesian Causal Forest (BCF) (Hahn et al., 2020; Caron et al., 2022b), estimation in TCNN is carried out in a single, more sample efficient step, although BCF is inherently Bayesian and relatively computationally intensive. To obtain better coverage properties in terms of uncertainty quantification in both TCNN and ICNN, we implement the MC dropout technique (Gal & Ghahramani, 2016) in both $\mu(\cdot)$ and $\tau(\cdot)$ blocks to perform approximate Bayesian inference, that is, we re-sample multiple times from the NN model with dropout layers to build an approximate posterior predictive distribution. This produces credible intervals around CATE estimates $\tau(\cdot)$ in a very straightforward way, and, in ICNN specifically, credible intervals around each inputs' score function, as we will show in the experimental section.

4. Experiments

We hereby present results from a simple simulated experiment on CATE estimation, to compare TCNN and ICNN performance against some of the state of the art methods. In addition, we demonstrate how ICNN with MC dropout in particular can be employed to produce highly interpretable score function measures, fully describing the estimated moderating effects of the covariates $\boldsymbol{x}_i$ in $\tau(\cdot)$, and uncertainty around them. For performance comparison we rely on the root *Precision in Estimating Heterogeneous Treatment Effects* (PEHE) metric (Hill, 2011), defined as:

$$\sqrt{\text{PEHE}_\tau} = \sqrt{\mathbb{E}\big[(\hat{\tau}_i(\boldsymbol{x}_i) - \tau_i(\boldsymbol{x}_i))^2\big]} \ , \qquad (4)$$

and the list of models we compare include: S-Learner version of NNs (S-NN); T-Learner version of NNs (T-NN);

Model	Train $\sqrt{\text{PEHE}_\tau}$	Test $\sqrt{\text{PEHE}_\tau}$
S-NN	1.046 ± 0.007	1.076 ± 0.007
T-NN	1.021 ± 0.002	1.074 ± 0.002
R-CF	1.467 ± 0.002	1.494 ± 0.002
R-NN	0.706 ± 0.003	0.712 ± 0.003
R-NAM	0.787 ± 0.002	0.787 ± 0.002
TCNN	$\mathbf{0.361 \pm 0.001}$	$\mathbf{0.362 \pm 0.001}$
ICNN	$\mathbf{0.328 \pm 0.001}$	$\mathbf{0.331 \pm 0.001}$

Table 1. Performance on simulated experiment, measured as 70%-30% train-test set $\sqrt{\text{PEHE}_\tau}$. Bold indicates better performance.

Causal Forest (Wager & Athey, 2018), a particular type of R-Learner (R-CF); a "unique-block", fully connected NN that uses Robinson's parametrization minimizing the loss function in (3) (R-NN); a "unique-block" NAM, again minimizing the loss function in (3) (R-NAM); our TCNN with fully connected NN blocks; and ICNN. S-NN, T-NN and R-NN all feature two [50, 50] hidden layers. R-NAM features two [20, 20] hidden layers for each input. TCNN features two [50, 50] hidden layers in the $\mu(\cdot)$ block, and one [20] hidden layer in the $\tau(\cdot)$ block. ICNN features two [20, 20] hidden layers in the $\mu(\cdot)$ block, and one [50] hidden layer in the $\tau(\cdot)$ block, for each input.

We simulate $N = 2000$ data points on $P = 10$ correlated covariates, with binary A_i and continuous Y_i. The experiment was run for $B = 100$ replications and results on 70%-30% train-test sets average $\sqrt{\text{PEHE}_\tau}$, plus 95% Monte Carlo errors, can be found in Table 1. The full description of the data generating process utilized for this simulated experiment can be found in the appendix Section A. NN models minimizing the Robinson loss function in (3) perform considerably better than S- and T-Learner baselines on this particular example, especially TCNN and ICNN that present the additional advantage of conveying targeted regularization. Considering the ICNN model only, we can then access the score functions on the $\tau(\cdot)$ NAM block that describe the moderating effects of the covariates $\boldsymbol{x}_i$. In particular in Figure 3 we plot the score function of the first covariate $X_{i,1}$ on CATE $\tau(\cdot)$, plus the approximate Bayesian credible intervals generated through MC dropout resampling (Gal & Ghahramani, 2016). In this specific simulated example, CATE function is generated as $\tau(\boldsymbol{x}_i) = 3 + 0.8 X_{i,1}^2$. So only $X_{i,1}$, out of all $P = 10$ covariates, drives the simple heterogeneity patterns in treatment response across individuals, in a quadratic form. As Figure 3 shows, ICNN is able in this example to learn a score function that very closely approximates the underlying true relationship $0.8 X_{i,1}^2$, and quantifies uncertainty around it. Naturally, in a different simulated setup with strong interaction terms among the covariates, performance of ICNN would inevitably deteriorate compared to the other versions of NN and models
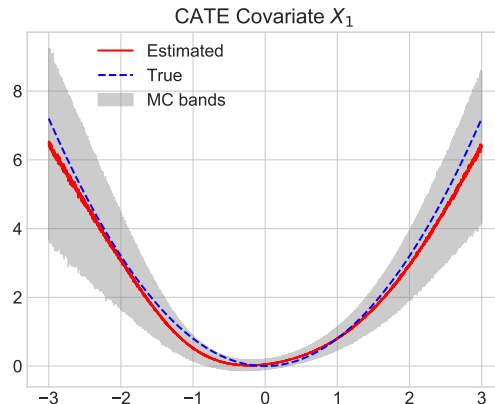

Figure 3. Score function output from ICNN model relative to covariate X_1, depicting its moderating effect on CATE, plus MC dropout generated credible intervals.

considered here. Thus, performance and interpretability in this type of scenario would certainly constitute a trade-off.

5. Conclusion

In this extended abstract paper, we have addressed the issue of interpretability and targeted regularization in causal machine learning models for the estimation of heterogeneous/individual treatment effects. In particular we have proposed a novel deep learning architecture (TCNN) that is able to convey regularization and quantify uncertainty when learning the CATE function, and, in its ICNN version, to output interpretable score function describing the estimated prognostic and moderation effects of the covariates $\boldsymbol{X}_i$. Finally, we have briefly demonstrated the use of TCNN and ICNN by comparing their performance against some of the popular methods for CATE estimation on a simple simulated experiment, where we have also illustrated how score functions are very intuitive and interpretable measures for moderation effects analysis.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. Neural additive models: Interpretable machine learning with neural nets. In *Proceedings of the 35th International Conference on Neu-*

ral Information Processing Systems, volume 34, pp. 4699–4711, 2021.

Alaa, A. and van der Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 129–138, 2018.

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task Gaussian Processes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3427–3435, 2017.

Athey, S. and Wager, S. Policy Learning With Observational Data. *Econometrica*, 89(1):133–161, January 2021.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pp. 1613–1622, 2015.

Caron, A., Baio, G., and Manolopoulou, I. Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, pp. 1–35, 2022a.

Caron, A., Baio, G., and Manolopoulou, I. Shrinkage bayesian causal forests for heterogeneous treatment effects estimation. *Journal of Computational and Graphical Statistics*, pp. 1–13, 2022b.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Chipman, H. A., George, E. I., and McCulloch, R. E. BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1): 266–298, 03 2010.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 1050–1059, 2016.

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.*, 13(1):163–182, 03 2018.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. *Bayesian Analysis*, 15(3):965 – 1056, 2020.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1414–1423, 2017.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hodson, R. Precision medicine. *Nature*, 547(7619), 2016.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 3020–3029, 2016.

Kitagawa, T. and Tetenov, A. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Künzel, S., Sekhon, J., Bickel, P., and Yu, B. Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116, 06 2017.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6405–6416. Curran Associates Inc., 2017.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 09 2020.

Nie, X., Brunskill, E., and Wager, S. Learning when-to-treat policies. *Journal of the American Statistical Association*, 0(ja):1–58, 2020.

Pearce, T., Leibfried, F., and Brintrup, A. Uncertainty in neural networks: Approximately bayesian ensembling. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 234–244, 2020.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Peters, J., Janzing, D., and Schlkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Rubin, D. B. Bayesian inference for causal effects: The role of randomization. *Ann. Statist.*, 6(1):34–58, 01 1978.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, volume 70, pp. 3076–3085, 2017.

Shapley, L. S. A value for n-person games. In *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems 31*, pp. 2633–2643, 2018.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

A. Data Generating Process

In this appendix section we briefly describe the data generating process utilized for the simulated experiment in Section 4. We generated $N = 2000$ data points on $P = 10$ correlated covariates, of which 5 continuous and 5 binary, drawn from a Gaussian Copula $C_\Theta^{\text{Gauss}}(u) = \Phi_\Theta\big(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_P)\big)$, where the covariance matrix is such that $\Theta_{jk} = 0.1^{|j-k|} + 0.1\mathbb{I}(j \neq k)$. The data generating process is fully described by the following quantities:

$$\mu(\boldsymbol{x}_i) = 6 + 0.3\exp(X_{i,1}) + 1X_{i,2}^2 + 1.5|X_{i,3}| + 0.8X_{i,4} ,$$

$$\tau(\boldsymbol{x}_i) = 3 + 0.8X_{i,1}^2 ,$$

$$\pi(\boldsymbol{x}_i) = \Lambda\left(-1.5 + 0.5X_{i,1} + \frac{\nu_i}{10}\right) , \tag{5}$$

$$A_i \sim \text{Bernoulli}\big(\pi(\boldsymbol{x}_i)\big) ,$$

$$Y_i = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)A_i + \varepsilon_i , \quad \text{where} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) ,$$

where: $\Lambda(\cdot)$ is the logistic cumulative distribution function; the error's standard deviation is $\sigma^2 = 0.5$; and $\nu_i \sim$ Uniform$(0,1)$. More details on the DGP and the models employed can be found at https://github.com/albicaron/ICNN, for full reproducibility.