CauseBox: A Causal Inference Toolbox for Benchmarking Treatment Effect Estimators with Machine Learning Methods

Paras Sheth[1], Ujun Jeong[1], Ruocheng Guo[2], Huan Liu[1], K. Selçuk Candan[1]
[1]Computer Science and Engineering, Arizona State University
Arizona, USA
[2]School of Data Science, City University of Hong Kong
Hong Kong, China
{psheth5,ujeong1,huanliu,candan}@asu.edu,ruocheng.guo@cityu.edu.hk

ABSTRACT

Causal inference is a critical task in various fields such as healthcare, economics, marketing and education. Recently, there have been significant advances through the application of machine learning techniques, especially deep neural networks. Unfortunately, to-date many of the proposed methods are evaluated on different (data, software/hardware, hyperparameter) setups and consequently it is nearly impossible to compare the efficacy of the available methods or reproduce results presented in original research manuscripts. In this paper, we propose a causal inference toolbox (`CauseBox`) that addresses the aforementioned problems. At the time of publication, the toolbox includes seven state of the art causal inference methods and two benchmark datasets. By providing convenient command-line and GUI-based interfaces, the `CauseBox` toolbox helps researchers fairly compare the state of the art methods in their chosen application context against benchmark datasets. The code is made public at `github.com/paras2612/CauseBox`.

CCS CONCEPTS

• **Computing methodologies** → *Causal reasoning and diagnostics*; *Machine learning algorithms*; *Neural networks*; *Learning latent representations.*

KEYWORDS

Causal Inference, Deep Learning, Treatment Effect Estimation

ACM Reference Format:
Paras Sheth[1], Ujun Jeong[1], Ruocheng Guo[2], Huan Liu[1], K. Selçuk Candan[1]. 2021. CauseBox: A Causal Inference Toolbox for Benchmarking Treatment Effect Estimators with Machine Learning Methods. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3481974

1 INTRODUCTION

The ability to assess the causal effects of treatments is vital across many domains, including healthcare [3, 12] and digital marketing [8, 24, 32, 34], and various methods machine learning methods have been proposed for the estimation of treatment effect [11, 21, 23, 37]. Over the past few years there has been an increasing proclivity towards developing methods for treatment effect estimation. These methods are evaluated on different subsets of various benchmark datasets (for instance IHDP, Jobs, News and so on). Currently, there is no unified platform where researchers can evaluate the efficiency of various state of the art baselines and compare it with their own methods. Since different baselines use different subsets of the benchmark datasets it is difficult to identify the strongest baselines. Treatment effect estimation, has been extensively studied in statistics for decades. However, traditional estimation methods may not well handle large-scale and high-dimensional heterogeneous data. Hence, there has been an increasing attention towards utilizing neural networks and ensemble learning methods for treatment effect estimation problems [4, 7, 13, 16, 17, 25, 29–31, 36].

Observational Data vs. Randomized Control Trials (RCTs). Treatment effects are generally estimated using two different types of data – observational data and data from randomized control trials (RCTs). *Observational data* consist of recorded values of treatment assignments, the covariates, and the observed outcomes. Such data often does not provide information of the mechanism by which the treatment was assigned. Randomized control trials (RCTs) are carried out to eliminate confounder bias. Unlike observational data, in RCTs the treatments are assigned in a bias-free randomized fashion among the subjects, which in turn eliminates any unintended relations between treatment and confounders. Unfortunately, conducting RCTs is expensive and time consuming. In contrast, observational data are often readily available for treatment effect estimation – but are subject to confounding bias.

Causal Inference from Observational Data. Recently, there have been significant advances in treatment effect estimation (in the presence of unknown confounders) through the application of advanced machine learning techniques, including deep neural networks [14, 15, 18, 19, 30, 35, 36]. Yet, many of the existing methods are evaluated on different (data, software/hardware, hyperparameter) setups and consequently it is very difficult to compare the efficacy of the available methods or reproduce results presented in original research manuscripts.

CauseBox. Starting from the premise that it is critical for researchers and practitioners to have access to a platform where they can compare the state of the art methods (or their proposed methods) on standard benchmark datasets, in this paper, we introduce a toolbox (`CauseBox`) for benchmarking treatment effect estimators that are based on machine learning methods.

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3481974

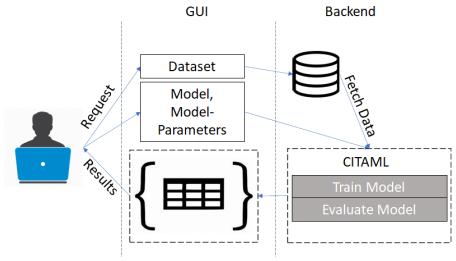

Figure 1: CauseBox toolbox overview

CauseBox currently provides command-line and GUI-based access to seven state of the art methods (including five that are deep learning based) and enables the users to run comparisons under different setups (including against their own algorithms).

2 RELATED WORK

Given the critical need for research on treatment effect estimation, there already are several attempts to develop toolboxes. In this section, we provide an overview of the existing efforts and highlight how CauseBox differs from them:

CausalML [6] implements an array of causal inference methods. It provides a standard interface for users to estimate Individual Treatment Effect (ITE) based on experimental or observational data, without strong assumptions on the model form. It currently supports Tree-based algorithms, Meta-learner algorithms (e.g., S-learner), and Instrumental Variable (IV) algorithms (e.g., 2-Stage Lease Squares). Covered evaluation metrics include Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

EconML [28] estimates heterogeneous treatment effects from observational data. The supported methods are at the intersection of econometrics and machine learning, including double machine learning, orthogonal random forests, meta-learners, doubly robust learners, orthogonal IV, and deep IV.

DoWhy [2] provides a unified interface for causal inference methods under the two fundamental frameworks – graphical models and potential outcomes. DoWhy helps in creating a causal graphical model for any scenario to describe the causal assumptions and then uses *do*-calculus to identify certain causal effects.

JustCause. The goal of JustCause [1] is to provide a fair way to benchmark new methods against several baselines and the state-of-the-art methods in causal effect estimation. The supported algorithms are doubly robust estimation, inverse propensity weighting, S-learner and T-learner. Evaluation metrics implemented include precision in estimation of heterogeneous effect (PEHE) score, MAE, EnoRMSE score, and Bias.

CauseBox differs from those mentioned above as it includes more recent, especially deep learning based, methods for treatment effect estimation. Moreover, CauseBox facilitates the evaluation of algorithms using standard metrics for treatment effect estimation, including PEHE [19] and policy risk [30].

3 CAUSEBOX SYSTEM OVERVIEW

CauseBox toolbox provides users a platform to evaluate the state of the art machine learning based methods for treatment effect estimation and compare consistent benchmark results against their own methods using benchmark datasets. As outlined in Fig. 1, the CauseBox toolbox consists of two major components: a frontend (command-line as well as graphical user interface), which takes user inputs and displays generated results to the user, and a backend, which evaluates the methods on datasets specified by the user using selected hyperparameters.

3.1 Frontend: User Interfaces

The frontend includes command-line based and a GUI-based user interfaces. The user specifies the models they aim to evaluate, along with the dataset and the necessary model hyperparameters. A screenshot of the user interface is provided in Fig. 2. Once the backend finishes training/testing the specified model with the provided data and hyperparameters, the frontend displays the evaluation results in a convenient format for comparative study.

3.2 Backend: Computational Platform

The backend, primarily built using TensorFlow and PyTorch libraries in python, along with R scripts for BART [19] and causal forest [33], consists of the implementation of various models and provides results back to frontend. After getting the parameters as input from the frontend, the backend fetches the dataset from the path provided and then evaluates the specified model on the fetched data. In the evaluation stage, the backend loads the test data and evaluates the performance of the model and passes the results to the frontend to be conveniently displayed to the user.

3.3 Benchmark Algorithms

As benchmarks, CauseBox includes seven cutting-edge methods for treatment effect estimation. These methods utilize neural network architecture for learning advanced representations which facilitate in better treatment effect estimation, making them the state of the art baselines for treatment effect estimation.

3.3.1 Counterfactual Regression Network. The central idea of (CFR-Net) [30] emphasizes the fact that the error for estimating the individual treatment effects is bounded by the sum of the standard generalization error of the representations and the distance between the treated and control distributions induced by the representations. The standard generalization error of representations is measured by the expected loss of the model and the distance between the treated and control distributions, is learnt by the help of an integral probabilistic metric (IPM) measuring similarity between probability distributions. For a more detailed reading, readers can refer to [30].

3.3.2 Causal Effect Variational Auto Encoder. The central idea emphasizes of (CEVAE) [25] on learning proxies to account for hidden confounders. To estimate these proxies the authors extend variational autoencoders to learn a latent variable causal model. Specifically, they try to infer the complex non-linear relationships between covariates (x), the treatment (t), the outcomes(y), and the confounders (z) and approximately recover the joint distribution $p(z; x; t, y)$. For a detailed reading readers could refer to [25].

3.3.3 Perfect Match. To simplify treatment effect estimation in situations where multiple treatments exist authors in [29], propose to utilize nearest neighbors based on propensity scores to balance

the covariates and this can help eliminate treatment assignment bias to lead to better effect estimation. The readers can refer to [29].

3.3.4 Disentangled Representation Network. The central idea of (DRNet) [17] is that if we can identify the underlying factors of a dataset, then we can leverage this knowledge to better estimate treatment effects. The authors propose to categorize the underlying factors into three sets, namely Γ, Δ, and Υ, where Γ accounts for the treatments only, Δ accounts for confounding variables that affect both treatments and outcomes, and υ accounts for the outcomes. They further show that by decomposing the dataset into these three components helps better estimate the treatment effects. For further details, readers can refer to [17].

3.3.5 Similarity Preserved Treatment Effect Estimation. The central idea of (SITE) [36] is to preserve similarity information while balancing treated and control group distributions. The authors argue that similar units should have similar outcomes, and to better estimate treatment effects, one should preserve the local similarity information. SITE model maps mini-batches of units from covariate space to latent space with the help of neural networks. The similarity information is preserved using the Position-Dependent Deep Metric [20]; to balance the data distributions Middle Point Distance Minimization is used. For further details, readers can refer to [36].

3.3.6 Causal Forests. Causal Forests are an extension of Random Forests [5] – each causal forest is an ensemble of causal trees and the forest outputs its prediction as the mean of the predictions for each tree. Given outcomes and treatments (y_i, t_i), each tree partitions feature space recursively such that it is partitioned into L leaves and, for each $i \epsilon L$, the data acts as if it comes from a randomized experiment. First, a classification tree is trained where the outcome is treatment assignment given (x, t) pairs for each data point. Given treatment t, the treatment effect is estimated on the leaf that contains covariate x. For details, readers can refer [33].

3.3.7 Bayesian Additive Regression Trees. Bayesian Additive Regression Tree (BART) [9], an ensemble/sum-of-trees based inference model, is built upon regression trees that partition the data into non-overlapping subsets such that the variance in response variable within each subset is minimized. To avoid overfitting of the data BART makes use of Bayesian priors. [19] applies BART to causal inference settings due to its ability to account for certain nonlinearities and interactions.

3.4 Benchmark Datasets

In addition to algorithms, CauseBox also provides datasets to be used as benchmarks. At the time of the writing, the following benchmark datasets are available to the users:

IHDP. The IHDP dataset [19] contains data from a randomised study on the impacts of specialist visits on the cognitive development of children. The study included 747 children with 25 covariates describing the properties of the children and their mothers. Children that did not receive specialist visits were part of a control group. IHDP is a semi-synthetic dataset where the simulated outcomes obtained from the NPCI package [10] are drawn from the noiseless outcome distributions to compute the ground truth individual treatment effect. In CauseBox, we use the version of the IHDP dataset that averages over 100 realizations with the 70/30 train/validation split [36].

Jobs. In the Jobs dataset [22], the treatment is job training, the measured outcomes are income and employment status after training. This dataset combines a randomized study with observational data to form a larger dataset. The presence of the randomized subgroup gives a way to estimate the ground truth causal effect. The data have 8 covariates, including age and education, as well as previous earnings. CauseBox provides a version of the dataset that averages over 10 realizations with the 56/24/20 train/validation/test split.

3.5 Benchmark Metrics

In addition to more conventional machine learning metrics commonly used in the literature, such as MAE and RMSE (see Section 2), CauseBox includes several metrics, Precision in Estimation of Heterogeneous Effects (PEHE), and Policy Risk, that are especially suitable for treatment effect estimation scenarios. Below we describe PEHE and policy risk metrics:

3.5.1 Precision in Estimation of Heterogeneous Effects (PEHE). PEHE is defined as:

$$\epsilon_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbb{E}_{y_1(i)\sim\mathcal{N}(\mu_1(i),1)}\,[y_1(i)] - \mathbb{E}_{y_0(i)\sim\mathcal{N}(\mu_0(i),1)}\,[y_0(i)] - [\hat{y}_1(i) - \hat{y}_0(i)])^2. \tag{1}$$

PEHE measures the accuracy of a causal inference model to estimate the individual treatment effect when ground truth is available (such as when using the IHDP data set). Here $y_1(i) - y_0(i)$ is the true ITE, drawn from the noiseless outcome distributions with mean μ_1 and μ_0, respectively, whereas $\hat{y}_1(i) - \hat{y}_0(i)$ is the estimated ITE.

3.5.2 Policy Risk. In some datasets (e.g., the Jobs dataset), the ground truth individual treatment effects are not available. To address the evaluation challenge in such cases, CauseBox provides another evaluation metric – policy risk, which is defined as the difference between the optimal average outcome and the average outcome when treatments are assigned according to the policy implied by a certain ITE estimator. More specifically, for a model f, if the policy to be treated is defined as, $\pi_{f(x)} = 1$ if $f(x,1) - f(x,0) > \lambda$ and $\pi_{f(x)} = 0$ otherwise, then policy risk is defined as,

$$R_{\text{Pol}}(\pi_f) = 1 - (\mathbb{E}[y_1|\pi_f(x) = 1].p(\pi_f = 1) + \mathbb{E}[y_0|\pi_f(x) = 0] \cdot p(\pi_f = 0)).$$

where, y_1 is the potential outcome under treatment, $\mathbb{E}[y_1|\pi_f(x) = 1]$ is the expected value of y_1 when $\pi_f(x) = 1$ and $p(\pi_f = 1)$ measures the probability of $\pi_f = 1$. The policy risk measures the expected loss if the treatment is taken according to the estimated ITE. In Fig. 3, we provide sample results of CauseBox for the available data sets, using the two metrics outlined above.

4 DEMONSTRATION SCENARIO

In this section, we provide a brief demonstration scenario outlining how the users can use the toolbox to train and evaluate different methods across benchmark datasets.

We note that the user can access CauseBox through commandline or GUI based interfaces. In the demonstration, we focus on

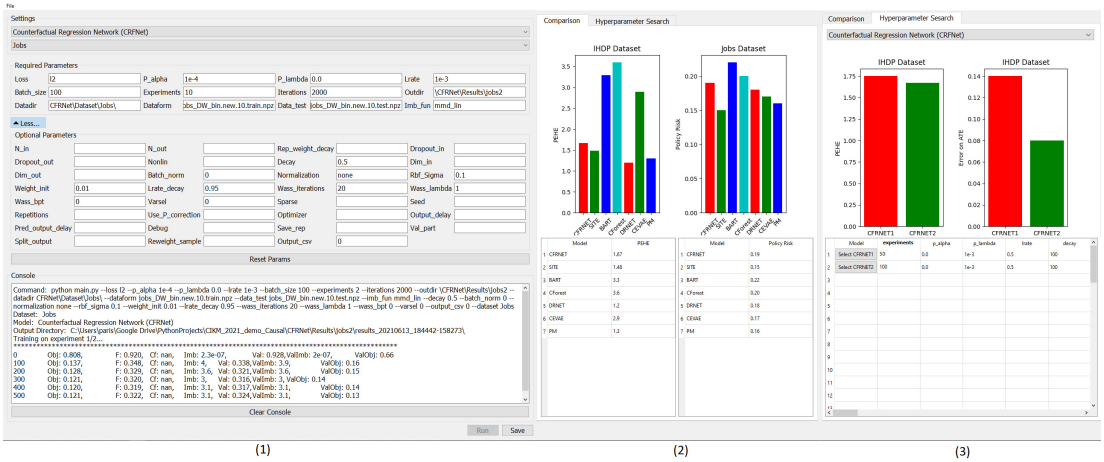

Figure 2: Components of the CauseBox GUI: Section (1) helps the user provide the evalaution context; Section (2) presents to the user an overview of model performances across benchmark datasets, and Section (3) enables the user explore the performance of a specified model across different hyperparameter settings

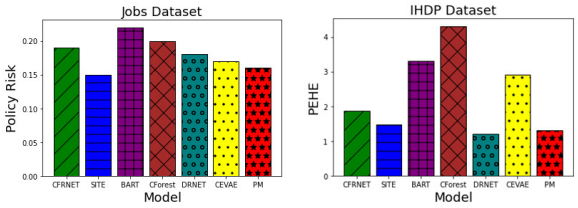

Figure 3: Results for the Jobs (left) and IHDP (right) datasets across different models: both for PEHE and policy risk, the lower the value is, the better the performance

the GUI, shown in Fig. 2, designed to help reduce the users' effort for comparing and analyzing various treatment effect estimation methods. Once the user accesses the GUI, she first selects a model along with a dataset to train and evaluate and provides values for the relevant hyperparameters (the user can also upload parameter text files for ease of use). In the GUI, the relevant parameters are divided into required and optional parameters. Once all the required parameters values are provides, the user can train and test the model against the selected benchmark data by clicking a single button. The results are made available to the user in the **comparison** and **hyperparameter search** tabs of the GUI.

The **comparison** tab provides the user a comprehensive overview of the experiments she has done so far and allows her to see (using bar charts and detailed tables) which model outperforms the rest on the chosen dataset(s) for different user selected casual inference metrics, such as PEHE and policy risk. The benchmark comparison against competitors can be done against CauseBox provided default values or can be carried out against user provided hyperparameter settings.

The user then interacts with the second, **hyperparameter search**, tab of interface to perform hyperparameter optimization for users' specified model. In this tab, the user can vary the hyperparameters of the model and observe the impact of her choices on the model performance for the selected data set. In Fig. 2, this tab visualizes results for the PEHE and error on ATE metrics. PEHE is measured as shown in 1 whereas error on ATE is the absolute error between the true average treatment effect and the predicted average treatment effect. The hyperparameter search tab also enables the user to perform *automated* search for optimal parameter settings for the given causal model under a user selected evaluation metric. This is achieved either through basic *grid search* or through *black-box Bayesian/stochastic optimization* [26, 27] (with user provided optimization settings). The interface also enables the user to investigate the sensitivity of the model against various hyperparamaters and visualize the complex interactions among the various hyperparameters.

5 CONCLUSIONS AND FUTURE WORK

In this paper we present the Causal Inference Toolbox (CauseBox). The toolbox implements several the state of the art causal inference methods and provides two benchmark datasets, IHDP and Jobs, to facilitate comparison and analysis of new methods against competitive benchmarks. We believe this toolbox can help researchers in comparing their methods against state of the art methods in an easy manner. The toolbox also provides command-line and GUI-based access to benchmarks. In addition to enabling direct hyperparameter configuration, CauseBox also provides an automated hyperparameter search function that relies on optimization techniques.

In future we plan to extend CauseBox's capabilities by allowing it to evaluate any customized algorithms. CauseBox currently supports two benchmark datasets IHDP and Jobs. We also plan to add more benchmark datasets (for instance, News) in the future.

6 ACKNOWLEDGEMENTS

This material is based upon work supported by, or in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF2110030 and W911NF2020124 as well as by the National Science Foundation (NSF) grant 1909555.

REFERENCES

[1] [n.d.]. JustCause. https://github.com/inovex/justcause

[2] 2019. DoWhy: A Python package for causal inference. https://github.com/microsoft/dowhy.

[3] Ahmed M Alaa and Mihaela van der Schaar. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *arXiv preprint arXiv:1704.02801* (2017).

[4] Onur Atan, James Jordon, and Mihaela van der Schaar. 2018. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[5] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[6] Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. 2020. CausalML: Python Package for Causal Machine Learning. arXiv:2002.11631 [cs.CY]

[7] Xiaohong Chen, Ying Liu, Shujie Ma, and Zheng Zhang. 2020. Efficient Estimation of General Treatment Effects using Neural Networks with A Diverging Number of Confounders. *arXiv preprint arXiv:2009.07055* (2020).

[8] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. 2013. Inference on counterfactual distributions. *Econometrica* 81, 6 (2013), 2205–2268.

[9] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2006. Bayesian Ensemble Learning. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*. 265–272.

[10] Vincent Dorie. 2016. NPCI: Non-parametrics for causal inference. *URL: https://github. com/vdorie/npci* (2016).

[11] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601* (2011).

[12] Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. 2013. Causal inference in public health. *Annual review of public health* 34 (2013), 61–75.

[13] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–37.

[14] Ruocheng Guo, Jundong Li, Yichuan Li, K Selçuk Candan, Adrienne Raglin, and Huan Liu. 2020. IGNITE: A minimax game toward learning individual treatment effects from networked observational data. In *29th International Joint Conference on Artificial Intelligence, IJCAI 2020*. International Joint Conferences on Artificial Intelligence, 4534–4540.

[15] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Counterfactual evaluation of treatment assignment functions with networked observational data. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 271–279.

[16] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Learning individual causal effects from networked observational data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 232–240.

[17] Negar Hassanpour and Russell Greiner. 2019. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

[18] Jingyu He, Saar Yalov, and P Richard Hahn. 2019. XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 1130–1138.

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

[20] Chen Huang, Chen Change Loy, and Xiaoou Tang. 2016. Local similarity-aware deep feature embedding. *arXiv preprint arXiv:1610.08904* (2016).

[21] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856* (2017).

[22] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* (1986), 604–620.

[23] Sheng Li and Yun Fu. 2017. Matching on balanced nonlinear representations for treatment effects estimation. In *NIPS*.

[24] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. 2016. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns.. In *IJCAI*. 3768–3774.

[25] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821* (2017).

[26] Logan Mathesen, Kaushik Keezhnagar Chandrasekar, Xinsheng Li, Giulia Pedrielli, and K. Selçuk Candan. 2019. Subspace Communication Driven Search for High Dimensional Optimization. In *2019 Winter Simulation Conference, WSC 2019, National Harbor, MD, USA, December 8-11, 2019*. 3528–3539.

[27] Logan Mathesen, Giulia Pedrielli, Szu Hui Ng, and Zelda B. Zabinsky. 2021. Stochastic optimization with adaptive restart: a framework for integrated local and global learning. *J. Glob. Optim.* 79, 1 (2021), 87–110.

[28] Microsoft Research. 2019. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML. Version 0.x.

[29] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. 2018. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656* (2018).

[30] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.

[31] Claudia Shi, David M Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120* (2019).

[32] Hal R Varian. 2016. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7310–7315.

[33] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[34] Pengyuan Wang, Wei Sun, Dawei Yin, Jian Yang, and Yi Chang. 2015. Robust tree-based causal inference for complex ad effectiveness analysis. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 67–76.

[35] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2020. A survey on causal inference. *arXiv preprint arXiv:2002.02770* (2020).

[36] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems* 31 (2018).

[37] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. 2015. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.