MIMICKING RANDOMIZED CONTROLLED TRIALS TO LEARN END-TO-END PATIENT REPRESENTATIONS THROUGH SELF-SUPERVISED COVARIATE BALANCING FOR CAUSAL TREATMENT EFFECT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A causal effect can be defined as a comparison of outcomes that result from two or more alternative actions, with only one of the action-outcome pairs actually being observed. The gold standard for causal effect measurements is Randomized Controlled Trials (RCTs), in which a target population is explicitly defined and each study sample is randomly assigned to either the treatment or control cohorts. The great potential to derive actionable insights from causal relationships has led to a growing body of machine-learning research applying causal effect estimators to Observational Data (OD) in the fields of healthcare, education, and economics. The primary difference between causal effect studies utilizing OD and RCTs is that for OD the study occurs after the treatment, and therefore we don't have control over the treatment assignment mechanism. This can lead to massive differences in covariate distributions between control and treatment samples, making a comparison of causal effects confounded and unreliable. Classical approaches have sought to solve this problem piecemeal, first by predicting treatment assignment and then treatment effect separately. Recent work extended part of these approaches to a new family of representation-learning based algorithms, showing that the upper bound of the expected treatment effect estimation error is determined by two factors: the outcome generalization-error of the representation and the distance between the treated and control distributions induced by the representation. Here we argue that to achieve minimal dissimilarity in learning such distributions, as RCTs are designed to do, a specific auto-balancing self-supervised objective should be used. Experiments on real and benchmark datasets revealed that our approach consistently produced less biased estimates than previously published state-of-art methods. We demonstrate that our reduction in error can be directly attributed to the ability to learn representations that explicitly reduce such dissimilarity. Thus, by learning representations that induce distributions analogous to RCTs, we provide empirical evidence to support the error bound dissimilarity hypothesis as well as providing a new state-of-the-art model for causal effect estimation.

1 INTRODUCTION

Causal effect estimation of a binary exposure on a continuous outcome from observational data is a fundamental problem faced by many researchers and has a broad range of applications across diverse disciplines. For example, in social economy, policy makers need to determine who would benefit most from subsidized job training. In precision medicine, doctors need to decide which medication will cause better outcomes for a specific patient affected by a disease, taking into account relevant information such as age and pre-existing chronic conditions. The gold standard for estimating causal relationships have been RCTs, which can be thought of as having three distinct stages: selection criteria to ensure that all samples are effectively equivalent prior the study starts so that differences in outcomes can be attributed to differences in treatments, randomization of each sample to a treatment arm, and outcome comparison between the treatment arms. In particular, by controlling the assignment mechanism, RCTs reduce the bias of treatment effect estimation due to factors that affect both

treatment and outcome (*confounders*). However, RCTs are not always feasible due to logistical, ethical, or financial considerations. Moreover, being based on restricted populations following strict protocols that frequently do not match daily standards of care, the results from RCTs do not always generalize to new patients in the real world (Munk NE, 2020; Klonoff, 2020). In the past decade, the viability of OD to extend to RCTs to infer causal relationships has been explored due to the increasingly available patient data captured in Electronic Health Records (EHRs), the remarkable advances of machine learning techniques, and considerably reduced cost of such studies.

Classical works in causal inference addressed the problem of estimating average treatment effects (ATE) from OD with covariate adjustment, also known as back-door adjustment (Belloni et al., 2013; Athey et al., 2018; Chernozhukov et al., 2017), or weighting methods (Austin, 2011), where an estimate of the probability of treatment, conditioned on covariates (*propensity score*), is used to reweight the units in the OD to make the treated and control populations more comparable. Targeted Learning (van der Laan & Rose, 2018) adjusts the estimation of an initial statistical model in a step targeted toward making an optimal bias-variance tradeoff of the causal effect. All those approaches are not end-to-end, meaning that models are trained separately without sharing a common representation. For example, in targeted learning the initial statistical model is trained separately from the one that adjusts its estimation. Shalit et al. (2017b) extended part of those approaches to a new family of representation-learning based algorithms (Bengio et al., 2013), and demonstrated that the expected error in learning individual treatment effect is upper bounded by the error of learning factual and counterfactual outcomes plus a term depending on the dissimilarity of the treated and untreated distributions induced by the learned representation.

In this work, we proposed an end-to-end model that, in addition to factual losses, uses a self-supervised auto-balancing objective specifically designed to minimize the dissimilarity of the learned representations for treated and untreated cohorts. We call this method BCAUSS (**B**alancing **C**ovariates **A**utomatically **U**sing **S**elf-**S**upervision). Adopting two widely used datasets in the casual inference community, such as IHDP and Jobs (see Section 4.1), we found that BCAUSS produced less biased estimates than previously published state-of-art methods. In particular, we compared BCAUSS to Dragonnet (Shi et al., 2019), the current state of the art on IHDP. We show here that BCAUSS produced less biased estimates than Dragonnet because of its ability to learn less dissimilar treated and untreated distributions, consistently to how they should be in RCTs.

2 Related work

Classical causal modeling typically involves the concept of *balancing score* (Rosenbaum & Rubin, 1983). Back-door adjustment methods (Belloni et al., 2013; Athey et al., 2018; Chernozhukov et al., 2017) and weighting methods (Austin, 2011) adopt a special balancing score, i.e. the *propensity score*, to reweight the units in OD to make the treated and control populations more comparable. Targeted Learning (van der Laan & Rose, 2018) adjusts the estimation of an initial statistical model with a second model, making an optimal bias-variance tradeoff of the causal effect. Treatment effect estimation has also been approached by designing treatment effect specific splitting criterions for recursive partitioning (Su et al., 2009; Athey & Imbens, 2016; Zhang et al., 2017), adopting ensemble algorithms and meta algorithms (Künzel et al., 2019; Wager & Athey, 2017). Other approaches like Propensity Dropout (Alaa et al., 2017) and Perfect Matching (Schwab et al., 2019) combines (pre-trained) propensity scores with neural networks.

Shalit et al. (2017b) introduced a new family of representation-learning based algorithms (Bengio et al., 2013), including TARNET. Shi et al. (2019) introduced Dragonnet extending TARNET in different ways. First, changing the network architecture so that the same learned representation used for learning factual and counterfactual outcomes is shared with the the propensity score estimator. Second, adopting a targeted regularization objective to achieve optimal asymptotical properties. BCAUSS adopts the same network architecture of Dragonnet with a specific auto-balancing self-supervised objective to achieve minimal dissimilarity in learning treated and untreated distributions. On the other hand, Belthangady & Norgeot (2021) adopted our objective on a single-task network architecture for propensity score estimation only (instead of causal treatment effect estimation). Further, such objective was adopted in conjunction with binary cross-entropy, which we show here is sub-optimal. Other works applied deep generative models to casual inference. For example, CEVEA (Louizos et al., 2017c), GANITE (Yoon et al., 2018), and CMPGP (Alaa & van der Schaar,

2017) use VAEs, GANs, and multi-task gaussian processes, respectively, to estimate treatment effects. Zhang et al. (2021) proposed a variational inference approach to infer latent factors from the observed variables. Our work does not aim to learn the joint probability distribution of covariates and outcome, but the conditional probability distribution of the outcome given covariates.

3 APPROACH

We assume a population P, where the i-th individual has covariates $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^{\mathbf{d}}$ and is subject to intervention $t_i \in \{0, 1\}$, also known as treatment, where $t_i = 0$ indicates the individual receives no treatment, while $t_i = 1$ indicates the individual receives the treatment. We use $Y_1^{(i)} \in \mathbb{R}$ to denote the potential outcome of the individual, if treated, and $Y_0^{(i)} \in \mathbb{R}$ to denote the potential outcome, if not treated. We assume $\left(Y^{(i)}, t_i, \mathbf{x}_i\right)$ are independent and identically distributed and we denote with P_n our sample data of size n, i.e. $P_n = \left\{Y^{(i)}, t_i, \mathbf{x}_i\right\}_{i=1}^n$, while we denote with I_n the subset of the input variables only, i.e. $I_n = \{t_i, \mathbf{x}_i\}_{i=1}^n$. Moreover, we assume that the following three fundamental assumptions for treatment effect estimations (Rosenbaum & Rubin, 1983) are satisfied.

Assumption 1. (SUTVA) *The Stable Unit Treatment Value Assumption requires that the potential outcomes for one individual are unaffected by the treatment of others.*

Assumption 2. (Unconfoundedness) *The distribution of treatment is conditional independent of the potential outcomes, given covariates, i.e.* $(\forall i) \left(t_i \perp\!\!\!\perp \left(Y_0^{(i)}, Y_1^{(i)}\right) | \mathbf{x}_i\right)$.

Assumption 3. (Positivity) *Every individual has a non-zero probability to receive either treatment or control, given covariates, i.e.* $(\forall i) \left(0 < P\left(t_i = 1 | \mathbf{x}_i\right) < 1\right)$.

When the unconfoundedness and the positivity assumption hold, then the treatment assignment is considered to be *strongly ignorable* (Rosenbaum & Rubin, 1983), implying that among people with the same covariates, we can think of treatment as being randomly assigned. In the related causal graph this means that any backdoor path from t to Y is blocked, among people with the same covariates. Hence, the average treatment effect (ATE) $\psi \in \mathbb{R}$ is defined as:

$$\psi = \mathbb{E}\left[Y_1 - Y_0\right] = \mathbb{E}\left[Y_1 | do\left(t = 1\right)\right] - \mathbb{E}\left[Y_0 | do\left(t = 0\right)\right]$$

, where $do\left(t = 1\right)$ denotes a manipulation on t by removing all its incoming edges on the related causal graph and setting $t = 1$, so that the effect of interest is causal (Pearl, 2009). To estimate ψ, the concept of *balancing score* is crucial in the sense of Rosenbaum & Rubin (1983), i.e. a function $b\left(\mathbf{x}\right)$ of the observed covariates $\mathbf{x}$ such that the conditional distribution of $\mathbf{x}$ given $b\left(\mathbf{x}\right)$ is the same for treated and control units; that is, $\mathbf{x} \perp\!\!\!\perp t | b\left(\mathbf{x}\right)$. Indeed, in Rosenbaum & Rubin (1983) the following theorem is proved.

Theorem 1. *Suppose treatment assignment is strongly ignorable and* $b\left(\mathbf{x}\right)$ *is a balancing score. Then the expected difference in observed responses to the two treatments at* $b\left(\mathbf{x}\right)$ *is equal to the average treatment effect at* $b\left(\mathbf{x}\right)$*, that is*

$$\psi = \mathbb{E}\left[\mathbb{E}\left[Y_t | b(\mathbf{x}), t = 1\right] - \mathbb{E}\left[Y_t | b\left(\mathbf{x}\right), t = 0\right]\right] = \mathbb{E}\left[Y_1 - Y_0 | b(\mathbf{x})\right].$$

3.1 LEARNING REPRESENTATIONS FOR CAUSAL TREATMENT EFFECT ESTIMATION

Following Shalit et al. (2017b), we will employ the following results and notations.

Definition 2. Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a bijective representation function, where $\mathcal{X}$ is the space of covariates and $\mathcal{R}$ is the representation space. Let Ψ be the inverse of Φ , i.e. for all $\mathbf{x} \in \mathcal{X}$ we have $\Psi\left(\Phi\left(\mathbf{x}\right)\right) = \mathbf{x}$. Let $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y} \subseteq \mathbb{R}$ the hypothesis.

Definition 3. Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ a loss function so that the expected loss for the unit and treatment pair $(\mathbf{x}, t)$ is $l_{h,\Phi}\left(\mathbf{x}, t\right) = \int_{\mathcal{Y}} L\left(y\left(t\right), h\left(\Phi\left(\mathbf{x}\right), t\right)\right) p\left(Y_t | \mathbf{x}\right) dY_t)$, the expected factual and counterfactual losses are:

$$\epsilon_F\left(h, \Phi\right) = \int_{\mathcal{X} \times \{0,1\}} l_{h,\Phi}\left(\mathbf{x}, t\right) p\left(x, t\right) dxdt$$

$$\epsilon_{CF}\left(h, \Phi\right) = \int_{\mathcal{X} \times \{0,1\}} l_{h,\Phi}\left(\mathbf{x}, t\right) p\left(x, 1 - t\right) dxdt$$

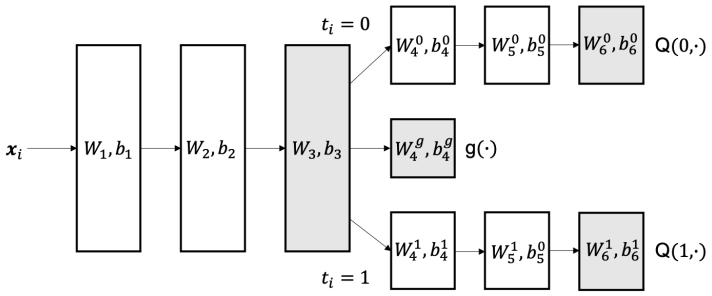


Figure 1: BCAUSS and Dragonnet network architecture.

, from which the expected factual treated and control losses are $\epsilon_F^{t=1}(h, \Phi) = \int_{\mathcal{X}} l_{h,\Phi}(\mathbf{x}, t)\, p^{t=1}(x)\, dx$ and $\epsilon_F^{t=0}(h, \Phi) = \int_{\mathcal{X}} l_{h,\Phi}(\mathbf{x}, t)\, p^{t=0}(x)\, dx$, where $p^{t=1} := p(x\,|t=1)$ and $p^{t=0} := p(x\,|t=0)$, respectively.

Definition 4. Given that the individual-level treatment effect (ITE) for unit $\mathbf{x}$ is $\tau(\mathbf{x}) := \mathbb{E}[Y_1 - Y_0\,|\mathbf{x})]$ and the treatment effect estimate of the hypothesis Q for unit $\mathbf{x}$ is $\tau_Q(\mathbf{x}) := Q(\mathbf{x}, 1) - Q(\mathbf{x}, 0)$, the expected Precision in Estimation of Heterogeneous Effect (PEHE) loss is

$$\epsilon_{PEHE}(Q) := \int_{\mathcal{X}} \left[\tau_Q(\mathbf{x}) - \tau(\mathbf{x})\right]^2 p(x)\, dx.$$

Also, the expected variance of Y_t with respect to a distribution $p(\mathbf{x}, t)$ is $\sigma^2_{Y_t}(p(\mathbf{x}, t)) = \int_{\mathcal{X}\times\mathcal{Y}} [Y_t - \mathbb{E}(Y_t\,|\mathbf{x}))]\, p(x, t)\, dY_t dx$, from which we define $\sigma^2_{Y_t} = \min\left\{\sigma^2_{Y_t}(p(\mathbf{x}, t)), \sigma^2_{Y_t}(p(\mathbf{x}, 1-t))\right\}$ and $\sigma^2_Y = \min\left\{\sigma^2_{Y_0}, \sigma^2_{Y_1}\right\}$.

For two probability density functions p, q defined over $S \subseteq \mathbb{R}^d$ and for a function family G of functions $g : \mathcal{S} \to \mathbb{R}$, we have that

$$IPM_G(p, q) := \sup_{g\in G} \left| \int_{\mathcal{S}} g(s)\, [p(s) - q(a)]\, ds \right|.$$

Integral probability metrics are symmetric, obey the triangle inequality, $IPM_G(p, p) = 0$ and, for rich enough function families G, we have that $IPM_G(p, q) = 0 \Rightarrow p = q$. Hence, IPM is a metric over the corresponding set of probabilities. Examples of function families are the 1-Lipschitz functions (Sriperumbudur, 2012) and the unit-ball of functions in a universal reproducing Hilbert kernel space (Gretton et al., 2012). In Shalit et al. (2017b) the following theorem is proved.

Theorem 5. *Under the above definitions and assuming the loss L is the squared loss, assuming there exists a constant $B_\Phi > 0$, such that for fixed $t \in \{0, 1\}$, $\frac{1}{B_\Phi} l_{h,\Phi}(\Psi(\mathbf{r}), t) \in G$, we have*

$$\epsilon_{PEHE}(h, \Phi) \leq 2\left(\epsilon_F^{t=0}(h, \Phi) + \epsilon_F^{t=1}(h, \Phi) + B_\Phi IPM_G\left(p_\Phi^{t=0}, p_\Phi^{t=1}\right) - 2\sigma^2_Y\right).$$

Theorem 5 states that the expected error in learning ITEs is upper bounded by the error of learning Y_0 and Y_1 , plus the IPM term, which depends on the dissimilarity of the learned treated and untreated distributions induced by the representation. The minimal upper bound for a model with given factual treated and untreated losses, is the one obtained when the IPM term is 0, reducing the causal treatment effect estimation problem to a standard regression problem.

3.2 BCAUSS

Fig. 1 depicts the three-task network architecture of BCAUSS, which is the same as Dragonnet (Shi et al., 2019). For $j \in \{0, 1, 2, 3\}$ and $i \in \{1, 2, \ldots, n\}$, assuming $\mathbf{a}_{0,i} = \mathbf{x}_i$, we have $\mathbf{z}_{i,j} = \mathbf{W}_j \mathbf{a}_{i,j-1} + \mathbf{b}_j$ and $\mathbf{a}_{i,j} = ReLU(\mathbf{z}_{i,j})$, where we use $ReLU(\cdot)$ to denote the ReLU activation function (Nair & Hinton, 2010). Hence, $g(\mathbf{x}_i) = \sigma(\mathbf{W}_4^g \mathbf{a}_{i,3} + \mathbf{b}_4^g)$, where $\sigma(\cdot)$ is the sigmoid function. For $j \in \{4, 5\}$, we have $\mathbf{z}_{i,j}^{t_i} = \mathbf{W}_\mathbf{j}^{\mathbf{t_i}} \mathbf{a}_{i,j-1} + \mathbf{b}_\mathbf{j}^{\mathbf{t_i}}$ and $\mathbf{a}_{i,j}^{t_i} = ReLU\left(\mathbf{z}_{i,j}^{t_i}\right)$. Hence, $Q(t_i, \mathbf{x}_i) = \mathbf{W}_\mathbf{6}^{\mathbf{t_i}} \mathbf{a}_{i,5} + \mathbf{b}_\mathbf{6}^{\mathbf{t_i}}$. We adopt mean squared error for factual loss,i.e.

$$L_{REG}^{(i)}\left(\theta; y_i, t_i, \mathbf{x}_i\right) = [t_i = 0]\left(Y_0^{(i)} - Q\left(0, \mathbf{x}_i\right)\right)^2 + [t_i = 1]\left(Y_1^{(i)} - Q\left(1, \mathbf{x}_i\right)\right)^2 \tag{1}$$

but to train the balancing score estimator, instead of adopting the binary cross-entropy objective assuming as label t_i, as it happens in the whole literature including Shi et al. (2019), we adopt the following *auto-balancing self-supervised term*

$$L_{BAL}\left(\theta; I_n\right) = \frac{1}{d}\sum_{1\leq j\leq d}\left(\frac{\sum_{1\leq i\leq n}\frac{t_i}{g(\mathbf{x}_i)}x_{i,j}}{\sum_{1\leq i\leq n}\frac{t_i}{g(\mathbf{x}_i)}} - \frac{\sum_{1\leq i\leq n}\frac{1-t_i}{1-g(\mathbf{x}_i)}x_{i,j}}{\sum_{1\leq i\leq n}\frac{1-t_i}{1-g(\mathbf{x}_i)}}\right)^2. \tag{2}$$

Notice that such term does not come with a superscript, implying that to be computed the entire trainset is required. Specifically, we called such term *auto-balancing*, as we're imposing the constraint that the learned representation achieves balance. In practice this is obtained by minimizing the squared deviation between a reweighed treated and untreated cohort, i.e. by minimizing such term, $g\left(\mathbf{x}\right)$ is incentivized to act like a balancing score in the sense of Rosenbaum & Rubin (1983). Also, we find such term *self-supervised* as it depends only on the decision of treatment t_i and covariates $\mathbf{x}_i$, which are input variables, and it does not depend on potential outcomes $Y_0^{(i)}, Y_1^{(i)}$. Hence, the optimization problem can be formulated as:

$$\hat{\theta} = \arg\min_{\theta} J\left(\theta; P_n\right), \tag{3}$$

$$J\left(\theta; P_n\right) = \lambda_{BAL} L_{BAL}\left(\theta; I_n\right) + \frac{1}{n}\sum_{1\leq i\leq n} L_{REG}^{(i)}\left(\theta; y_i, t_i, \mathbf{x}_i\right), \tag{4}$$

where λ_{BAL} controls the relative importance of the two objective terms.

3.3 Comparator: Dragonnet

Dragonnet (Shi et al., 2019) is based on the same network architecture depicted in Fig. 1 but, while mean squared error is adopted for factual loss (equation 1), to train the propensity score estimator the binary cross-entropy objective is adopted, i.e.

$$L_{BCE}^{(i)}\left(\theta; t_i, \mathbf{x}_i\right) = t_i \log\left(g\left(\mathbf{x_i}\right)\right) + \left(1 - t_i\right)\log\left(1 - g\left(\mathbf{x_i}\right)\right). \tag{5}$$

Additionally, to achieve asymptotically robustness and efficiency, a further targeted regularization term is used, i.e.

$$L_{T-REG}^{(i)}\left(\theta; y_i, t_i, \mathbf{x}_i\right) = \left[Y_{t_i}^{(i)} - Q\left(t_i, \mathbf{x}_i\right) + \epsilon\left(\frac{t_i}{g\left(\mathbf{x_i}\right)} + \frac{1 - t_i}{1 - g\left(\mathbf{x_i}\right)}\right)\right]^2 \tag{6}$$

, where ϵ is a further hyperparameter. Notice that BCAUSS and Dragonnet not only are based on the same architecture, but they have the same factual losses, hence the upper bound of the expected error in learning ITEs of Theorem 5 will depend only on the IPM term, i.e. on the dissimilarity of treated and untreated representations.

4 Experiments

4.1 Benchmark and real-world datasets

The IHDP benchmark dataset The Infant Health and Development Program (IHDP) is a randomized controlled study designed to evaluate the effect of home visit from specialist doctors on the cognitive test scores of premature infants. The datasets is first used for benchmarking treatment effect estimation algorithms in Hill (2011), where selection bias is induced by removing non-random subsets of the treated individuals to create an observational dataset, and the outcomes are generated using the original covariates and treatments. It contains 747 subjects and 25 variables. Following Shi et al. (2019); Shalit et al. (2017b), we use the simulated outcome implemented as setting "A" and, in order to make our results reproducible, we adopted the one available for download at `https://www.fredjo.com/`, which is composed by 1000 repetitions of the experiment. We averaged over 1000 train/validation/test splits with ratios 70/20/10.

The Jobs real-world dataset The study by LaLonde (1986) is a widely used benchmark in the causal inference community, where the treatment is job training and the outcomes are income and employment status after training. The study includes 8 covariates such as age and education, as well as previous earnings. Our goal is to predict unemployment, using the feature set of Dehejia & Wahba (2002). Following Shalit et al. (2017b), we use the LaLonde experimental sample (297 treated, 425 control) and the PSID comparison group (2490 control). We averaged over 10 train/validation/test splits with ratios 62/18/20. The dataset is available for download at `https://www.fredjo.com/`.

4.2 Evaluation criteria

ATE For evaluating the performance of ATE estimation, the ground truth can be calculated by averaging the differences of the outcomes in the treated and control groups. Then, comparing the ground truth ATE $\tilde{\psi}_{ATE}$ with the related estimate obtained from a sample of the dataset, performance can then be evaluated using the mean absolute error in ATE (Hill, 2011; Shalit et al., 2017c; Louizos et al., 2017a; Yao et al., 2018), i.e. $\epsilon_{ATE} = \left|\tilde{\psi}_{ATE} - \frac{1}{n}\sum_{1\leq i\leq n} Q\left(1, \mathbf{x}_i\right) - Q\left(0, \mathbf{x}_i\right)\right|$.

ATT We adopted this metric on Jobs. Because all the treated subjects T were part of the original randomized sample E, following Shalit et al. (2017b), we can compute the true average treatment effect on the treated (ATT): $\tilde{\psi}_{ATT} = \frac{1}{|T|}\sum_{\mathbf{i}\in T} Y^{(i)} - \frac{1}{|C\cap E|}\sum_{\mathbf{i}\in C\cap E} Y^{(i)}$, where C is the control group. Hence, $\epsilon_{ATT} = \left|\tilde{\psi}_{ATT} - \frac{1}{|T|}\sum_{\mathbf{i}\in T}\left[Q\left(1, \mathbf{x}_i\right) - Q\left(0, \mathbf{x}_i\right)\right]\right|$.

IPMs To measure the dissimilarity between treated and untreated distributions, we adopt Maximum Mean Discrepancy (Gretton et al., 2012) and Wasserstein distance (Villani, 2008; Cuturi & Doucet, 2014), which are IPMs used in Shalit et al. (2017b) (see Theorem 5). Additionally, we adopt Kolmogorov-Smirnov (KS) statistic Kolmogorov (1933); Smirnov (1948) to show statistical significance whether or not the two samples have different distribution, i.e. if p-value $> 1\%$ the difference between the two samples is not significant enough to say that they have different distribution.

4.3 Experimental details

Unless otherwise specified (e.g. in the ablation study of Section 5.1), with regard to figure 1, $\mathbf{W}_1 \in \mathbb{R}^{200\times d}$, $\mathbf{b_1} \in \mathbb{R}^{\mathbf{200}\times\mathbf{1}}$, for $i \in \{2, 3\}$, $\mathbf{W}_i \in \mathbb{R}^{200\times 200}$, $\mathbf{b_i} \in \mathbb{R}^{\mathbf{200}\times\mathbf{1}}$, while for $i \in \{4, 5\} \wedge t_i \in \{0, 1\}$, $\mathbf{W}_i^{t_i} \in \mathbb{R}^{100\times 100}$, $\mathbf{b_i^{t_i}} \in \mathbb{R}^{\mathbf{100}\times\mathbf{1}}$. Experiments adopt learning rate 1e-5, batch size equal to the train-set length, stochastic gradient descent with momentum (Robbins, 2007), $\lambda_{BAL} = 1$. The ablation study 5.1 shows the importance of these settings to achieve superior performance. Models were trained on the optimal number of epochs by adopting early stopping (Yao et al., 2007), using 22% of the train-set for cross-validation.

4.4 Results

We compared BCAUSS with Dragonnet and the other methods considered in Shi et al. (2019) on IHDP. BCAUSS had the best in-sample and out-sample performance (Table 1). The best performance was achieved adopting the self-supervised auto-balancing term alone (eq. 2) to train the balancing score estimator. Adding the Dragonnet targeted regularization term (eq. 6) with or without binary cross-entropy is sub-optimal. On the other hand, adding only the binary cross-entropy is also sub-optimal. In the supplementary Section A.2 we analyze the effect of back-propagating the gradient with respect to our self-supervised auto-balancing objective compared to binary cross-entropy (Section A.3), showing why if we adopt the former instead of the latter we have a lower upper bound of the expected error in learning ITE (Theorem 5) in case the treated and untreated covariate distributions are highly dissimilar, as unfortunately might happen for OD. Additionally, on Jobs we compared BCAUSS with GANITE, which is the current state-of-the-art on this dataset, including the other methods considered in Yoon et al. (2018). BCAUSS had the best out-sample performance (Table 1).

Method	Variation	IHDP ϵ^{tr}_{ATE}	IHDP ϵ^{te}_{ATE}
GANITE (Yoon et al., 2018)	—	0.43±.05	0.49 ±.05
CEVAEs (Louizos et al., 2017b)	—	0.34±.01	0.46±.02
AIPW	—	0.13±.00	0.29±.01
BNN (Johansson et al., 2016)	—	0.37±.03	0.42±.03
TARNET (Shalit et al., 2017b)	—	0.26±.01	0.28±.01
CFR Wass (Shalit et al., 2017a)	—	0.25±.01	0.27±.01
Dragonnet (Shi et al., 2019)	—	0.14±.01	0.20±.01
baseline (Dragonnet)	-T_REG	0.13±.00	0.21±.01
	—	0.15±.01	0.20±.01
BCAUSS	+T-_REG	0.15±.00	0.19±.01
	+BCE +T_REG	0.16±.00	0.20±.01
	+BCE	0.13±.00	0.17±.01
	—	**0.10±.00**	**0.15±.01**

Method	Jobs ϵ^{tr}_{ATT}	Jobs ϵ^{te}_{ATT}
OLS/LR-1	.01±.00	.08±.04
OLS/LR-2	.01±.01	.08±.03
BLR	.01±.01	.08±.03
k-NN	.21±.01	.13±.05
BART(Chipman et al., 2010)	.02±.00	.08±.03
Rand. Forest	.03±.01	.09±.04
Caus. Forest (Wager & Athey, 2018)	.03±.01	.09±.04
AIPW	**.00±.01**	.09±.02
BNN (Johansson et al., 2016)	.04±.01	.09±.04
TARNET (Shalit et al., 2017b)	.05±.02	.11±.04
CFR Wass (Shalit et al., 2017a)	.04±.01	.09±.03
GANITE (Yoon et al., 2018)	.01±.01	.06±.03
BCAUSS	**.**02±.00	**.05±.02**

Table 1: Results on IHDP (left) and Jobs (right). **Left**: BCAUSS is state-of-the-art on the IHDP benchmark dataset. "+BCE" and "+T_REG" mean adding the binary cross-entropy of eq. 5 and the targeted regularization term of eq. 6 to the overall objective. Similarly, "-T_REG" means removing the targeted regularization term of eq. 6 from the objective. AIPW adopts as propensity score estimator (Belthangady & Norgeot, 2021) and two linear regressors (for control and treatment respectively) with an L2 penalty. **Right**: Methods compared with BCAUSS are described in Shalit et al. (2017b); Yoon et al. (2018). Lower is better.

5 Analysis

5.1 Ablations

The models shown in Tab. 2 refer to the best models that we trained in our experiments described in Section 3.2. These models adopt the experimental settings specified in Section 4.3 except the ones explicitly indicated for each row of Table 2.

5.1.1 Effect of optimizer and batch size

The effect of the optimizer can be analyzed by comparing row 5-8 (SGD) to row 9-12 (Adam) in Table 2. Adopting SGD consistently improves ATE estimation compared to Adam by 0.05 in-sample and 0.08 out-of-sample, except for the lowest batch size where the out-of-sample difference is 0.07. The effect of batch size can be analyzed by comparing row 5 to row 6-8 (SGD) and row 9 to 10-12 (Adam) in Table 2. While for Adam we don't observe any improvement with higher batch sizes, for SGD we don't observe any improvement for batch size equal or higher than $\lfloor \frac{5}{12} \cdot n \rfloor$, but we observe a test performance degradation of 0.01 for lower values. In the supplementary A.2 we discuss in detail why working with large batch-size and SGD is beneficial.

5.1.2 Effect of activation function

The effect of activation function can be analyzed by comparing for each row in Table 2 the columns ReLU/ELU/Tanh, corresponding to in-sample and out-of-sample performance of the activation functions ReLUs, ELUs and Tanhs. We find that overall the best performance is observed with ReLUs. Specifically, in case of SGD optimizer, ReLUs consistently outperform ELUs by 0.02 on trainset and 0.01 on test-set. In turn, ELUs outperform Tanhs by 0.01 on test-set. On the other hand, in case of Adam optimizer, while in-sample performance is pretty homogeneous across the three, out-sample performance of ELUs outperform Tanhs by 0.01 on test-set. In turn, Tanhs outperform ReLUs by 0.03 on test-set.

Model	Optimizer	Batch Size	λ_{BCE}	λ_{BAL}	λ_{TAR}	ReLU		ELU		Tanh	
						ϵ_{ATE}^{tr}	ϵ_{ATE}^{te}	ϵ_{ATE}^{tr}	ϵ_{ATE}^{te}	ϵ_{ATE}^{tr}	ϵ_{ATE}^{te}
(0)	SGD	n	1	0	1	0.16±.01	0.20±.01	0.14±.01	0.18±.01	0.15±.01	0.19±.01
(1)	SGD	n	1	0	0	0.12±.00	0.16±.01	0.13±.00	0.17±.01	0.13±.00	0.18±.01
(2)	SGD	n	1	0.5	0	0.13±.00	0.16±.01	0.13±.00	0.17±.01	0.13±.00	0.18±.01
(3)	SGD	n	1	1	0	0.13±.00	0.16±.01	0.13±.01	0.17±.01	0.13±.00	0.18±.01
(4)	SGD	n	1	1.5	0	0.13±.00	0.17±.01	0.13±.01	0.17±.01	0.13±.00	0.18±.01
(5)	SGD	n	0	1	0	**0.10±.00**	**0.15±.01**	0.12±.00	0.16±.01	0.13±.00	0.17±.01
(6)	SGD	$\lfloor \frac{n}{2} \rfloor$	0	1	0	**0.10±.00**	**0.15±.01**	0.12±.00	0.16±.01	0.12±.00	0.17±.01
(7)	SGD	$\lfloor \frac{5}{12} \cdot n \rfloor$	0	1	0	**0.10±.00**	**0.15±.01**	0.12±.00	0.16±.01	0.12±.00	0.17±.01
(8)	SGD	$\lfloor \frac{n}{3} \rfloor$	0	1	0	**0.10±.00**	0.16±.01	0.12±.00	0.16±.01	0.12±.00	0.17±.01
(9)	Adam	n	0	1	0	0.15±.00	0.23±.01	0.15±.00	0.19±.01	0.15±.00	0.20±.01
(10)	Adam	$\lfloor \frac{n}{2} \rfloor$	0	1	0	0.15±.00	0.23±.01	0.15±.00	0.19±.01	0.15±.00	0.20±.01
(11)	Adam	$\lfloor \frac{5}{12} \cdot n \rfloor$	0	1	0	0.15±.00	0.23±.01	0.15±.00	0.19±.01	0.15±.00	0.20±.01
(12)	Adam	$\lfloor \frac{n}{3} \rfloor$	0	1	0	0.15±.00	0.23±.01	0.16±.00	0.20±.01	0.15±.00	0.20±.01
(13)	SGD	n	0	0.5	0	**0.10±.00**	**0.15±.01**	0.12±.00	0.16±.01	0.13±.00	0.17±.01
(14)	SGD	n	0	1.5	0	**0.10±.00**	**0.15±.01**	0.12±.00	0.16±.01	0.13±.00	0.17±.01

Table 2: Ablation results of the different variants described in Section 5.1 on IHDP dataset (n is the total number of observations on train-set). $\lambda_{BAL} = 1$ corresponds to add to the objective the self-supervised auto-balancing term of eq. 2, $\lambda_{BCE} = 1$ corresponds to add to the objective the binary cross-entropy term of eq. 5, $\lambda_{TAR} = 1$ corresponds to add to the objective the targeted regularization term of eq. 6. Row 5 with ReLU activation function corresponds to BCAUSS. Row 0 with ELU activation function corresponds to Dragonnet. **Bold** indicates the best performance overall.

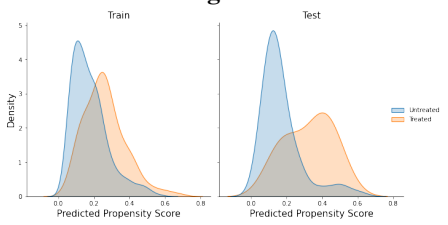
5.1.3 Effect of self-supervised auto-balancing term, binary cross-entropy term and targeted regularization term

The effect of the self-supervised auto-balancing term can be analyzed by comparing row 0-4 to row 5-14 in Table 2. We can see that optimal train and test performance can be achieved when the network is regularized only with such term. Additionally, we can also see that same optimal results can be achieved with different values of λ_{BAL}. The effect of the binary cross-entropy term can be analyzed by comparing row 1-4 to row 5-12. We can see that for ReLUs not adopting the binary cross-entropy improves the performance, at least, by 0.03 on trainset and 0.01 on test-set. The effect of the targeted regularization term can be analyzed by comparing row 0 (Dragonnet) to row 1. We can see that adopting such term is less than optimal for all the activation functions. However, our best test result for Dragonnet is better than the one reported in the original paper (Shi et al., 2019), obtained with a lower batch-size. In the supplementary Section A.3 we discuss in detail why working with large batch-size and SGD is beneficial for the binary cross-entropy objective.

5.2 Treated and untreated distributions induced by the learned representation

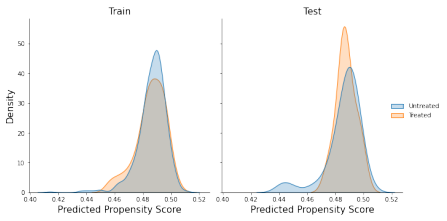
To explain the superior performance of BCAUSS we show that the learned treated and untreated distributions of BCAUSS are less dissimilar than the ones of Dragonnet which implies, thanks to Theorem 5, a lower upper bound of the related expected error in learning ITEs. We perform such analysis on the IHDP dataset in this section. In the supplementary section A.1 the same analysis is repeated on the Jobs dataset, showing, again, that BCAUSS learns less dissimilar treated and untreated representations than Dragonnet. In Table 3, both on train-set and test-set, for Dragonnet we have KS test with p-value $< 1\%$, while for BCAUSS we don't. Wasserstein distance is one order of magnitude lower on train-set and two order of magnitudes lower on test-set for BCAUSS compared to Dragonnet. MMD measures are several order of magnitudes lower for BCAUSS compared to Dragonnet, both on train-set and test-set.

Dragonnet



	Train	Test
KS test	0.37	0.59
p-value	1.33E-13	0.0071
Wasserstein dist.	0.0826	0.1583
MMD (linear)	0.0068	0.0238
MMD (rbf)	0.0127	0.0453
MMD (poly)	0.0015	0.0062

BCAUSS



	Train	Test
KS test	0.0851	0.2574
p-value	0.4026	0.6321
Wasserstein dist.	0.0016	0.0042
MMD (linear)	5.960E-07	3.541E-06
MMD (rbf)	1.123E-06	7.102E-06
MMD (poly)	5.584E-07	2.925E-06

Table 3: Comparison between the treated and untreated distributions induced by the learned representation on Dragonnet and BCAUSS from one experiment of IHDP dataset. Prior covariate treated and control groups have KS test with p-value $< 1\%$, Wasserstein distance 0.1045, MMD(linear) 0.0108, MMD(rbf) 0.0203, MMD(poly) is 0.0023.

Furthermore, we can notice that the variance of the distribution induced by the learned representation of BCAUSS is one order of magnitude lower, consistently to how they should be in RCTs. For example, on test-set for BCAUSS the standard deviation on treated is 0.0128 vs. 0.0064 for untreated, while for Dragonnet, the standard deviation on treated is 0.1137 vs. 0.1204 for untreated. Table 4 shows the same comparison averaging 1,000 experiments of the IHDP dataset, confirming that treated and untreated distributions learned by BCAUSS are less dissimilar than the ones learned by Dragonnet. Hence, adopting the auto-balancing self-supervised objective of eq. 2, the network learns representations that induce distributions analogous to RCTs, where $IPM_G\left(p_\Phi^{t=0}, p_\Phi^{t=1}\right) \approx 0$. As a consequence, the back-door path of the related causal graph from treatment to outcome tends to be blocked, enabling unbiased estimate of causal treatment effect.

	Dragonnet		BCAUSS	
	Train	Test	Train	Test
KS (p-value $< 1\%$)	100%	20.5%	9.9%	1.7%
Wasserstein dist.	0.0940	0.0712	0.0020	0.0043
MMD (linear)	0.0090	0.0051	3.900E-06	1.584E-05
MMD (rbf)	0.0168	0.0097	7.795E-06	3.167E-05
MMD (poly)	0.0020	0.0011	3.809E-06	1.580E-05

Table 4: Comparison between the distributions of treated vs. untreated induced by the learned representation on Dragonnet and BCAUSS averaging 1,000 experiments of the IHDP dataset.

6 Conclusion

We introduced BCAUSS, a multi-task deep neural network for causal treatment effect estimation able to achieve minimal dissimilarity in learning treated and untreated distributions, thanks to the adoption of an auto-balancing self-supervised objective. Experiments on real and benchmark datasets show BCAUSS consistently produces less biased estimates than previously published state-of-art methods. Empirical analysis shows that such a property is due to the representations learned by the network, that is incentivized by our objective to learn less dissimilar treated and untreated distributions, consistently to how they should be in RCTs.

Acknowledgments

We would like to thank Claudia Shi for her guidance in replicating Dragonnet results.

Reproducibility Statement

To contribute to the methods' reproducibility, we provide in the supplementary material the code for replicating experiments and Jupyter Notebooks used to perform our analysis. In the final paper we

will report proper repository link. In Section 4.1 we report the links to download benchmarks and real-world datasets for replicating experiments and we describe the data processing steps referencing the original papers where such datasets were first introduced. Also, in Section 4.2 the evaluation metrics for each dataset are defined explaining how to compute them.

REFERENCES

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6a508a60aa3bf9510ea6acb021c94b48-Paper.pdf. 2

Ahmed M. Alaa, Michael Weisz, and Mihaela van der Schaar. Deep counterfactual networks with propensity-dropout, 2017. 2

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1510489113. 2

Susan Athey, Guido W. Imbens, and Stefan Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions, 2018. 1, 2

P. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46:399 – 424, 2011. 1, 2

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on Treatment Effects after Selection among High-Dimensional Controlsâ . *The Review of Economic Studies*, 81(2): 608–650, 11 2013. ISSN 0034-6527. doi: 10.1093/restud/rdt044. URL https://doi.org/10.1093/restud/rdt044. 1, 2

C.S. Will Belthangady and Beau Norgeot. Minimizing bias in massive multi-arm observational studies with bcaus: Balancing covariates automatically using supervision. *BMC Medical Research Methodology*, 2021. 2, 1

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2013. 1, 2

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and causal parameters, 2017. 1, 2

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266 – 298, 2010. doi: 10.1214/09-AOAS285. URL https://doi.org/10.1214/09-AOAS285. 1b

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 685–693, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/cuturi14.html. 4.2

Rajeev Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161, 2002. URL https://EconPapers.repec.org/RePEc:tpr:restat:v:84:y:2002:i:1:p:151-161. 4.1

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html. 3.1, 4.2

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 01 2011. doi: 10.1198/jcgs.2010.08162. URL https://doi.org/10.1198/jcgs.2010.08162. 4.1, 4.2

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 3020–3029, 2016. 1a, 1b

David C. Klonoff. The expanding role of real-world evidence trials in health care decision making. *Journal of Diabetes Science and Technology*, 2020. 1

Andrey Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. In *Giornale dell'Istituto Italiano degli Attuari*, 1933. 4.2

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, Feb 2019. ISSN 1091-6490. doi: 10.1073/pnas.1804597116. 2

Robert LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–20, 1986. URL https://EconPapers.repec.org/RePEc:aea:aecrev:v:76:y:1986:i:4:p:604-20. 4.1

Christos Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *NIPS*, 2017a. 4.2

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models, 2017b. 1a

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017c. URL https://proceedings.neurips.cc/paper/2017/file/94b5bde6de888ddf9cde6748ad2523d1-Paper.pdf. 2

Pottegård A Witte DR Thomsen RW Munk NE, Knudsen JS. Differences between randomized clinical trial participants and real-world empagliflozin users and the changes in their glycated hemoglobin levels. *JAMA*, 2020. 1

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077. 3.2

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X. 3

H. Robbins. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 2007. 4.3

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41. 2, 3, 3.2, A.2, A.3

Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks, 2019. 2

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3076–3085, 2017a. 1a, 1b

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3076–3085, 2017b. 1, 2, 3.1, 3.1, 4.1, 4.1, 4.2, 4.2, 1a, 1b, 1

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 06–11 Aug 2017c. URL http://proceedings.mlr.press/v70/shalit17a.html. 4.2

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett (eds.), *NeurIPS*, pp. 2503–2513, 2019. URL http://dblp.uni-trier.de/db/conf/nips/nips2019.html#ShiBV19. 1, 2, 3.2, 3.2, 3.3, 4.1, 4.4, 1a, 5.1.3, A.3

Nikolai Smirnov. Table for estimating the goodness of fit of empirical distributions. In *Annals of Mathematical Statistics*, 1948. 4.2

Fukumizu Kenji Gretton Arthur SchÂšolkopf Bernhard Lanckriet Gert RG et al. Sriperumbudur, Bharath K. On the empirical estimation of integral probability metrics. In *Electronic Journal of Statistics*, 2012. 3.1

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(5):141–158, 2009. URL http://jmlr.org/papers/v10/su09a.html. 2

Mark J. van der Laan and Sherri Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Publishing Company, Incorporated, 1st edition, 2018. ISBN 3319653032. 1, 2

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science Business Media, 2008. 4.2

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests, 2017. 2

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. 1b

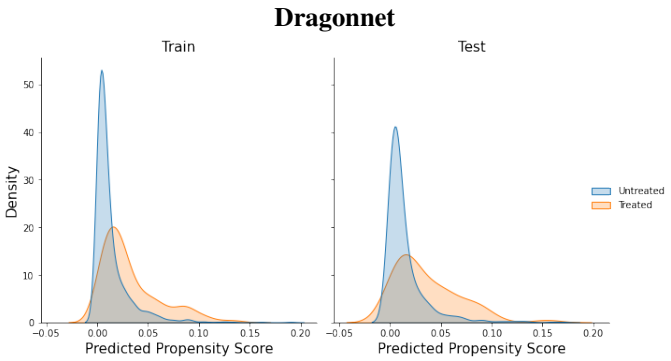
Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/a50abba8132a77191791390c3eb19fe7-Paper.pdf. 4.2

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constr. Approx*, pp. 289–315, 2007. 4.3

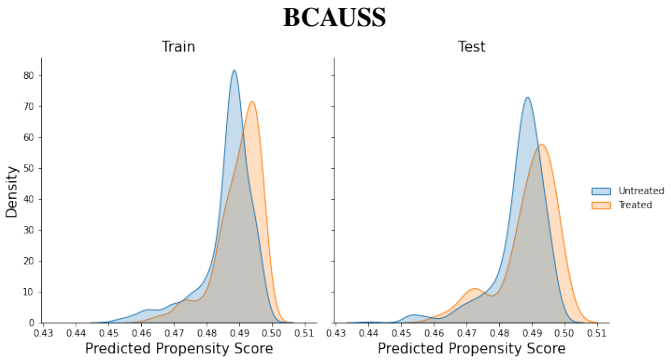
Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: estimation of individualized treatment effects using generative adversarial nets. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=ByKWUeWA-. 2, 4.4, 1a, 1b, 1

Weijia Zhang, Thuc Duy Le, Lin Liu, Zhi-Hua Zhou, and Jiuyong Li. Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 33(15):2372–2378, 2017. doi: 10.1093/bioinformatics/btx174. 2

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors, 2021. 2



	Train	Test
Wasserstein dist.	0.0192	0.0217
MMD (linear)	0.0003	0.0004
MMD (rbf)	0.0007	0.0009
MMD (poly)	1.839E-06	2.715E-06



	Train	Test
Wasserstein dist.	0.0035	0.0030
MMD (linear)	1.223E-05	7.134E-06
MMD (rbf)	2.446E-05	1.425E-05
MMD (poly)	1.150E-06	6.787E-06

Table 5: Comparison between the distributions of treated vs. untreated induced by the learned representation on Dragonnet and BCAUSS from one experiment of Jobs dataset.

A Appendix

We start providing further empirical evidence that adopting our auto-balancing self-supervised objective the network learns representations analogous to RCTs (Section A.1). Then, in the Section A.2 we analyze the effect of back-propagating the gradient with respect to our self-supervised auto-balancing objective compared to binary cross-entropy (Section A.3).

A.1 Comparison between learned treated and untreated distributions on Jobs

Tab. 5 shows the comparison between the distributions of treated vs. untreated induced by the learned representation on Dragonnet and BCAUSS from one experiment of Jobs dataset, where we find confirmation of what already observed on the IHDP dataset, i.e. treated and untreated distributions learned by BCAUSS are more similar than the ones learned by Dragonnet. Specifically, Wasserstein distance is one order of magnitude lower on train-set and two order of magnitudes lower on test-set compared to Dragonnet. MMD with the considered kernels are one order of magnitude lower for BCAUSS compared to Dragonnet, both on train-set and test-set, except for the polynomial kernel where they are the same order of magnitude.

A.2 The effect of back-propagating the gradient with respect to the auto-balancing objective

The parameters of the network of fig. 1 are updated at each iteration to minimize $J(\theta; P_n)$, which in case of batch gradient descent means

$$\theta^{(i+1)} := \theta^{(i)} - \alpha \frac{\partial J}{\partial \theta} = \theta^{(i)} - \alpha \left(\lambda_{BAL} \frac{\partial L_{BAL}}{\partial \theta} + \frac{1}{n} \sum_{1 \leq j \leq n} \frac{\partial L_{REG}}{\partial \theta}^{(j)} \right) \tag{7}$$

, where α is the learning rate. Hence, regarding the factual and counterfactual losses, using chain rule for partial derivatives[1], we have

[1]If $f : \mathbb{R} \to \mathbb{R}$ is a differentiable function applied element-wise to a vector $\mathbf{x}$, i.e. $z = f(\mathbf{x}) = [f(x_1), f(x_2), \ldots, f(x_d)]^T$, then it is possible to show that the Jacobian $\frac{\partial z}{\partial \mathbf{x}} = diag(f'(\mathbf{x})) =$

$$\frac{\partial L_{REG}}{\partial z_{j,5}^{t_j}}^{(j)} = \frac{\partial L_{REG}}{\partial z_{j,6}^{t_j}}^{(j)} \frac{\partial z_{j,6}^{t_j}}{\partial a_{j,5}^{t_j}} \frac{\partial a_{j,5}^{t_j}}{\partial z_{j,5}^{t_j}} = \quad 2\left[Y_{t_j}^{(j)} - Q\left(t_j, \mathbf{x}_i\right)\right] W_6^{t_j} \odot H\left(z_{j,5}^{t_j}\right) \tag{8}$$

$$\frac{\partial L_{REG}}{\partial z_{j,4}^{t_j}}^{(j)} = \frac{\partial L_{REG}}{\partial z_{j,5}^{t_j}}^{(j)} \frac{\partial z_{j,5}^{t_j}}{\partial a_{j,4}^{t_j}} \frac{\partial a_{j,4}^{t_j}}{\partial z_{j,4}^{t_j}} = \quad \frac{\partial L_{REG}}{\partial z_{j,5}^{t_j}}^{(j)} W_5^{t_j} \odot H\left(z_{j,4}^{t_j}\right) \tag{9}$$

$$\frac{\partial L_{REG}}{\partial z_{j,3}}^{(j)} = \frac{\partial L_{REG}}{\partial z_{j,4}^{t_j}}^{(j)} \frac{\partial z_{j,4}^{t_j}}{\partial a_{j,3}} \frac{\partial a_{j,3}}{\partial z_{j,3}} = \quad \frac{\partial L_{REG}}{\partial z_{j,4}^{t_j}}^{(j)} W_4^{t_j} \odot H\left(z_{j,3}\right) \tag{10}$$

$$\frac{\partial L_{REG}}{\partial z_{j,2}}^{(j)} = \frac{\partial L_{REG}}{\partial z_{j,3}}^{(j)} \frac{\partial z_{j,3}}{\partial a_{j,2}} \frac{\partial a_{j,2}}{\partial z_{j,2}} = \quad \frac{\partial L_{REG}}{\partial z_{j,3}}^{(j)} W_3 \odot H\left(z_{j,2}\right) \tag{11}$$

$$\frac{\partial L_{REG}}{\partial z_{j,1}}^{(j)} = \frac{\partial L_{REG}}{\partial z_{j,2}}^{(j)} \frac{\partial z_{j,2}}{\partial a_{j,1}} \frac{\partial a_{j,1}}{\partial z_{j,1}} = \quad \frac{\partial L_{REG}}{\partial z_{j,2}}^{(j)} W_2 \odot H\left(z_{j,1}\right) \tag{12}$$

where $H\left(\cdot\right)$ is the Heaviside step function. In order to compute, for example, $\frac{\partial L_{REG}}{\partial W_1}^{(j)}$, we have

$$\delta_{1,j} = \frac{\partial L_{REG}}{\partial z_{j,1}}^{(j)} \tag{13}$$

$$\frac{\partial L_{REG}}{\partial W_1}^{(j)} = \frac{\partial L_{REG}}{\partial z_{j,1}}^{(j)} \frac{\partial z_{j,1}}{\partial W_1} = \quad \delta_{1,j}^T \mathbf{x}_j^T \tag{14}$$

In eq. 14 it has been used the practice (in a slight abuse of notation) of making the Jacobian $\frac{\partial L_{REG}}{\partial W_1}^{(j)}$ of the same shape as W_1. This is a well consolidated practice in deep learning literature, since this matrix has the same shape as W_1 we could just subtract it (times the learning rate) from W_1 when doing gradient descent. Relationships for other matrices and bias terms can be derived with similar reasoning. In the same way, regarding the auto-balancing loss, applying the product rule and the chain rule, we have

$$\frac{\partial L_{BAL}}{\partial W_1} = \frac{2}{d} \sum_{1 \leq j \leq d} \left\{ f_j\left(g, I_n\right) \sum_{1 \leq i \leq n} \frac{\partial f_j}{\partial g_i} \frac{\partial g_i}{\partial z_i^g} \frac{\partial z_i^g}{\partial z_{i,3}} \frac{\partial z_{i,3}}{\partial z_{i,2}} \frac{\partial z_{i,2}}{\partial z_{i,1}} \frac{\partial z_{i,1}}{\partial W_1} \right\} \tag{15}$$

$$= \frac{2}{d} \sum_{1 \leq j \leq d} \left\{ f_j\left(g, I_n\right) \sum_{1 \leq i \leq n} \frac{\partial f_j}{\partial g_i} \sigma\left(z_i^g\right) \left[1 - \sigma\left(z_i^g\right)\right] \left[W_4^q \odot H\left(z_{i,3}\right) W_3 \odot H\left(z_{i,2}\right) W_2 \odot H\left(z_{i,1}\right)\right]^T \mathbf{x_i^T} \right\} \tag{16}$$

where $f_j\left(g, I_n\right) = \left(\frac{\sum_{1 \leq i \leq n} \frac{t_i}{g(\mathbf{x}_i)} x_{i,j}}{\sum_{1 \leq i \leq n} \frac{t_i}{g(\mathbf{x}_i)}} - \frac{\sum_{1 \leq i \leq n} \frac{1-t_i}{1-g(\mathbf{x}_i)} x_{i,j}}{\sum_{1 \leq i \leq n} \frac{1-t_i}{1-g(\mathbf{x}_i)}} \right)$ and it is used the derivative of the sigmoid function, i.e. $\frac{\partial \sigma(a)}{\partial a} = \sigma\left(a\right)\left[1 - \sigma\left(a\right)\right]$. Relationships for other matrices and bias terms can be derived with similar reasoning. When we plug eq. 14 and 16 into 7 for each matrix and bias term of the network, we can understand how back-propagation works on BCAUSS. First, the term of eq. 8 is a standard regression term depending on the difference between observed and predicted outcome. This term is the same for networks like Dragonnet and TARNET. The term of eq. 16 is our adjustment due to the dissimilarity between learned treated and untreated distributions. Specifically, if there is no dissimilarity between learned treated and untreated distributions, then this term is zero and the causal problem degenerates into a standard regression problem, as it should be. In this case,

$\begin{pmatrix} \frac{\partial f(x_1)}{\partial x_1} & 0 & \cdots & 0 \\ 0 & \frac{\partial f(x_2)}{\partial x_2} & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \frac{\partial f(x_d)}{\partial x_d} \end{pmatrix}$. Since multiplication by a diagonal matrix is the same as doing element-wise multiplication by the diagonal, we could also write $\odot f'\left(\mathbf{x}\right)$ when applying the chain rule.

because the learned representation of the network is able to re-create the same conditions of RCTs, there is no need of adjustments and *the original causal problem defined on the covariates space becomes a standard regression problem on the learned representation space*. On the other hand, if there is dissimilarity between learned treated and untreated distributions, i.e. $\exists j \in \{1, ..., d\}$ so that $f_j(g, I_n) \neq 0$, then the updates of the network parameters are adjusted with the term of eq. 16, which depends on $f_j(g, I_n)$ and the derivative of $f_j(g, I_n)$ with respect to the predicted propensity scores of each observation. Now, we can understand why batch-gradient descent or SGD with large batch-size works better than SGD with small batch-size and even Adam. For example, if for Adam for a given j we have that $f_j(g, I_n) = 0$ but $f_j(g, I_k) \neq 0$, where k is the batch-size ($k < n$), we update our parameters with the exponentially weighted average of the terms of eq. 16 and the exponentially weighted average of the (element-wise) squares of the magnitudes of same terms. These updates are compensated with the updates of the following batches not in a linear way, but in a exponentially weighted averaging way leading, in general, to sub-optimal states of the network. Further, if the batch size is too small, we can observe numerical problem for SGD due to specific values of $f_j(g, I_k)$, which are smoothed for Adam. The same reasoning holds for Adam if $f_j(g, I_n) \neq 0$ on the whole train-set, i.e. the network is updated with a "bias estimate" between learned treated and untreated distributions which is not the actual one. We might wonder what is the causal interpretation of $f_j(g, I_n)$ for $j \in \{1, ..., d\}$ and why it is so important. Such term is the difference of the g-weighted means of j-th covariate for the treatment and control groups. Hence, by minimizing such term for each $j \in \{1, ..., d\}$, $g(\mathbf{x})$ is incentivized to act like a balancing score, in the sense of Rosenbaum & Rubin (1983). And, if $g(\mathbf{x})$ is a balancing score in the sense of Rosenbaum & Rubin (1983), at any value of $\mathbf{x} \in \mathcal{X}$, the difference between the treatment and control predicted outcome is an unbiased estimate of the treatment effect at that value of the balancing score.

A.3 Comparison with the binary cross-entropy objective

A different objective used in literature to train the propensity score estimator is the binary cross-entropy objective of eq. 5. Even Rosenbaum & Rubin (1983) claim that the propensity score "*may be estimated from observed data, perhaps using a model such as a logit model*". We find very illuminating such use of "*perhaps*" in the previous sentence. Maybe, what the authors mean is that for estimating the propensity score adopting a logit model, we need a train-sample very representative of the underlying covariate distribution. With this in mind let analyze the effect of back-propagating the gradient with respect to the binary cross-entropy objective. The parameters of the network are updated at each iteration according to the following optimization step

$$\theta^{(i+1)} := \theta^{(i)} - \alpha \frac{\partial J}{\partial \theta} = \theta^{(i)} - \frac{\alpha}{n} \left[\sum_{1 \leq j \leq n} \left(\frac{\partial L_{REG}}{\partial \theta}^{(j)} + \lambda_{BCE} \frac{\partial L_{BCE}}{\partial \theta}^{(j)} \right) \right] \tag{17}$$

, where α is the learning rate. The components $\frac{\partial L_{REG}}{\partial \theta}$ can be derived like in Section A.2, while for components $\frac{\partial L_{BCE}}{\partial \theta}$ we have

$$\frac{\partial L_{BCE}}{\partial W_1}^{(j)} = \left\{ \left[g(\mathbf{x_j}) - t_j \right] \mathbf{W}_4^g \odot H(z_{j,3}) W_3 \odot H(z_{j,2}) W_2 \odot H(z_{j,1}) \right\}^T \mathbf{x_j^T} \tag{18}$$

Relationships for other matrices and bias terms can be derived with similar reasoning. The term of eq. 18 is the adjustment (times λ_{BCE}) for the gradient update $\frac{\partial L_{REG}}{\partial \theta}^{(j)}$ depending on the difference of the decision of treatment of the j-th observation and the related predicted propensity score. Hence, we can understand why working with large batch-size and SGD is beneficial also for the binary cross-entropy objective, as recalled already in (Shi et al., 2019). For example, if the true propensity score for a given $\mathbf{x_j} \in \mathcal{X}$ is 1/3 but in the current batch there is only one observation belonging to the treated group having covariate $\mathbf{x_j}$, then we update the parameters of the network with a component $\left\{ \left[g(\mathbf{x_j}) - t_j \right] \mathbf{W}_4^g \odot H(z_{j,3}) W_3 \odot H(z_{j,2}) W_2 \odot H(z_{j,1}) \right\}^T \mathbf{x_j^T}$ which can be zero, in general, only in case $g(\mathbf{x_j}) = 1$, which is the wrong value for $g(\mathbf{x_j})$ to converge. On the other hand, adopting SGD such component can be hopefully compensated in the following batches with two observations like $\left\{ g(\mathbf{x_j}) \mathbf{W}_4^g \odot H(z_{j,3}) W_3 \odot H(z_{j,2}) W_2 \odot H(z_{j,1}) \right\}^T \mathbf{x_j^T}$, if the train-set is representative of the underlying covariate distribution. On the other hand, assuming the network

was already able to estimate $g(\mathbf{x_j})$ as 1/3, instead of having no components for all the observations having covariates $\mathbf{x_j}$, we have the three ones mentioned above, hopefully compensating each other (this is not guaranteed as they belong to different batches, with the network having different initial states). Ideally, at each optimization step, we need a batch-size large enough to represent the underlying covariate distribution. Indeed, in our experiments we noticed improvements in models adopting the binary cross-entropy objective with large batch-sizes. As final step, let assume we adopt the maximum possible batch-size and let ask what are the implications for the network if we adopt the the binary cross-entropy objective. The network should learn an internal representation so that $g(\mathbf{x}) \approx p(t=1|\mathbf{x})$ which, in general, highly depends on prior treated an control distributions, e.g. $logit(p(t=1|\mathbf{x})) = logit(p(\mathbf{x}|t=1)) + logit(p(t=1))$. In particular, if such prior distributions are highly dissimilar, as unfortunately might happen in observational datasets, the related learned distributions are affected as well, and Theorem 5 states that in this case the upper bound of the related expected error in learning individual treatment effect increases. On the contrary, if we adopt an objective explicitly designed to learn internal representations that re-create the same (or very similar) conditions of RCTs, where treated and untreated distributions are by design the same, we end up with a lower upper bound.