DR-VIDAL - Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning for Estimation of Counterfactuals and Treatment Effects

ABSTRACT

Randomized controlled trials (RCTs) have been the gold-standard, in health sciences among others, to assess causal effects of interventions on outcomes, e.g., medical treatments or lifestyle changes. However, conducting RCTs is not always feasible due to operational or ethical constraints. Alternatively, causal effects can be drawn from observational, real-world data, but the data generation and collection process might not be randomized and contain underlying bias. Several techniques for addressing bias in treatment assignments, and for predicting individualized treatment effects (ITEs) –with counterfactuals, i.e., outcomes for alternative treatment scenarios– have been proposed, from propensity score matching, to ensemble tree-based learning, to recent breakthrough in deep learning, e.g., the Causal Effect Variational Autoencoder (CEVAE) or the Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE). In this work, we propose a novel deep learning approach, the Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning (DR-VIDAL) that incorporates the following key characteristics: (1) disentangled variational autoencoder with latent variables under a causal structure substantially different from the CEVAE; (2) feature attention on high-dimensional datasets and noise-reduction; (3) information-theoretic optimization for instance generation and prediction of counterfactuals; and (4) doubly robust ITE estimation. Tests performed on synthetic and real-world datasets showed that the DR-VIDAL outperforms several other state-of-the-art techniques. The utility of DR-VIDAL is not only with respect to prediction, for which the doubly robustness is assured, but also for more general inference tasks. The code is available under the MIT license at: https://bitbucket.org/goingdeep2406/dr-vidal/src/master/

CCS CONCEPTS

• **Deep Learning** → **Representation Learning**; *Variational Autoencoder*; Generative Adversarial Network; • **Causal Inferences** → Doubly Robust Estimation .

KEYWORDS

causal AI, biomedical informatics, generative adversarial networks, variational inference, information theory, doubly robust

ACM Reference Format:

. 2018. DR-VIDAL - Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning for Estimation of Counterfactuals and Treatment Effects. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Evaluating causal effects of interventions on outcomes is key to knowledge and progress in many fields, e.g., medicine, psychology, public health, and policy making. A typical scenario in medicine is to determine whether a treatment (e.g., lipid-lowering medication) is effective to reduce risk or cure an illness (e.g., cardiovascular disease). Randomized controlled trials (RCTs) are often considered to be the best practice for evaluating causal effects [33]. In an RCT, a treatment is assigned randomly to individuals, with some getting it and some being assigned a placebo, making it independent of an individual's characteristics, thus avoiding selection bias and influence on the outcome from other sources rather then the treatment (albeit other possible bias, e.g., people understanding if they have been given a placebo). However, RCTs are not always feasible due to ethical and legal constraints. For instance, if one wanted to evaluate whether college education is the cause of good salary, it would not be ethical to randomly pick teenagers and randomize their attendance to college. So, in many cases, observational data, i.e., real-world data collected retrospectively and not randomized, are the only usable source. Unfortunately, observational data are often plagued with various biases –because the data generation processes are largely unknown– such as confounding (i.e., spurious causal effects on outcomes by features that are correlated with a true unmeasured cause) and colliders (i.e., mistakenly including effects of an outcome as predictors), making it difficult to infer causal claims [16]. In other words, without randomization, it is quite difficult to distinguish conditional association from causation. Another problem with causal effect estimation is that, in both RCTs and observational datasets, only factual outcomes are available, since clearly an individual cannot be treated and not-treated at the same time. Counterfactuals are alternative predictions that respond to the question "what outcome would have been observed if a person had been given the placebo instead of the treatment?" If models are biased, counterfactual predictions can be wrong, and interventions can be ineffective or harmful [28].

Traditional statistical approach for estimating treatment effects, taking into account possible bias from pre-treatment characteristics, are propensity score matching (PSM) and inverse probability weighting (IPW) [5]. The propensity score is a scalar estimate representing the conditional probability of receiving a certain treatment, given a set of measured pre-treatment covariates. By matching (or

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

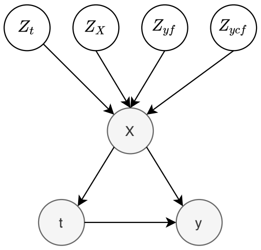

Figure 1: Directed acyclic graph modelling the causal relationships among a treatment t, outcome y and pre-treatment covariates X, under a latent space Z.

weighting) treated and control subjects according to their propensity score, a balance in pre-treatment covariates is induced, mimicking a randomization of the treatment group assignment. However, traditional PSM approach accounts only for measured covariates, and latent bias may remain after matching [14]. PSM has been implemented historically with logistic (linear) regression, coupled with different feature selection methods in presence of high-dimensional datasets [34]. A problem with PSM is that it often decreases the sample size due to matching, while IPW can be affected by skewed, heavy-tailed weight distributions. Machine learning approaches have been introduced more recently, e.g., Bayesian additive regression trees [17] and counterfactual random forests [35]. Big data also led to the flourishing of deep learning tailored to causal inference [19]. Notable examples include Treatment-Agnostic Representation Network (TARNet) [31], Dragonnet [32], Deep Counterfactual Network with Propensity-Dropout (DCN-PD) [2], Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE) [36], and Causal Effect Variational Autoencoder (CEVAE) [23].

1.1 Contribution

This work introduces a novel deep learning approach for treatment effect estimation and counterfactual prediction, named the *Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning* (DR-VIDAL). The main features of DR-VIDAL are:

- Incorporation of an underlying causal structure where the observed pre-treatment covariate set X is decomposed into four independent latent variables $Z_t, Z_X, Z_{yf}, Z_{ycf}$, inducing confounding on both the treatment and the outcome (Figure 1).
- Latent variables are inferred using a variational autoencoder (VAE) [21] with disentanglement.
- A generative adversarial network (GAN) [15] with variational information maximization [8] generates (synthetic) complete tuples of covariates, treatment, factual and counterfactual outcomes.
- Individual treatment effects (ITE) are estimated on complete datasets with a downstream, four-headed deep learning block which is doubly robust [12, 13].

To our knowledge, this is the first time in which VAE, GAN, information theory and doubly robustness are amalgamated together into a counterfactual prediction method. By performing test runs on synthetic and real-world datasets, we show that DR-VIDAL can outperform a number of state-of-art tools for estimating ITE.

2 PROBLEM FORMULATION

We utilize the *potential outcomes* framework [29, 30]. Let us consider a treatment T (binary for ease of reading, but the theory can be extended to multiple treatments) which can be prescribed to a population sample of size N. The individuals are characterized by a set of pre-treatment background covariates $\boldsymbol{X}$, and a health outcome Y is measured after treatment. We define each subject i with the tuple $\{X, T, Y\}_{i=1}^{N}$, where Y_i^0 and Y_i^1 are the potential outcomes when applying treatments $T_i = 0$ and $T_i = 1$, respectively. The ITE $\tau(\boldsymbol{x})$ for subject i with pre-treatment covariates $\boldsymbol{X}_i = \boldsymbol{x}$, is defined as the difference in the average potential outcomes under both treatment interventions (i.e., treated vs. not treated), conditional on $\boldsymbol{x}$, i.e.,

$$\tau(\boldsymbol{x}) = \mathbb{E}[Y_i^1 - Y_i^0 \mid \boldsymbol{X}_i = \boldsymbol{x}] \tag{1}$$

The ITE cannot be calculated directly give the inaccessibility to both potential outcomes, as only factual outcomes can be observed, while the others (counterfactual) can be considered as missing values. However, when the potential outcomes are made independent of the treatment assignment, conditionally on the pre-treatment covariates, i.e., $\{Y^1, Y^0\} \perp T \mid \boldsymbol{X}$, the ITE can then be estimated as $\tau(\boldsymbol{x}) = \mathbb{E}[Y^1 \mid T = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y^0 \mid T = 0, \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y \mid T = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid T = 0, \boldsymbol{X} = \boldsymbol{x}]$. Such assumption is called strongly ignorable treatment assignment (SITA) [18, 26]. By further averaging over the distribution of $\boldsymbol{X}$, the average treatment effect (ATE) τ_{01} can be calculated as

$$\tau_{01} = \mathbb{E}[\tau(\boldsymbol{X})] = \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] \tag{2}$$

ITE and ATE can be calculated straightforward with stratification matching of $\boldsymbol{x}$ in treatment and control groups, but the calculation becomes unfeasible as the covariate space increases in dimensions.

The propensity score $\pi(x)$ represents the probability of receiving the treatment $T = 1$ conditioned on the pre-treatment covariates $X = x$ [29], denoted as

$$\pi(\boldsymbol{x}) = P(T = 1 | \boldsymbol{X} = \boldsymbol{x}). \tag{3}$$

The propensity score can be calculated using a regression function, e.g., logistic, and then ITE/ATE can be calculated by matching (PSM) or weighting (IPW) instances through $\pi(\boldsymbol{x})$, in a doubly robust way [27], or with other approaches [4, 9–11, 24, 25, 27, 35]. In the next section, we describe those based on deep learning.

3 RELATED WORK

A comprehensive work on characterizing the conditions and the limits of heterogeneous treatment effect estimation using deep learning has been provided [1]. The sample size plays an important role, e.g., estimations on small sample sizes are affected by selection bias, whilst on large sample sizes by algorithmic design. Our work is motivated mostly from the advancement in ITE estimation brought by DCN-PD [2], CEVAE [23], GANITE [36], TARNet [31], and Dragonnet [32]. DCN-PD is a doubly robust, multitask network for counterfactual prediction, where propensity scores are used to determine a dropout probability of samples to regularize training, carried out in alternating phases, using treated and control

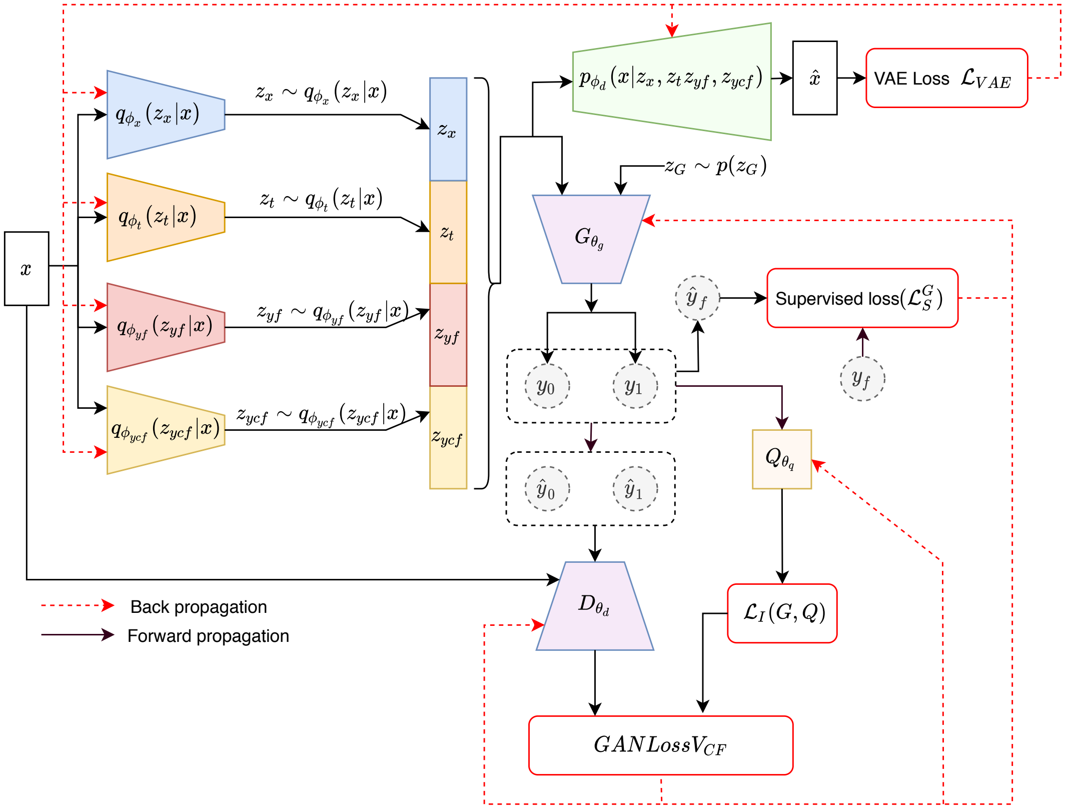
x
$q_{\phi_x}(z_x|x)$
$q_{\phi_t}(z_t|x)$
$q_{\phi_{yf}}(z_{yf}|x)$
$q_{\phi_{ycf}}(z_{ycf}|x)$
$z_x \sim q_{\phi_x}(z_x|x)$
$z_t \sim q_{\phi_t}(z_t|x)$
$z_{yf} \sim q_{\phi_{yf}}(z_{yf}|x)$
$z_{ycf} \sim q_{\phi_{ycf}}(z_{ycf}|x)$
z_x
z_t
z_{yf}
z_{ycf}
$p_{\phi_d}(x|z_x, z_t z_{yf}, z_{ycf})$
$\hat{x}$
VAE Loss $\mathcal{L}_{VAE}$
$z_G \sim p(z_G)$
G_{θ_g}
$\hat{y}_f$
Supervised loss($\mathcal{L}_S^G$)
y_f
y_0
y_1
$\hat{y}_0$
$\hat{y}_1$
Q_{θ_q}
D_{θ_d}
$\mathcal{L}_I(G, Q)$
$GANLossV_{CF}$
Back propagation
Forward propagation

Figure 2: Architecture of the counterfactual network to estimate the counterfactual outcome.

batches. CEVAE uses VAE and GAN to identify latent variables from an observed pre-treatment vector and to generate counterfactuals. GANITE generates proxies of counterfactual outcomes using a GAN, and feeds them to an ITE generator. In TARNet, each sample from the treated and control group is associated with a weight indemnifying group imbalance. Dragonnet is a modified TARNet that introduces targeted regularization based on propensity scores. In the following sections, we discuss in detail the novelty and the differences in architectural design and training mechanisms of DR-VIDAL with respect to the aforementioned approaches.

4 PROPOSED METHODOLOGY

DR-VIDAL architecture can be decomposed into three main parts: (1) a VAE to disentanglement of the latent variables, (2) a GAN to generate the counterfactual outcomes (3) a doubly robust module to estimate ITE. The architectural layout and the training algorithm of the first two components (VAE and GAN) are illustrated in Figure 2 and Algorithm 1, respectively. For the third component (doubly robust estimator), architecture and training algorithm are shown in Figure 3 and Algorithm 2, respectively.

4.1 Latent variable disentanglement with VAE

We assume that the observed covariates x are generated from an independent latent space z, composed by $z_x \sim p(z_x)$, $z_t \sim p(z_t)$, $z_{yf} \sim p(z_{yf}), and z_{ycf} \sim p(z_{ycf})$, which denote the latent variables for the covariates x, treatment indicator t, and factual outcomes y_f and y_{cf}, respectively. This decomposition follows the causal structure shown in Figure 1. The goal is to infer the posterior distribution $p(z_x, z_t, z_{yf}, z_{ycf}|x)$, which is however harder to optimize. So, we use the theory of variational inference [6] to learn the variational posteriors $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x), q_{\phi_{ycf}}(z_{ycf}|x)$, using 4 different neural network encoders with parameters $\phi_x, \phi_t, \phi_{yf}, and \phi_{ycf}$, respectively. Using the latent factors sampled from the learned variational posteriors, we reconstruct x by estimating the likelihood $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})$ via a single decoder parameterized by ϕ_d. The latent factors, assumed Gaussian, are defined as follows:

$$p(z_x) = \prod_{i=1}^{D_{z_x}} \mathcal{N}(z_{x_i}|0,1); \qquad p(z_t) = \prod_{i=1}^{D_{z_t}} \mathcal{N}(z_{t_i}|0,1) \tag{4}$$

$$p(z_{yf}) = \prod_{i=1}^{D_{z_{yf}}} \mathcal{N}(z_{yf_i}|0,1); \quad p(z_{ycf}) = \prod_{i=1}^{D_{z_{ycf}}} \mathcal{N}(z_{ycf_i}|0,1) \tag{5}$$

where $D_{z_x}, D_{z_t}, D_{z_{yf}}, D_{z_{ycf}}$ are the dimensions of the latent factors $z_x, z_t, z_{yf}, z_{ycf}$, respectively. The variational posteriors of

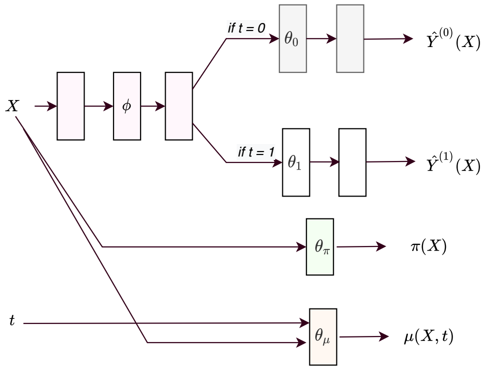

Figure 3: Architecture of the four-headed, doubly robust neural network to calculate individual treatment effects.

the inference of models are defined as:

$$q_{\phi_x}(z_x|x) = \prod_{i=1}^{D_{z_x}} \mathcal{N}(\mu = \hat{\mu}_x, \sigma^2 = \hat{\sigma}_x^2) \tag{6}$$

$$q_{\phi_t}(z_t|x) = \prod_{i=1}^{D_{z_t}} \mathcal{N}(\mu = \hat{\mu}_t, \sigma^2 = \hat{\sigma}_t^2) \tag{7}$$

$$q_{\phi_{yf}}(z_{yf}|x) = \prod_{i=1}^{D_{z_{yf}}} \mathcal{N}(\mu = \hat{\mu}_{yf}, \sigma^2 = \hat{\sigma}_{yf}^2) \tag{8}$$

$$q_{\phi_{ycf}}(z_{ycf}|x) = \prod_{i=1}^{D_{z_{ycf}}} \mathcal{N}(\mu = \hat{\mu}_{ycf}, \sigma^2 = \hat{\sigma}_{ycf}^2) \tag{9}$$

where $\hat{\mu}_x, \hat{\mu}_t, \hat{\mu}_{yf}, \hat{\mu}_{ycf}$ and $\hat{\sigma}_x^2, \hat{\sigma}_t^2, \hat{\sigma}_{yf}^2, \hat{\sigma}_{ycf}^2$ are the means and variances of the Gaussian distributions parameterized by encoders $E_{\phi_x}, E_{\phi_t}, E_{\phi_{yf}}, E_{\phi_{ycf}}$ with parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$ respectively.

The overall evidence lower bound (ELBO) loss of the VAE is expressed as $\mathcal{L}_{ELBO}$ in the following equation,

$$\begin{aligned}
&\mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf}) \\
&= \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}} [\log p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})] \\
&\quad - KL\big(q_{\phi_x}(z_x|x)||p_{\phi_d}(z_x)\big)\big) \\
&\quad - KL\big(q_{\phi_t}(z_t|x)||p_{\phi_d}(z_t)\big)\big) \\
&\quad - KL\big(q_{\phi_{yf}}(z_{yf}|x)||p_{\phi_d}(z_{yf})\big)\big) \\
&\quad - KL\big(q_{\phi_{ycf}}(z_{ycf}|x)||p_{\phi_d}(z_{ycf})\big)\big)
\end{aligned}$$

We minimize the optimization function of the VAE as $\mathcal{L}_{VAE}$ to obtain the optimal parameter of the encoders $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$, and of the decoder ϕ_d as $\mathcal{L}_{VAE}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf}) = -\mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf})$. A detailed derivation of the ELBO loss for the VAE is given in the Appendix A.1.

4.2 Generation of counterfactuals via GAN

After learning the hidden latent codes $z_x, z_t, z_{yf}, z_{ycf}$ from the VAE, we concatenate the latent codes to form z_c, passed to the generator of the GAN block G_{θ_g}, along with a random noise $z_G \sim \mathcal{N}(0, Id)$. G_{θ_g} is parameterized by θ_g, and it outputs the vector $\overline{y}$ of the potential (factual and counterfactual) outcomes. We replace the factual outcome y_f in the generated outcome vector $\overline{y}$ to form $\hat{y}_0$ and $\hat{y}_1$, which are passed to the counterfactual discriminator D_{θ_d}, along with the true covariate vector x. D_{θ_d} is parameterized by θ_d, and is responsible to predict the treatment variable, similarly to GANITE. The loss of the GAN block is defined as:

$$\begin{aligned}
V_{GAN}(G, D) = \mathbb{E}_{x, z_G, z_c} \big[& t^T \log D(x, G(z_G, z_c)) \\
& + (1-t)^T \log(1 - D(x, G(z_G, z_c)) \big]
\end{aligned}$$

where $x \sim p(x), z_G \sim p(z_G)$ and z_c denote the concatenated latent codes $z_x \sim q_{\phi_x}(z_x|x)$, $z_t \sim q_{\phi_t}(z_t|x)$, $z_{yf} \sim q_{\phi_{yf}}(z_{yf}|x)$ and $z_{ycf} \sim q_{\phi_{ycf}}(z_{ycf}|x)$. From $\overline{y}$, we also calculate the predicted

Algorithm 1 Training of the generative adversarial network for counterfactual outcome calculation

Input: Training set $\boldsymbol{X} = \{(x^{(1)}, t^{(1)}, y_f^{(1)}), ..., (x^{(n)}, t^{(n)}, y_f^{(n)})\}$; hyper-parameters $\gamma > 0$; $\lambda > 0$; Encoders: E_{ϕ_x}, E_{ϕ_t}, $E_{\phi_{yf}}$, $E_{\phi_{ycf}}$ with parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$ respectively; Decoder D_{ϕ_d} with parameter D_{ϕ_d}; Generator G_{θ_g}, Discriminator D_{θ_d}, Q network D_{θ_q} with parameters $\theta_g, \theta_d, \theta_q$ respectively

```
Initialize parameters: $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}, \phi_d, \theta_g, \theta_d, \theta_q$
while training do
    $x \leftarrow$ batch of samples from the dataset
    $z_{\mu_x}, z_{\sigma_x} \leftarrow E_{\phi_x}(x)$
    $z_{\mu_t}, z_{\sigma_t} \leftarrow E_{\phi_t}(x)$
    $z_{\mu_{yf}}, z_{\sigma_{yf}} \leftarrow E_{\phi_{yf}}(x)$
    $z_{\mu_{ycf}}, z_{\sigma_{ycf}} \leftarrow E_{\phi_{ycf}}(x)$
    $z_x \leftarrow z_{\mu_x} + \epsilon z_{\sigma_x}$ , where $\epsilon \sim \mathcal{N}(0, Id)$
    $z_t \leftarrow z_{\mu_t} + \epsilon z_{\sigma_t}$ , where $\epsilon \sim \mathcal{N}(0, Id)$
    $z_{yf} \leftarrow z_{\mu_{yf}} + \epsilon z_{\sigma_{yf}}$ , where $\epsilon \sim \mathcal{N}(0, Id)$
    $z_{ycf} \leftarrow z_{\mu_{ycf}} + \epsilon z_{\sigma_{ycf}}$ , where $\epsilon \sim \mathcal{N}(0, Id)$
    Concatenate $z_x, z_t, z_{yf}, z_{ycf}$ to form $z_c$
    $\hat{x} \leftarrow D_{\phi_d}(z_c)$
    Calculate $\mathcal{L}_{VAE}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf})$
    $\phi_x \overset{-}{\longleftarrow} \nabla_{\phi_x}\mathcal{L}_{VAE}$; $\phi_t \overset{-}{\longleftarrow} \nabla_{\phi_t}\mathcal{L}_{VAE}$; $\phi_{yf} \overset{-}{\longleftarrow} \nabla_{\phi_{yf}}\mathcal{L}_{VAE}$;
     $\phi_{ycf} \overset{-}{\longleftarrow} \nabla_{\phi_{ycf}}\mathcal{L}_{VAE}$; $\phi_d \overset{-}{\longleftarrow} \nabla_{\phi_d}\mathcal{L}_{VAE}$
    $z_G \sim \mathcal{N}(0, Id)$
    $y_0, y_1 \leftarrow G_{\theta_g}(z_G, z_c)$
    $\hat{y}_0 = ((1-T) * y_f + T * y_0)$
     $\hat{y}_1 = (T * y_f + (1-T) * y_1)$
    $d_{logit} \leftarrow D_{\theta_d}(x, \hat{y}_0, \hat{y}_1)$
    Calculate $\mathcal{L}^D(\theta_d)$
    $\theta_d \overset{-}{\longleftarrow} \nabla_{\theta_d}\mathcal{L}^D(\theta_d)$
    $\hat{y}_f \leftarrow T * y_1 + (1-T) * y_0$
    Compute$\mathcal{L}_S^G(y_f, \hat{y}_f)$
    Concatenate $y_0, y_1$ to form $q_{input}$
    $q_\mu, q_\sigma \leftarrow Q_{\theta_q}(q_{input})$
    Compute $\mathcal{L}_I(G, Q)$ by treating $Q(c|x)$ as factored Gaussian
     using $q_\mu, q_\sigma$ and $z_c$
    Compute $\mathcal{L}^G(\theta_g)$
    $\theta_g \overset{-}{\longleftarrow} \nabla_{\theta_g}\mathcal{L}^G(\theta_g)$
end while
```

factual outcome $\hat{y}_f$. As also done in GANITE, we make sure to include the supervised loss $\mathcal{L}_S^G(y_f, \hat{y}_f)$, which enforces the predicted factual outcome $\hat{y}_f$ to be as close as to the true factual outcome y_f.

$$\mathcal{L}_S^G(y_f, \hat{y}_f) = \frac{1}{n}\sum_{i=1}^{n}\big(y_f(i) - \hat{y}_f(i)\big)^2 \tag{10}$$

The complete loss function of counterfactual GAN is given by

$$V_{CF}(G, D) = V_{GAN}(G, D) + \gamma\mathcal{L}_S^G(y_f, \hat{y}_f)$$

We also employ an additional regularization $\lambda I(z_c; G(z_G, z_c))$ to maximize the mutual information between the learnt concatenated latent code z_c and the generated output by the generator $G(z_G, z_c)$, as in [8]. So, we propose to solve the following minimax game:

$$\min_G \max_D V_{CF\_I}(G, D) = V_{CF}(G, D) + \lambda I(z_c; G(z_G, z_c)) \tag{11}$$

$I(z_c; G(z_G, z_c))$ is harder to solve because of the presence of the posterior $p(z_c|x)$ [8], so we obtain the lower bound of it using an auxiliary distribution $Q(z_c|x)$ to approximate $p(z_c|x)$.

Finally, the optimization function of the counterfactual information-theoretic GAN –*InfoGAN*– incorporating the variational regularization of mutual information and hyperparameter λ is given by:

$$\min_{G,Q} \max_D V_{CF\_infoGAN}(G, D, Q) = V_{CF}(G, D) - \lambda\mathcal{L}_I(G, Q) \tag{12}$$

The counterfactual InfoGAN is used to generate the missing counterfactual outcome y_{cf} to form the quadruple $\{X, t, y_f, y_{cf}\}_{i=1}^N$ and sent to the doubly robust block to estimate the ITE.

4.3 Information-theoretic GAN optimization

The GAN generator G_{θ_g} works to fool the discriminator D_{θ_d}. To get the optimal Discriminator $D^*_{\theta_d}$, we maximize $V_{CF\_infoGAN}$

$$\max_D \mathcal{L}^D(\theta_d) = V_{CF\_infoGAN}(G, D, Q) \tag{13}$$

To get the optimal generator $G^*_{\theta_g}$, we maximize $V_{CF\_infoGAN}$

$$\min_{G,Q} \mathcal{L}^G(\theta_g) = V_{CF\_infoGAN}(G, D, Q) \tag{14}$$

A detailed derivation the information-theoretic loss of the GAN is given in the Appendix A.2.

4.4 Doubly robust ITE estimation

As above introduced, the propensity score $\pi(x)$ represents the probability of receiving a treatment $T = 1$ (over the alternative $T = 0$) conditioned on the pre-treatment covariates X. By combining IPW through $\pi(x)$ with outcome regression by both treatment variable and the covariates, Jonsson defined the doubly robust estimation of causal effect [13] as follows:

$$\hat{\delta}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\bigg[\frac{y_i t_i - (t_i - \pi(x_i))\mu(x_i, t_i)}{\pi(x_i)} - \frac{y_i(1 - t_i) - (t_i - \pi(x_i))\mu(x_i, t_i)}{1 - \pi(x_i)}\bigg] \tag{15}$$

where $\mu(x, t) = \hat{\alpha_0} + \hat{\alpha_1} x_1 + \hat{\alpha_2} x_2 + \cdots + \hat{\alpha_n} x_n + \hat{\delta} t$, and $(t_i - \pi(x_i))\mu(x_i, t_i)$ is used for the IPW estimator.

Algorithm 2 Training of the doubly robust multitask network for ITE estimation

```
Input: Complete dataset $\tilde{X} = \{(x^{(1)}, t^{(1)}, y_f^{(1)}, y_{cf}^{(1)}),..., (x^{(n)}, t^{(n)}, y_f^{(n)}, y_{cf}^{(n)})\}$ after training the GAN module for counterfactual prediction; hyper-parameters $\alpha > 0$; $\beta > 0$; outcome heads with shared parameters $\phi$ and outcome specific parameters $\theta_0, \theta_1$; propensity head with parameters $\theta_\pi$; regressor head with parameters $\theta_\mu$
Initialize parameters: $\theta_0, \theta_1, \theta_\pi, \theta_\mu$
while training do
    $x \leftarrow$ batch of samples from the dataset
    Calculate $\hat{y}_i^{(0)}, \hat{y}_i^{(1)}, \hat{y}_f^{(i)}, \hat{y}_{cf}^{(i)}$
    Calculate the predicted loss $\mathcal{L}_i^p(\theta_1, \theta_0, \phi)$
    Calculate $\hat{y}_{fDR}^{(i)}, \hat{y}_{cfDR}^{(i)}$
    Calculate the Doubly Robust loss $\mathcal{L}_i^{DR}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi)$
    Calculate the final loss $\mathcal{L}_{ITE}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi)$
    Calculate gradients of the loss $\mathcal{L}_{ITE}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi)$
    Update the parameters $\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi$
end while
```

After getting the counterfactual outcome y_{cf} from the counterfactual GAN to form the quadruple $\{X, t, y_f, y_{cf}\}_{i=1}^N$, we pass this as the input to the doubly robust multitask network to estimate the ITE, using the architecture shown in Figure 3. To predict the outcomes $y^{(0)}$ and $y^{(1)}$, we use a configuration similar to TARNet, which contains a number of shared layers, denoted by f_ϕ, parameterized by ϕ, and two outcome specific heads f_{θ_0} and f_{θ_1}, parameterized by θ_0 and θ_1.

To ensure doubly robustness, we introduce two more heads that predict the propensity score $\pi(X) = \mathbb{P}(t = 1|X)$ and the regressor $\mu(X, t)$. These two are calculated using two neural networks, parameterized by θ_π and θ_μ respectively. The factual and counterfactual outcome $y_i^{(0)}$ and $y_i^{(1)}$ of the i^{th} sample are then calculated as:

$$\hat{y}_i^{(0)} = f_{\theta_0}(f_\phi(x_i)) \qquad \text{if } t_i = 0 \tag{16}$$

$$\hat{y}_i^{(1)} = f_{\theta_1}(f_\phi(x_i)) \qquad \text{if } t_i = 1 \tag{17}$$

$$\hat{y}_f^{(i)} = t_i\hat{y}_i^{(1)} + (1 - t_i)\hat{y}_i^{(0)} \tag{18}$$

$$\hat{y}_{cf}^{(i)} = (1 - t_i)\hat{y}_i^{(1)} + t_i\hat{y}_i^{(0)} \tag{19}$$

Next, the predicted loss $\mathcal{L}_i^p(\theta_1, \theta_0, \phi)$ is calculated as:

$$\mathcal{L}_i^p(\theta_1, \theta_0, \phi) = (\hat{y}_f^{(i)} - y_f^{(i)})^2 + (\hat{y}_{cf}^{(i)} - y_{cf}^{(i)})^2 + \alpha\text{BinaryCrossEntropy}(\pi(x_i), t_i) \tag{20}$$

where α is a hyperparameter. With the help of the propensity score $\pi(X)$ and the regressor $\mu(X, t)$, the doubly robust outcomes are calculated as

$$\hat{y}^{(i)}_{fDR} = t_i\left[\frac{t_i\hat{y}^{(1)}_i - (t_i - \pi(x_i)\mu(x_i,t_i))}{\pi(x_i)}\right] + (1-t_i)\left[\frac{(1-t_i)\hat{y}^{(0)}_i - (t_i - \pi(X_i)\mu(x_i,t_i))}{1-\pi(x_i)}\right] \quad (21)$$

$$\hat{y}^{(i)}_{cf_{DR}} = (1-t_i)\left[\frac{(1-t_i)\hat{y}^{(1)}_i - (t_i - \pi(x_i)\mu(x_i,t_i))}{\pi(x_i)}\right] + t_i\left[\frac{t_i\hat{y}^{(0)}_i - (t_i - \pi(x_i)\mu(x_i,t_i))}{1-\pi(x_i)}\right] \quad (22)$$

The doubly robust loss $\mathcal{L}^{DR}_i(\theta_1,\theta_0,\theta_\phi,\theta_\mu,\phi)$ is calculated as:

$$\mathcal{L}^{DR}_i(\theta_1,\theta_0,\theta_\pi,\theta_\mu,\phi) = (\hat{y}^{(i)}_{fDR} - y^{(i)}_f)^2 + (\hat{y}^{(i)}_{cf_{DR}} - y^{(i)}_{cf})^2 \quad (23)$$

Finally, the loss function of the ITE is:

$$\mathcal{L}^{ITE}(\theta_1,\theta_0,\theta_\pi,\theta_\mu,\phi) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}^p_i + \beta\mathcal{L}^{DR}_i\right) \quad (24)$$

where β is a hyperparameter and the whole network is trained using end-to-end strategy .

4.5 Differences with CEVAE and GANITE

The counterfactual outcome predictor of DR-VIDAL uses together VAE and GAN, used in CEVAE and GANITE, respectively. CEVAE also incorporates a causal graph, but it is more simplistic than DR-VIDAL, as it infers only the observed proxy X from Z. We instead consider multiple latent variables causally related to the treatment and the outcome in addition to the direct link to the pre-treatment covariates. Upon that, we use GAN to generate counterfactual examples, but, unlike GANITE, we first disentangle the multiple latent factors using a VAE, then we optimize the GAN with the mutual information, finally generating the entire potential outcome vector.

4.6 Differences with TARNet and Dragonnet

The design of the doubly robust module block of DR-VIDAL is closely related to that of TARNet and Dragonnet (Figure 3. However, TARNet uses a two-headed network, which is not doubly robust. Dragonnet includes a third head that incorporates the propensity score. DR-VIDAL exploits the doubly robustness adding two heads, i.e., the propensity score and the regressor head, to the the basic two-headed TARNet configuration. Further, in TARNet the weights corresponding to each of samples are calculated as the crude probability of the treatment assignment, whereas DR-VIDAL accounts for the pre-treatment covariates. For Dragonnet, the targeted regularization is implemented without taking into account the regressed outcome, which instead is estimated by DR-VIDAL in the fourth head, as a function of treatment and pre-treatment covariates. Another major difference between TARNet/Dragonnet and DR-VIDAL is the training strategy. For both TARNet and Dragonnet, the counterfactual outcome does not exist, so for each sample the overall loss function has to be estimated with the factual outcome only, updating the parameters of the outcome head of the factual outcome during training. In DR-VIDAL, we have the entire potential outcome vector, comprising both the factual and the counterfactual outcomes. So, for each training sample, the loss function is calculated for both outcomes, and the corresponding parameters of both the outcome heads are updated, as shown in Algorithm 2.

	IHDP	Jobs	Twins
	$\sqrt{\epsilon^{out-of-s}_{PEHE}}$	$R^{out-of-s}_{Pol}$	$\sqrt{\epsilon^{out-of-s}_{PEHE}}$
DR-VIDAL	**0.62 ± 0.06**	**0.102 ± 0.01**	**0.318 ± 0.008**
DR-VIDAL (w/o DR loss)	0.85 ± 0.06	0.110 ± 0.01	0.324 ± 0.007
DR-VIDAL (w/o Info loss)	0.67 ± 0.04	0.109 ± 0.01	0.318 ± 0.012
DR-VIDAL (w/o DR + Info loss)	0.81 ± 0.05	0.113 ± 0.01	0.326 ± 0.008

Table 1: Performance of the all the different DR-VIDAL configurations on the IHDP, Jobs and Twins datasets (1000, 10, and 100 realizations, respectively). Results show the out-of-sample (mean ± st.dev) error (PEHE) and policy risk (R_{Pol}).

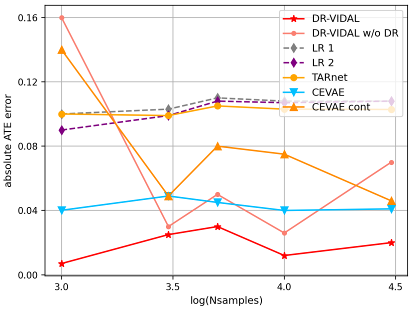

Figure 4: Comparison of the performance (ATE) of DR-VIDAL vs. all other models on samples from the generative process of CEVAE, defined in equation 25.

5 EXPERIMENTAL SETUP

5.1 Datasets

Synthetic datasets. We conduct performance tests on two synthetic data experiments. The first uses the same data generation process devised for CEVAE [23]. We generate a marginal distribution x as a mixture of Gaussians from the 5-dimensional latent variable z, indicating each mixture component, specifically:

$$\begin{aligned} z_i &\sim Bern(0.5) \\ x_i|z_i &\sim \mathcal{N}(z_i, \sigma_5^2 z_i + \sigma_3^2(1-z_i)) \\ t_i|z_i &\sim Bern(0.75z_i + 0.25(1-z_i)) \\ y_i|t_i,z_i &\sim Bern(Sigmoid(3(z_i + 2(2t_i - 1)))) \end{aligned} \quad (25)$$

Datasets of sample size N ∈ {1000, 3000, 5000, 10000, 30000} are generated, and divided into 80-20 % train-test split.

In the second experimental setting, we amalgamate the synthetic data generation process by CEVAE with that of GANITE [36], to model the more complex causal structure illustrated in Figure 1. We sample 7-, 1-, 1-, and 1-dimensional vectors for z_x, z_t, z_{yf}, and z_{ycf} from Bernoulli distributions, and then collate them into x, i.e.,

$$
\begin{aligned}
z_x &\sim Bern(0.5); \qquad z_t \sim Bern(0.5)\\
z_{yf} &\sim Bern(0.5); \qquad z_{ycf} \sim Bern(0.5)\\
x_x|z_x &\sim \mathcal{N}(z_x, 5(z_x) + 3(1 - z_x))\\
x_t|z_t &\sim \mathcal{N}(z_x, 2(z_t) + 0.5(1 - z_t))\\
x_{yf}|z_{yf} &\sim \mathcal{N}(z_{yf}, 10(z_{yf}) + 6(1 - z_{yf}))\\
x_{ycf}|z_{ycf} &\sim \mathcal{N}(z_{ycf}, 10(z_{ycf}) + 6(1 - z_{ycf}))\\
w_t^T &\sim \mathcal{U}((-0.1, 0.1)^{10x1})\\
n_t &\sim \mathcal{N}(0, 0.1)\\
t|x &\sim Bern(Sigmoid(w_t^T x + n_t))\\
w_y^T &\sim \mathcal{U}((-1, 1)^{10x2}); \qquad n_y \sim \mathcal{N}(0^{2x1}, 0.1x\mathcal{I}^{2x2})\\
y|x &\sim w_y^T x + n_y \qquad (26)
\end{aligned}
$$

From the covariates x, we simulate the treatment assignment t and the potential outcomes y as described in the GANITE paper. We generate multiple synthetic datasets for sample sizes N ∈ {1000, 3000, 5000, 10000, 30000}, also divided into 80-20 % splits.

Real-world datasets. We use three popular real-world benchmark datasets: the Infant Health and Development Program (IHDP) dataset [17], the Twins dataset [3], and the Jobs dataset [22]. The first two are semi-synthetic, and simulated counterfactuals to the real factual data are available. These datasets have been also designed and collated to meet specific treatment overlap condition, nonparallel treatment assignment, and nonlinear outcome surfaces [17, 23, 31, 36]. The IHDP datasets is composed by 110 treated subjects and 487 controls, with 25 covariates. The Twins dataset comprises 4553 treated, 4567 controls, with 30 covariates. The Jobs dataset comprises 237 treated, 2333 controls, with 17 covariates. For all the real-world datasets, we use the same experimental settings described in GANITE, where the datasets are divided into 56/24/20 % train-validation-test splits. We run 1000, 10 and 100 realizations of IHDP, Jobs and Twins datasets, respectively.

5.2 Performance metrics

Consistently with prior studies [17, 31, 36], we report the error on the ATE ϵ_{ATE}, and the expected Precision in Estimation of Heterogeneous Effect (PEHE), ϵ_{PEHE}, for IHDP and Twins datasets, since factual and the counterfactual outcomes are available. For the Jobs dataset, as the counterfactual outcome does not exist, we report the policy risk $R_{pol}(\pi)$, and the error on the average treatment effect on the treated (ATT) ϵ_{ATT}, as indicated in [31, 36]. All the equations are included in the Appendix 29.

5.3 Training and implementation of DR-VIDAL

Adversarial module. To reduce the model complexity and parameters for the encoder of the VAE, we have a shared neural network connected to 4 other networks for estimating the four posterior distributions $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$. The shared neural network has 3 layers, each with 15 nodes. The networks with $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$ as outputs have a single layer with 5, 1, 1, 1 nodes, respectively. The decoder is a 4-layer neural network, each with 15 nodes to calculate the data likelihood $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})$. For the GAN, the generator network has 2 shared layers and 2 outcome-specific layers, each with 100 nodes. The discriminator and the network for information maximization (Q network in Figure 2) is a 3-layered neural network, each with 30 nodes and 8 nodes respectively. All the layers of the VAE and GAN use Rectified Linear Unit (ReLU) activation functions and the parameters are updated using the Adam optimizer [20]. The random noise z_G is sampled from a 92-dimensional standardized Gaussian distribution $\mathcal{N}(0, 1)$. The hyperparameter γ is set as 1 for all datasets, while λ is set as 0.2, 0.01 and 10 for IHDP, Jobs and Twins, respectively. The batch sizes of IHDP, Jobs, and Twins are 64, 64, and 256, respectively. The learning rates of the VAE, generator and discriminator are 1e-3, 1e-4, and 5e-4, respectively.

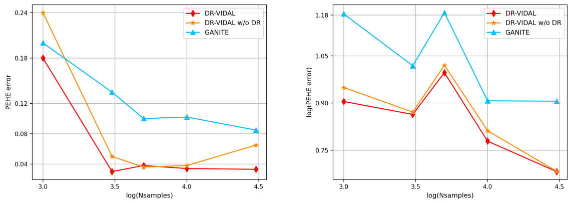

Figure 5: Comparison of the performance (PEHE) of GANITE vs. DR-VIDAL, with the doubly robust (DR) or without the doubly robust (w/o DR) feature, to predict the absolute error for PEHE on samples from the generative process of CEVAE (left) and DR-VIDAL (right) defined in equations 25 and 26 respectively.

Doubly robust module. For the doubly robust module, the shared network f_ϕ and outcome specific networks f_{θ_0} and f_{θ_1} are both 3-layer neural network, each with 200 and 100 nodes. The propensity network π has 2 layers each with 200 nodes. The regressor network μ has 6 layers with 200 nodes and 100 nodes in the first and last 3 layers. All the layers of the VAE and GAN use ReLU activation and the Adam optimizer. The batch sizes are the same as for the adversarial module. We set the learning rate of all the networks as 1e-4 and the hyperparameters α and β are set at 1 for all 3 datasets.

Implementation and availability. DR-VIDAL's code is written in Pytorch (https://pytorch.org/) and is available under the MIT license at: https://bitbucket.org/goingdeep2406/dr-vidal/src/master/.

5.4 Comparison with other methods.

On all three datasets, DR-VIDAL is compared against TARNet, CEVAE, and GANITE. In addition, we show performance results for: least squares regression using treatment as a covariate (OLS/LR1); separate least squares regressions for each treatment (OLS/LR2); balancing linear regression (BLR) [19]; k-nearest neighbor (k-NN) [10]; Bayesian additive regression trees (BART) [9]; random forest (R Forest) [7]; causal forest (C Forest) [35]; balancing neural network (BNN) [19]; and counterfactual regression with Wasserstein distance (CFR$_{WASS}$) [31].

	IHDP				Twins			
Methods	$\sqrt{\epsilon_{PEHE}^{within-s}}$	$\epsilon_{ATE}^{within-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$	$\sqrt{\epsilon_{PEHE}^{within-s}}$	$\epsilon_{ATE}^{within-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$
OLS/LR1	5.8 ± 0.3	0.73 ± 0.04	5.8 ± 0.3	0.94 ± 0.06	0.319 ± 0.005	0.0038 ± 0.0025	0.297 ± 0.016	0.0069 ± 0.0056
OLS/LR2	2.4 ± 0.1	0.14 ± 0.01	2.5 ± 0.1	0.31 ± 0.02	0.320 ± 0.001	0.0039 ± 0.0025	0.318 ± 0.007	0.0070 ± 0.0059
BLR	5.8 ± 0.3	0.72 ± 0.04	5.8 ± 0.3	0.93 ± 0.05	0.312 ± 0.002	0.0057 ± 0.0036	0.320 ± 0.003	0.0334 ± 0.0092
k-NN	2.1 ± 0.1	0.14 ± 0.01	4.1 ± 0.2	0.79 ± 0.05	0.333 ± 0.003	0.0028 ± 0.0021	0.323 ± 0.018	0.0051 ± 0.0039
BART	2.1 ± 0.2	0.23 ± 0.01	2.3 ± 0.1	0.34 ± 0.02	0.347 ± 0.009	0.1206 ± 0.0236	0.338 ± 0.016	0.1265 ± 0.0234
R Forest	4.2 ± 0.2	0.73 ± 0.05	6.6 ± 0.3	0.96 ± 0.06	0.306 ± 0.002	0.0049 ± 0.0034	0.321 ± 0.005	0.0080 ± 0.0051
C Forest	3.8 ± 0.2	0.18 ± 0.01	3.8 ± 0.2	0.40 ± 0.03	0.366 ± 0.003	0.0286 ± 0.0035	0.316 ± 0.011	0.0335 ± 0.0083
BNN	2.2 ± 0.1	0.37 ± 0.03	2.1 ± 0.1	0.42 ± 0.03	0.325 ± 0.003	0.0056 ± 0.0032	0.321 ± 0.018	0.0203 ± 0.0071
TARNET	0.88 ± 0.02	0.26 ± 0.01	0.95 ± 0.02	0.28 ± 0.01	0.317 ± 0.005	0.0108 ± 0.0017	0.315 ± 0.003	0.0151 ± 0.0018
CFR$_{WASS}$	0.71 ± 0.0	0.25 ± 0.01	0.76 ± 0.0	0.27 ± 0.01	0.315 ± 0.007	0.0112 ± 0.0016	0.313 ± 0.008	0.0284 ± 0.0032
GANITE	1.9 ± 0.4	0.43 ± 0.05	2.4 ± 0.4	0.49 ± 0.05	0.289 ± 0.12	0.0058 ± 0.0017	0.297 ± 0.05	0.0089 ± 0.0075
CEVAE	2.7 ± 0.1	0.34 ± 0.01	2.6 ± 0.1	0.46 ± 0.02	n.r	n.r	n.r	n.r
DR-VIDAL	**0.69 ± 0.05**	**0.57 ± 0.07**	**0.69 ± 0.06**	**0.57 ± 0.08**	**0.317 ± 0.002**	**0.0102 ± 0.0128**	**0.318 ± 0.008**	**0.0111 ± 0.0137**

Table 2: Performance on the within-sample and out-of-sample test sets (mean ± st.dev) of various models on the IHDP and Twins datasets.

Methods	$R_{Pol}^{within-s}$	$\epsilon_{ATT}^{within-s}$	$R_{Pol}^{out-of-s}$	$\epsilon_{ATT}^{out-of-s}$
OLS/LR1	0.22 ± 0.0	0.01 ± 0.00	0.23 ± 0.0	0.08 ± 0.04
OLS/LR2	0.21 ± 0.0	0.01 ± 0.01	0.24 ± 0.0	0.08 ± 0.03
BLR	0.22 ± 0.0	0.01 ± 0.01	0.25 ± 0.0	0.08 ± 0.03
k-NN	0.02 ± 0.0	0.21 ± 0.01	0.26 ± 0.0	0.13 ± 0.05
BART	0.23 ± 0.0	0.02 ± 0.00	0.25 ± 0.0	0.08 ± 0.03
R Forest	0.23 ± 0.0	0.03 ± 0.01	0.28 ± 0.0	0.09 ± 0.04
C Forest	0.19 ± 0.0	0.03 ± 0.01	0.20 ± 0.0	0.07 ± 0.03
BNN	0.20 ± 0.0	0.03 ± 0.01	0.24 ± 0.0	0.09 ± 0.04
TARNET	0.17 ± 0.0	0.05 ± 0.02	0.21 ± 0.0	0.11 ± 0.04
CFR$_{WASS}$	0.17 ± 0.0	0.04 ± 0.01	0.21 ± 0.0	0.09 ± 0.03
GANITE	0.13 ± 0.01	0.01 ± 0.01	0.14 ± 0.01	0.06 ± 0.03
CEVAE	0.15 ± 0.0	0.02 ± 0.01	0.26 ± 0.0	0.03 ± 0.01
DR-VIDAL	**0.09 ± 0.005**	**0.04 ± 0.03**	**0.10 ± 0.01**	**0.05 ± 0.02**

Table 3: Performance on the within-sample and out-of-sample test sets (mean ± st.dev) of various models on the Jobs dataset.

6 RESULTS

6.1 Synthetic datasets

In the first synthetic dataset, which uses the generative assumptions of CEVAE defined in equation (25), the doubly robust version of DR-VIDAL demonstrates lower ATE error at all sample sizes with respect to all models, as shown in Figure 4. When comparing PEHE, DR-VIDAL (both with and without the doubly robust feature) largely outperforms GANITE, as displayed in Figure 5, left panel. In the second synthetic dataset, generated under the more complex assumptions according to equation (26), DR-VIDAL, both with and without the doubly robust feature, outperforms GANITE in terms of PEHE, as shown in Figure 5 (right panel). It is worth noting the potential of DR-VIDAL to better disentangle hidden representations in comparison to GANITE irrespective of the presence of the doubly robust module.

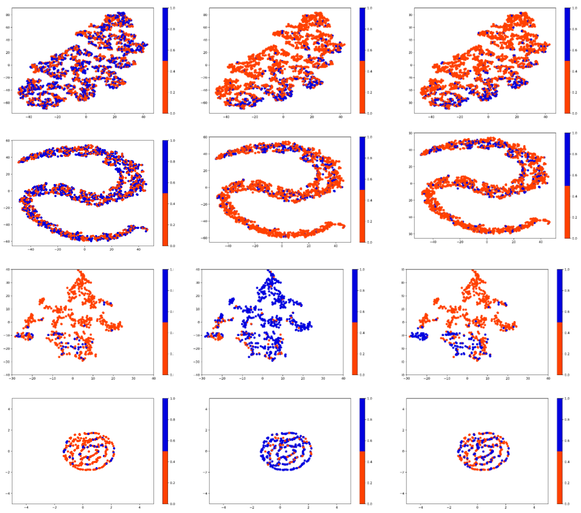

Figure 6: Visualization of the latent representation learned by the adversarial module of DR-VIDAL for the Twins and Jobs dataset using t-distributed stochastic neighbor embedding (t-SNE). The 1st and 2nd panels show the t-SNE before and after training the network for Twins dataset. The 3rd and 4th panels show the same for Jobs dataset. From left to right, the plots show the t-SNE of treatment, factual and counterfactual outcomes.

6.2 Real world datasets

The performance of the all the different DR-VIDAL configurations – with/without information-theoretic optimization and with/without

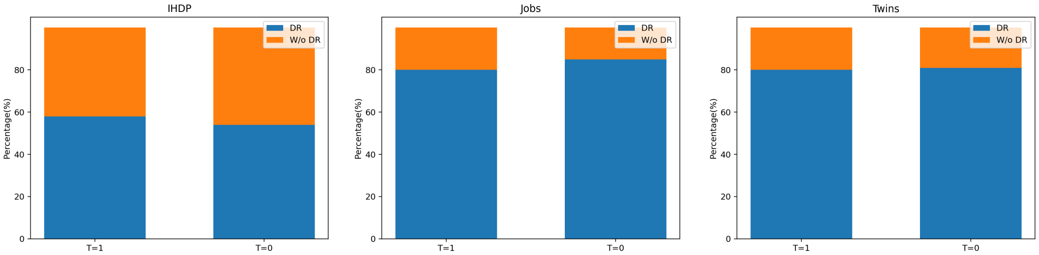

Figure 7: Performance comparison of doubly robust vs. non-doubly robust version of DR-VIDAL. The bar plots show how many times one model setup is better than the other in terms of error on the factual outcome (y_f). Panels, from left to right, show results on IHDP, Jobs and Twins datasets (100, 10, 100 iterations), respectively.

doubly robust loss– on the IHDP, Jobs and Twins datasets (out-of-sample) are shown in Table 1. In all three datasets, across all realizations, the information-theoretic, doubly robust configuration yields the best results. The doubly robust loss seems to be responsible for most of the improvement. The absolute gain is small, in the order of 1%, but the relative gain with respect to the non-doubly robust setup is significant, as shown in Figure 7, where the doubly robust module always outperforms its non-doubly robust version (from 55-60% in IHDP to over 80% in Twins and Jobs datasets).

Tables 2 and 3 show the comparison with CEVAE, GANITE, TARNet and other methods on the three datasets (in-sample and out-of-sample). DR-VIDAL outperformed the other methods on all datasets. On the IHDP and Jobs dataset, DR-VIDAL was the best over all by a larger margin. Instead, performance increment in the Twins dataset was mild. Even if DR-VIDAL has a large number of parameters, the disentanglement of hidden factors and the adversarial training make it appropriate for datasets with relatively small sample size like IHDP. It is worth noting that DR-VIDAL converges much faster than other causal generative models CEVAE and GANITE which is possibly due to the incorporation of doubly robustness in DR-VIDAL.

The t-distributed stochastic neighbor embedding (t-SNE) of representations learned by the VAE of the adversarial module of DR-VIDAL for Twins and Jobs datasets –before and after training– are shown in Figures 6 and **??**. For all datasets, the t-SNE shows reorganization and cluster tightness (i.e., the data reside on a smaller space) on the treatment, factual and counterfactual outcomes spaces.

7 DISCUSSION

DR-VIDAL is an original deep learning approach to causal effect estimation and counterfactual prediction that combines disentangled adversarial representation learning, information-theoretic optimization, and doubly robust regression. Our approach fuses several key properties of existing methods with a distinctive causal structure and robust regression design. On all benchmark datasets, DR-VIDAL outperforms other tools, and both the doubly robust property and information-theoretic optimization improve performance over the basic disentangled adversarial setup.

This work has some limitations. First, the causal graph, even if more elaborated than CEVAE, is still relatively basic, with straigthforward confounding and a unique adjustment set. For instance, a slight modification to the causal graph that connects the Z to X and only to their respective treatment, factual and counterfactual outcome nodes would already imply two adjustments set. Another limitation is that the encoded representation in the VAE does not employ any attention mechanism, that could help identifying the most important covariates for the propensity scores, especially with of high-dimensional and noisy datasets. Finally, one thing that would be worth evaluating is how Dragonnet would perform as a downstream module for DR-VIDAL, substituting it to our four-head dobly robust block (as they have same input and architecture, but Dragonnet has three heads instead of four).

In conclusion, DR-VIDAL framework is a comprehensive approach to predict counterfactuals and estimate ITE, and its flexibility (modifiable causal structure and modularity) allows for further expansion and improvement.

REFERENCES

[1] Ahmed Alaa and Mihaela van der Schaar. 2018. Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 129–138. http://proceedings.mlr.press/v80/alaa18a.html

[2] Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. 2017. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966* (2017).

[3] Douglas Almond, Kenneth Y Chay, and David S Lee. 2005. The costs of low birth weight. *The Quarterly Journal of Economics* 120, 3 (2005), 1031–1083.

[4] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

[5] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46, 3 (2011), 399–424.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association* 112, 518 (2017), 859–877.

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

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems* 29 (2016), 2172–2180.

[9] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.

[10] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2008. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* 90, 3 (2008), 389–405.

[11] Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* 84, 1 (2002),

151–161.

[12] Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. 2014. Doubly robust policy evaluation and optimization. *Statist. Sci.* 29, 4 (2014), 485–511.

[13] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *American journal of epidemiology* 173, 7 (2011), 761–767.

[14] MM Garrido et al. 2014. Methods for Constructing and Assessing Propensity Scores. *Health Services Research* 49, 5 (2014), 1701-–20.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014), 2672–2680.

[16] M.A. Hernan and J.M. Robins. 2019. *Causal Inference.* Taylor & Francis. https://books.google.com/books?id=_KnHIAAACAAJ

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

[18] Guido W Imbens. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 3 (2000), 706–710.

[19] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning.* 3020–3029.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

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

[23] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in neural information processing systems.* 6446–6456.

[24] Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. 2018. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics* 27, 1 (2018), 209–219.

[25] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 23, 19 (2004), 2937–2960.

[26] J. Pearl, M. Glymour, and N.P. Jewell. 2016. *Causal Inference in Statistics: A Primer.* Wiley. https://books.google.com/books?id=L3G-CgAAQBAJ

[27] Kristin E Porter, Susan Gruber, Mark J Van Der Laan, and Jasjeet S Sekhon. 2011. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics* 7, 1 (2011).

[28] Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2 (2020), 369–375. https://doi.org/10.1038/s42256-020-0197-y

[29] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (04 1983), 41–55. https://doi.org/10.1093/biomet/70.1.41

[30] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688–701.

[31] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 3076–3085. http://proceedings.mlr.press/v70/shalit17a.html

[32] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. In *Advances in neural information processing systems.* 2507–2517.

[33] Bonnie Sibbald and Martin Roland. 1998. Understanding controlled trials: Why are randomised controlled trials important? *BMJ* 316, 7126 (1998), 201. https://doi.org/10.1136/bmj.316.7126.201

[34] Yuxi Tian, Martijn J Schuemie, and Marc A Suchard. 2018. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International journal of epidemiology* 47, 6 (2018), 2005–2014.

[35] 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.

[36] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. 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. https://openreview.net/forum?id=ByKWUeWA-

A APPENDIX

A.1 Derivation of the Loss function ELBO loss of VAE

From Figure 1, $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{y_cf})$ and $p_{\phi_d}(z_x, z_t, z_{yf}, z_{y_cf}|x)$ are the true likelihood and true posterior respectively. The posterior is hard to evaluate, so we have to approximate the true posterior to the product of the factorized known distributions $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$ and $q_{\phi_{ycf}}(z_{ycf}|x)$ by minimising the KL divergence as follows,

$$
\begin{aligned}
&KL\big(q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)||\\
&\quad p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf}|x)\big)\\
&= \int\int\int\int q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad \big[\log \frac{q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)}{p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf}|x)}\big] dz_x dz_t dz_{yf} dz_{ycf}\\
&= \int\int\int\int q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad \big[\log q_{\phi_x}(z_x|x) + \log q_{\phi_t}(z_t|x)\\
&\quad + \log q_{\phi_{yf}}(z_{yf}|x) + \log q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad - \log p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf}|x)\big] dz_x dz_t dz_{yf} dz_{ycf}\\
&= \int\int\int\int q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad \big[\log q_{\phi_x}(z_x|x) + \log q_{\phi_t}(z_t|x)\\
&\quad + \log q_{\phi_{yf}}(z_{yf}|x) + \log q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad - \log p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf}) - \log p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf})\\
&\quad + \log p_{\phi_d}(x)\big] dz_x dz_t dz_{yf} dz_{ycf}\\
&= \int q_{\phi_x}(z_x|x) \log \frac{q_{\phi_x}(z_x|x)}{p_{\phi_d}(z_x)} dz_x\\
&\quad + \int q_{\phi_t}(z_t|x) \log \frac{q_{\phi_t}(z_t|x)}{p_{\phi_d}(z_t)} dz_t\\
&\quad + \int q_{\phi_{yf}}(z_{yf}|x) \log \frac{q_{\phi_{yf}}(z_{yf}|x)}{p_{\phi_d}(z_{yf})} dz_{yf}\\
&\quad + \int q_{\phi_{ycf}}(z_{ycf}|x) \log \frac{q_{\phi_x}(z_{ycf}|x)}{p_{\phi_d}(z_x)} dz_{ycf}\\
&\quad - \int\int\int\int \big[q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)\\
&\quad q_{\phi_{ycf}}(z_{ycf}|x) \log p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})\big] dz_x dz_t dz_{yf} dz_{ycf}\\
&\quad + \log p_{\phi_d}(x)\\
&= KL\big(q_{\phi_x}(z_x|x)||p_{\phi_d}(z_x))\big) + KL\big(q_{\phi_t}(z_t|x)||p_{\phi_d}(z_t))\big)\\
&\quad + KL\big(q_{\phi_{yf}}(z_{yf}|x)||p_{\phi_d}(z_{yf}))\big)\\
&\quad + KL\big(q_{\phi_{ycf}}(z_{ycf}|x)||p_{\phi_d}(z_{ycf}))\big)\\
&\quad - \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}}[\log p(x|z_x, z_t, z_{yf}, z_{ycf})]\\
&\quad + \log p_{\phi_d}(x)
\end{aligned}
$$

where, the distributions $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$ and $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf}|x)$ are parameterized by the parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}, \phi_d$. The KL divergence of two distributions is always greater than or equal to zero. So,

$$
\begin{aligned}
& KL\big(q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)|| \\
& \qquad p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf}|x)\big) \geq 0, \\
& \log p_{\phi_d}(x) \geq \mathcal{L}_{ELBO} \qquad \text{where,} \\
& \mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf}) \\
& = \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}}[\log p(x|z_x, z_t, z_{yf}, z_{ycf})] \\
& \quad - KL\big(q_{\phi_x}(z_x|x)||p_{\phi_d}(z_x))\big) - KL\big(q_{\phi_t}(z_t|x)||p_{\phi_d}(z_t))\big) \\
& \quad - KL\big(q_{\phi_{yf}}(z_{yf}|x)||p_{\phi_d}(z_{yf}))\big) \\
& \quad - KL\big(q_{\phi_{ycf}}(z_{ycf}|x)||p_{\phi_d}(z_{ycf}))\big)
\end{aligned}
$$

A.2 Variational information maximization

$$
\begin{aligned}
& I(z_c; G(z_G, z_c)) = H(z_c) - H(z_c|G(z_G, z_c)) \\
& = H(z_c) + \int\int p(Z_c = z'_c, X = G(z_G, z_c)) \\
& \qquad \log p(Z_c = z'_c|X = G(z_G, z_c))\, dz_c dx \\
& = H(z_c) + \mathbb{E}_{x\sim G(z_G, z_c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log(p(z'_c|x)) \\
& = H(z_c) + \mathbb{E}_{x\sim G(z_G, z_c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[\frac{p(z'_c|x)}{Q(z'_c|x)}Q(z'_c|x)\right] \\
& = H(z_c) + \mathbb{E}_{x\sim G(z_G, z_c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[\frac{p(z'_c|x)}{Q(z'_c|x)}\right] \\
& \qquad + \mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[Q(z'_c|x)\right] \\
& = H(z_c) + \mathbb{E}_{x\sim G(z,c)} \int p(z'_c|x) \log\frac{p(z'_c|x)}{Q(z'_c|x)} dc' \\
& \qquad + \mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{c'\sim p(z_c|x)} \log\left[Q(z'_c|x)\right] \\
& = H(z_c) + \mathbb{E}_{x\sim G(z,c)}\left[KL\big(p(z'_c|x||Q(z'_c|x))\big)\right] \\
& \qquad + \mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[Q(z'_c|x)\right] \\
& \geq H(z_c) + \mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[Q(z'_c|x)\right] \\
& \geq H(z_c) + \mathbb{E}_{z_c\sim p(z_c)}\mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[Q(z'_c|x)\right] \\
& \geq H(z_c) + \mathbb{E}_{z_c\sim p(z_c)}\mathbb{E}_{x\sim G(z,c)} \log\left[Q(z_c|x)\right] \\
& \big(\text{by Lemma 5.1 of [8]}\big) \\
& = L_I(G, Q)
\end{aligned}
$$

A.3 Performance metric

$$
\begin{aligned}
\epsilon_{PEHE} = \frac{1}{N}\sum_{n=0}^{N}\Big(\mathbb{E}_{y_j(n)\sim\mu_j(n)}\big[y_1(n) - y_0(n)\big] \\
- \big[\hat{y_1}(n) - \hat{y_0}(n)\big]\Big)^2
\end{aligned}
\tag{27}
$$

$$
\epsilon_{ATE} = ||\frac{1}{N}\sum_{n=0}^{N}\mathbb{E}_{y(n)\sim\mu(n)}[y(n)] - \frac{1}{N}\sum_{n=0}^{N}\hat{y}(n)||_2^2 \tag{28}
$$

$$
\begin{aligned}
R_{pol}(\pi) = \frac{1}{N}\sum_{n=0}^{N}\Big[1 - \Big(\sum_{i=1}^{k}\Big[\frac{1}{|\Pi_i \cap T_i \cap E|} \\
\sum_{x(n)\in\Pi_i\cap\, T_i\cap E} y_i(n) \times \frac{|\Pi_n \cap E|}{|E|}\Big]\Big)\Big]
\end{aligned}
\tag{29}
$$

where $\pi_i = \{\boldsymbol{x(n)} : i = \arg\max \hat{\boldsymbol{y}}$, $T_i = \boldsymbol{x(n)} : t_i(n) = 1\}$, and E is the randomized sample.

The true average treatment effect on the treated (ATT) and its error ϵ_{ATT} are defined as follows:

$$
ATT = \frac{1}{|T_1 \cap E|}\sum_{x_i\in T_1\cap E} Y_1(x_i) - \frac{1}{|T_0 \cap E|}\sum_{x_i\in C\cap E} Y_0(x_i) \tag{30}
$$

$$
\epsilon_{ATT} = \Big|ATT - \frac{1}{|T_1 \cap E|}\sum_{x_i\in T_1\cap E} \hat{Y}_1(x_i) - \hat{Y}_0(x_i)\Big| \tag{31}
$$

where T_1, T_0 and E are the subsets corresponding to treated, controlled samples, and randomized controlled trials, respectively.