ADVERSARIALLY BALANCED REPRESENTATION FOR CONTINUOUS TREATMENT EFFECT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the individual treatment effect (ITE) requires covariate balance among different treatment groups, and machine learning models have shown great promise in learning a balanced representation of covariates. In contrast with binary treatments for which learning such a representation has been widely studied, the more practical yet complicated continuous treatment setting has remained relatively under-explored. Adopting an information-theoretic approach, we introduce a novel mutual information (MI)-based objective for continuous treatment effect estimation. Leveraging variational approximation to optimize MI terms in our objective, we propose a method called Adversarial CounterFactual Regression (ACFR). ACFR aligns the representation of covariates through an adversarial game and predicts the potential outcomes using a contribution-constraining hypothesis network. Comparison of ACFR against state-of-the-art methods on semi-synthetic datasets demonstrates its superiority in individual-level metrics.

1 INTRODUCTION

Conducting Randomized Control Trial (RCT), the unconfounded source for computing the treatment effect is expensive, time-consuming, and in many cases infeasible or unethical (Pearl, 2009) (Yao et al., 2021). Alternatively, observational datasets in which the treatment assignment mechanism is unknown, have been extensively used to estimate the causal treatment effect (Robins et al., 1994) (Chipman et al., 2010) (Künzel et al., 2017). An observational dataset records units with their covariates(X), their assigned treatment (T), and their outcome after intervening (Y). Under the strong ignorability assumption (Rosenbaum & Rubin, 1983), the treatment effect is identifiable from an observational dataset (Imbens & Wooldridge, 2009). Nonetheless, two main challenges are associated with using these datasets. 1) In real-world observational datasets, we only observe the outcome of a treatment that the unit received (factual outcome), and hence it is impossible to access the ground-truth treatment effect on units. 2) There may exist some variables known as confounders that affect both treatment assignment and outcome. The existence of confounders leads to selection bias among units ($p(T) \neq p(T|X)$), and implies covariate shift among treatment groups ($P(X) \neq P(X|T)$). Therefore, a model trained for predicting factual outcomes will not be as accurate in the prediction of counterfactuals(Yao et al., 2021).

In recent years, many treatment effect estimation methods have been developed based on the notion of deep representation learning (Bengio et al., 2013). Methods proposed for individual treatment effect (ITE) estimation are mainly grounded in learning a balanced representation of the covariates, i.e. a representation in which the covariate shift is reduced while the expressive power for predicting the factual outcomes is preserved. From the theoretical viewpoint, the authors of Shalit et al. (2017) provided a theorem that in the binary treatment setting ($T \in \{0, 1\}$), the counterfactual outcome prediction error is bounded by the sum of the factual outcome prediction error and the integral probability metric (IPM) distance of samples having received different treatments. Various ITE estimation methods have been developed based on this theorem, resulting in models that align the distributions through IPM distance in the representation space and predict the outcome from the representation using hypothesis heads (one head for each treatment option)(Yao et al., 2018) (Hassanpour & Greiner, 2019a) (Hassanpour & Greiner, 2019b). We also focus on the individual effect of treatments, but unlike the majority of the previous works, we consider continuous-valued treatments (e.g. different dosages of a medication). Although it is a more general and practical setup, extending the above framework from binary to continuous treatments is challenging. This is due to

the infinite number of treatment options in the continuous setting, which makes it hard to extract a balanced representation and also to maintain the influence of the treatment value on the predicted outcome.

To be more specific, the covariate shift in the binary case is commonly decreased by minimizing a distributional distance from the IPM class between the conditional distributions $p(Z|T = 0)$ and $p(Z|T = 1)$. In the continuous case, recently Bellot et al. (2022) presented an upper bound on the counterfactual error which consists of the factual error term plus an IPM distance term between $p(Z, T)$ and $p(Z)p(T)$. Marginal distributions being unknown, the IPM term is approximated as in Johansson et al. (2016), by computing the distance between a factual pair (z_i, t_i) and treatment-permutated pairs $(z_i, t_j)_{j\neq i}$. However, the impact of this approximation is negligible in the entire objective since the only difference between pairs is the scalar treatment value while the higher-dimensional representation is the same (Hassanpour & Greiner, 2019a). Moreover, it is unclear how the hypothesis model should consider the continuous treatment value when having separate hypothesis heads as in the binary setting is not possible. Regarding the treatment as input without any adjustment also creates the risk of overfitting the representation. Methods such as Varying Coefficient Network (VCNet) (Nie et al., 2021) and Dose-Response Network (DRNet) (Schwab et al., 2020) included the treatment value into the architecture of the hypothesis model to prevent overfitting. However, this increases the number of parameters of the hypothesis network significantly, and the best choice of how to involve the treatment is problem-dependent.

This paper investigates the problem of continuous individual treatment effect estimation[1]. Inspired by the application of information theory to representation learning (Alemi et al., 2016)(Tishby & Zaslavsky, 2015), we formulate the two goals of the balanced representation learning approach with mutual information (MI) terms. Specifically, we propose to learn a latent representation Z of the covariates X defined by a parametric encoder $p_\phi(z|x)$ that has the following two properties: 1) We want to remove relevant information about the treatment T from the latent representation Z, i.e. we aim to minimize the mutual information [2] between T and Z, $I(T, Z; \phi)$. 2) We want Z to contain the necessary information about the outcome such that given the treatment T we can predict Y accurately, i.e. we aim to maximize the mutual information of Z and Y given T, $I(Z, Y|T; \phi)$. We propose a neural network model called Adversarial CounterFactual Regression (ACFR) that optimizes the two terms jointly. ACFR consists of three sub-networks: the encoder $p_\phi(z|x)$ that extracts the representation given the covariates, a treatment-predictor network that regresses the treatment given the representation, and a hypothesis network that regresses the outcome given the representation and treatment. Our model addresses two major challenges in the estimation of continuous treatments. It reduces the selection bias of the representation through an adversarial game (Goodfellow et al., 2014) between the encoder and the treatment-predictor, and it also maintains the influence of the treatment in the hypothesis network by constraining the impact of the representation. Our experimental comparison on TCGA and News datasets demonstrates that ACFR matches or outperforms the state-of-the-art for the continuous ITE task. Our contributions are summarized in the following:

- We present a novel information-theoretic objective for continuous individual treatment effect estimation. The objective minimizes the selection bias as well as the factual outcome error with two mutual information terms.
- Leveraging variational approximation, we optimize our proposed objective with a neural network-based model called ACFR. ACFR takes the covariates of a unit and a continuous-valued treatment as inputs and predicts the potential outcome.
- We evaluate the performance of ACFR and state-of-the-art methods on TCGA and News datasets, and also analyze the robustness of ACFR to varying levels of selection bias.

2 Problem Statement

We assume a dataset of the form $D = \{x_i, t_i, y_i\}_{i=1}^N$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes the covariates of the ith unit, $t_i \in [0, 1]$ is the continuous treatment that unit i received and $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ denotes

[1] Continuous ITE estimation is also known as individual dose-response function estimation(Silva, 2016).

[2] Mutual information can capture non-linear dependency between two variables. For two variables X and Y, $I(X, Y)$ is the KL-divergence between $P(X, Y)$ and $P(X)P(Y)$ defined as $I(X, Y) = \int p(x, y) \log(\frac{p(x,y)}{p(x)p(y)}) dx\, dy$

the outcome of interest for unit i after receiving treatment t_i. N is the total number of units and d is the dimension size of covariates. We are interested to learn a machine learning model to predict the causal quantity $\mu(x,t) = E_{\mathcal{Y}}[Y(t)|X = x]$, which is the potential expected outcome under treatment t for individual with covariates x. Note that, unlike binary ITE, the goal is to predict all potential outcomes, and not the difference between them.[3] Same as previous works, we rely on the following assumptions to make treatment effect identifiable from an observational dataset.

Assumption a) Unconfoundedness (Rosenbaum & Rubin, 1983) ($\{Y_t\}_{t\in T} \perp\!\!\!\perp T|X$) which means given covariates, treatment and potential outcomes are conditionally independent.

Assumption b) Overlap (Imbens, 2004) ($P(T = t|X = x) > 0, \forall t \in [0,1], \forall x \in X$) which means every unit receives an arbitrary treatment level t with a probability greater than zero.

Having the above assumptions, $\mu(x,t)$ can be rewritten as follows, and we are able to estimate it using the observational dataset.

$$\mu(x,t) = E_{\mathcal{Y}}[Y(t)|X = x] = E_{\mathcal{Y}}[Y|X = x, T = t]$$

To estimate the above term, we propose to learn a predictive model $f : \mathcal{X} \times [0,1] \to \mathcal{Y}$ which takes covariate and treatment and predicts the outcome. We report the performance of model f in terms of the mean integrated square error (MISE) and the policy error (PE) metrics (Schwab et al., 2020).

3 METHOD

We assume the underlying causal structure of variables to be as in Figure 1. The latent representation Z extracted via a parametric encoder $p_\phi(z|x)$ is assumed to be causally dependent to covariate X, and to be conditionally independent of treatment T and outcome Y given X, i.e. $p(Z|X) = p(Z|X,T,Y)$. As outlined earlier, the balanced representation learning approach aims to extract a representation with minimum selection bias which maintains necessary information for predicting the factual outcomes. We measure the expressive power of the defined latent representation in the prediction of the factual outcome with the mutual information between Z and Y given T. We also measure the selection bias of the representation with the mutual information between Z and T. We aim to jointly maximize $I(Z,Y|T;\phi)$ and minimize $I(T,Z;\phi)$ with respect to encoder parameters ϕ as shown in Formula 1, where γ_1 controls the trade-off between two terms. We discuss the optimization of each MI term separately in the following.

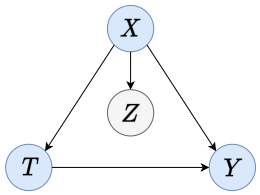

Figure 1: The underlying causal structure of variables.

$$\max_{\phi} I(Z,Y|T;\phi) - \gamma_1 I(Z,T;\phi) \tag{1}$$

3.1 SELECTION BIAS MINIMIZATION

As mentioned previously, to reduce the selection bias we want to minimize $I(T;Z)$ with respect to encoder parameters ϕ. $I(T,Z;\phi)$ can be written as $I(T,Z;\phi) = H(T) - H(T|Z;\phi)$, in which the entropy of treatment $H(T)$ is independent from encoder parameterization. Therefore, we focus on maximizing $H(T|Z;\phi)$ with respect to ϕ, defined as:

$$\max_{\phi} H(T|Z) = \max_{\phi} \mathrm{E}_{p(t,z)}[-\log p(t|z)] \tag{2}$$

To maximize the right-hand side of equation 2, we need to compute $p(t|z)$ which is determined from the encoder and the underlying causal graph $p(t|z) = \int \frac{p(x)p(t|x)p(z|x)}{p(z)}\, dx$. But since it is intractable, let $q_\pi(t|z)$ be the variational distribution defined over the same space to approximate it. $q_\pi(t|z)$ is the ACFR's treatment-predictor network that produces a distribution over the treatment given the representation. From conditional entropy definition in Farnia & Tse (2016), we have:

$$H(T|Z) = \inf_{\pi} \mathrm{E}_{p(t,z)}[-\log q_\pi(t|z)] \tag{3}$$

[3] In binary ITE, the goal is to estimate $\tau(x) = \mu(x,1) - \mu(x,0)$.

It means to compute the conditional entropy we have to minimize the above negative log-likelihood term with respect to π, and based on that the entropy maximization in equation 2 can be rewritten as follows:

$$\max_{\phi} H(T|Z) = \max_{\phi} \inf_{\pi} \mathrm{E}_{p(t,z)}[-\log q_\pi(t|z)] \tag{4}$$

Based on the causal graph in Figure 1, in which $p(Z|X) = p(Z|X,T)$, we can then incorporate covariate X in the above equation, which results in equation 5.

$$\max_{\phi} H(T|Z) = \max_{\phi} \inf_{\pi} \int p(x,t)p_\phi(z|x)[-\log q_\pi(t|z)]\, dz\, dx\, dt \tag{5}$$

In words, to reduce the mutual information (equivalently to increase the conditional entropy) between treatment and latent representation, an encoder parameterized by ϕ and a treatment-predictor parameterized by π should act adversarially against each other. Treatment-predictor attempts to minimize the negative log-likelihood term in the right-hand side of equation 5 by predicting the treatment accurately from Z, and the encoder attempts to maximize the term by excluding the information about treatment from the representation Z. We employ a deterministic encoder $z = g_\phi(x)$[4], and we assume $q_\pi(t|z)$ to be normal distribution $\mathcal{N}(f_\pi(z), \sigma^2) = \mathcal{N}(f_\pi(g_\phi(x)), \sigma^2)$ with fixed variance σ^2 and the mean is the output of the treatment-predictor network. Putting the probability density function of $q_\pi(t|z)$ in the right-hand side of equation 5, it will be:

$$\max_{\phi} H(T|Z) = \max_{\phi} \inf_{\pi} \frac{1}{2\sigma^2} \int p(x,t)[(t - f_\pi(g_\phi(x)))^2]\, dx\, dt + \log(\sigma\sqrt{2\Pi}) \tag{6}$$

Then, by substituting joint distribution $p(x,t)$ with the empirical data, and removing the constants from the right-hand side of equation 6, we can obtain a minimax loss term L_{adv} between the encoder and the treatment-predictor shown in Formula 7.

$$\max_{\phi} \inf_{\pi} \underbrace{\frac{1}{N} \Sigma_{i=1}^{N} (t_i - f_\pi(g_\phi(x_i)))^2}_{L_{adv}} \tag{7}$$

L_{adv} is optimized by alternatingly fixing the parameters of one network and optimizing the loss with respect to the parameters of the other one. Assuming that the treatment-predictor reaches its optimum with respect to any fixed encoder parameter during the training, we can obtain a useful property for our latent representation from the theoretical results of Wang et al. (2020): The global maximum of L_{adv} is achieved if and only if $\mathrm{E}[T|Z] = \mathrm{E}[T]$. It means the encoder is being optimized toward extracting a representation where the expectation of conditional distribution $p(T|Z)$ matches the expectation of marginal distribution $p(T)$. This reduces the discrepancy between $P(T|Z)$ and $P(T)$, and thus decreases the selection bias of representation considerably.

Nonetheless, since the distributions themselves are not guaranteed to be aligned, we expect that some amount of selection bias in the representation is not removed with the adversarial game. Without loss of generality, treatment-predictive covariates can be classified into instrumental variables (treatment-predictive only covariates) and confounders Hassanpour & Greiner (2019b). We expect that only the information of instrumental variables to be removed and information of confounders remains in the representation. The presence of the information related to confounders in the representation helps factual outcome prediction but hurts covariate balancing and counterfactual prediction [5]. In the next section, we propose a loss term L_{con} that allows the model to learn how to constrain the impact of confounding factors such that results in the least factual performance drop.

3.2 Factual Error Minimization

In this section, we view the prediction of outcome Y from representation Z and treatment T as maximization of $I(Y, Z|T; \phi)$. The conditional mutual information can be written as $I(Y, Z|T; \phi) = H(Y|T) - H(Y|Z, T; \phi)$. Since the conditional entropy $H(Y|T)$ is independent of the choice of

[4]Given covariate $x \in \mathcal{X}$ encoder maps it to some point in representation space $z \in \mathcal{Z}$ with the probability of one and to any other points with the probability of zero.

[5]This is a common downside of the balanced representation learning approach. Removing confounders from the representation hurts performance on outcome prediction and its presence hurts the generalizability of the model for counterfactual prediction.

ϕ, maximizing $I(Y, Z|T; \phi)$ is equivalent to minimizing $H(Y|Z, T; \phi)$, i.e. the uncertainty about outcome given representation and treatment.

$$\min_{\phi} H(Y|Z,T) = \min_{\phi} \mathrm{E}_{p(z,t,y)}[-\log p(y|z,t)] \tag{8}$$

$p(y|z,t) = \int \frac{p(x,y,t)p(z|x)}{p(z,t)} dx$ in the above equation is fully determined from joint distribution and our encoder, however, because it is intractable regard $q_\theta(y|z,t)$ as a variational approximation to $p(y|z,t)$. The variational distribution $q_\theta(y|z,t)$ is ACFR's hypothesis network taking representation and treatment as inputs and producing a distribution over the outcome. Leveraging the fact that Kullback Leibler divergence between $p(y|z,t)$ and its variational distribution is non-negative, we have: $\mathrm{E}_{p(z,t,y)}[-\log p(y|z,t)] \leq \mathrm{E}_{p(z,t,y)}[-\log q_\theta(y|z,t)]$. That means, we have an upper bound for $H(Y|Z,T)$ and after incorporating X based on the assumption that $p(Z|X) = p(Z|X,Y,T)$, the upper bound will be:

$$\min_{\phi} H(Y|Z,T) \leq \min_{\phi,\theta} \int p(z,t,y)[-\log q_\theta(y|z,t)]\, dz\, dt\, dy \tag{9.1}$$

$$= \min_{\phi,\theta} \int p(x,t,y)p_\phi(z|x)[-\log q_\theta(y|z,t)]\, dx\, dz\, dt\, dy \tag{9.2}$$

The upper bound (right-hand side of inequality 9.2) suggests that we can indirectly minimize $H(Y|Z,T)$ by minimizing the negative log-likelihood of factual outcomes with respect to the encoder and hypothesis network parameters. As mentioned earlier, hypothesis network $q_\theta(y|z,t)$ needs to consider a particular role for treatment in order to predict the outcome accurately. Unlike previous works that designed a special architecture for maintaining the treatment influence (Nie et al., 2021) (Schwab et al., 2020), we give treatment and representation as inputs to the hypothesis network and increase the treatment influence by constraining the representation impact. We add a term to the upper bound which encourages the hypothesis network to keep the predicted outcome distribution close to the original one even when the latent representation input has been perturbed slightly by some noise ϵ. Adding negative log-likelihood term for the perturbed representation $z + \epsilon$ to the upper bound in9.2 (the negative log-likelihood for the original representation z) results in a new upper bound for $H(Y|Z,T)$ shown in inequality 10. γ_2 is a hyper-parameter to be tuned and choosing ϵ will be discussed below.

$$\min_{\phi} H(Y|Z,T) \leq \min_{\theta,\phi} \int p(x,t,y)p_\phi(z|x)\,[-\log q_\theta(y|z,t) - \gamma_2 \log q_\theta(y|z+\epsilon,t)]\, dz\, dx\, dt\, dy \tag{10}$$

As discussed in the previous section, confounding factors exist in the representation. Pearl (2011) and Johansson et al. (2016) concluded that using the confounding features that are more associated with the treatment than the outcome is not desirable. Therefore, it is beneficial to add noise to dimensions according to their association with the treatment. This encourages the hypothesis network not only to increase treatment influence but also to rely on confounding factors as less as possible. To identify confounding dimensions in the representation, we utilize the treatment-predictor network because the association of a dimension with the treatment can be viewed as the contribution of the dimension in this network. We apply a simple gradient-based attribution method proposed in Simonyan et al. (2013) on the treatment-predictor to obtain the contribution of dimensions. Therefore, the noise ϵ for unit i is defined as $\epsilon_i = [\epsilon_i^1, ..., \epsilon_i^k]^T = [\mathcal{N}(0, C_i^1), ..., \mathcal{N}(0, C_i^k)]^T$. C_i is the contribution vector constructed by taking the absolute value of the gradient with respect to the representation of the ith unit, and C_i^j shows the contribution of the jth dimension.

Similar to the previous section, we derive a mean squared error for each negative log-likelihood term in the right-hand side of equation 10. We use the defined deterministic encoder $g_\phi(x)$, and we assume $q_\theta(y|z,t)$ to be normal distribution $\mathcal{N}(h_\theta(z,t), \sigma^2) = \mathcal{N}(h_\theta(g_\phi(x),t), \sigma^2)$ with fixed variance σ^2 and mean equals to the output of hypothesis network. Then, approximating $p(x,t,y)$ with empirical data, the right-hand side of inequality 10 turns into Formula 11 consisting of two loss terms. L_{fo} and L_{com} are MSE between the factual outcome and the predicted outcome, using the original and the perturbed representation respectively. Also, ϵ is replaced with ϵ_i since the added noise is computed for each unit exclusively.

$$\min_{\phi,\theta} \underbrace{\frac{1}{N}\Sigma_{i=1}^{N}(y_i - h_\theta(g_\phi(x_i), t_i))^2}_{L_{fo}} + \gamma_2 \underbrace{\frac{1}{N}\Sigma_{i=1}^{N}(y_i - h_\theta(g_\phi(x_i) + \epsilon_i, t_i))^2}_{L_{con}} \tag{11}$$

3.3 ADVERSARIAL COUNTERFACTUAL REGRESSION

We discussed minimizing $I(T;Z)$ and maximizing $I(Z,Y|T)$ in previous sections. We want to jointly optimize both terms, and thus we put the defined subnetworks together and propose Adversarial CounterFactual Regression network (ACFR). ACFR is trained end-to-end and its architecture is shown in Figure 2. Also, by substituting the mutual information terms in the objective 1 with the losses derived in section 3.1 and 3.2, we obtain the following loss function for ACFR.

$$\mathcal{L}_{ACFR} = \min_{\phi,\theta} \max_{\pi} L_{fo} + \gamma_2 L_{con} - \gamma_1 L_{adv} \tag{12}$$

ACFR pseudocode for optimizing the loss function is shown in Algorithm 1. The algorithm consists of three following stages per iteration.

1) A batch of units is sampled randomly, and the covariates of the batch are mapped to the latent representation using encoder g_ϕ.

2) For M times, L_{adv} loss for the encoded batch is computed and π is updated. [6]. Then, ϵ is constructed using the backpropagated gradient of L_{adv} with respect to z.

3) Perturbed representation $\tilde{z}$ is constructed using z and noise vector ϵ. All three loss terms L_{fo}, L_{con}, and L_{adv} are computed and ϕ and θ are updated.[7]

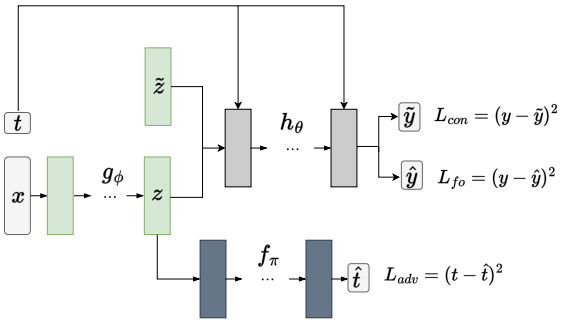

Figure 2: The architecture of ACFR network.

Algorithm 1 ACFR: Adversarial Counterfactual Regression

Input: Factual samples $(x_i, t_i, y_i)_{i=1}^N$, encoder network with initial parameter ϕ, treatment-predictor network with initial parameters π, hypothesis network with initial parameters θ, batch size b, iteration number I, internal loop size M, and hyper-parameters γ_1 and γ_2. The step sizes η_1 and η_2 are obtained using Adam (Kingma & Ba, 2015).

1: **for** $iter$ = 1 to I **do**
2: Sample a mini-batch $B = \{i_1, i_2, ..., i_b\} \subset \{1, 2, ..., N\}$.
3: Encode mini-batch covariates into latent representation.
$z_B = g_\phi(x_B)$.
4: **for** m = 1 to M **do**
5: Computing L_t and updating π.
$\hat{t}_B = f_\pi(z_B) \quad L_{adv} = \frac{1}{b}\Sigma_{i\in B}(t_i - \hat{t_i})^2 \quad \pi \leftarrow \pi - \eta_1 \nabla_\pi L_{adv}$
6: **end for**
7: Constructing noise vector ϵ_B.
8: Computing L_{fo}, L_{con}, L_{adv} and updating ϕ and θ.
$\tilde{z}_B = z_B + \epsilon_B \quad \hat{y}_B = h_\theta(z_B, t_B) \quad \tilde{y}_B = h_\theta(\tilde{z}_B, t_B) \quad \hat{t}_B = f_\pi(z_B)$
$L_{fo} = \frac{1}{b}\Sigma_{i\in B}(y_i - \hat{y_i})^2 \quad L_{con} = \frac{1}{b}\Sigma_{i\in B}(y_i - \tilde{y_i})^2 \quad L_{adv} = \frac{1}{b}\Sigma_{i\in B}(t_i - \hat{t_i})^2$
$[\phi, \theta] \leftarrow [\phi - \eta_2 \nabla_\phi (L_{fo} + \gamma_2 L_{con} - \gamma_1 L_{adv}), \theta - \eta_2 \nabla_\theta (L_{fo} + \gamma_2 L_{con})]$
9: **end for**

Output: ϕ, θ for test phase.

[6] π is updated M times to ensure it reaches the possible optimum for a fixed ϕ as discussed in section 3.1

[7] Note that as discussed earlier ACFR minimizes minimax loss L_{adv} w.r.t π in the second stage while ϕ is fixed, and maximizes it w.r.t. ϕ in the third stage while π is fixed.

Table 1: Datasets and data generating functions.

Dataset	#Samples	#Covariates	Treatment assignment	Outcome function
TCGA	9659	4000	$\beta = \frac{2(\alpha-1)v_2^T x}{v_3^T x} + 2 - \alpha$	$y = 10(v_1^T x + 12 v_2^T x t - 12 v_3^T x t^2)$
News	5000	3477	$t = \text{Beta}(\alpha, \beta)$	$y = 10(v_1^T x + \sin(\frac{v_2^T x}{v_3^T x}\pi t))$

4 EXPERIMENTS

Treatment effect estimation methods have to be validated in predicting potential outcomes including counterfactuals which are unavailable in real-world observational datasets. Inevitably, synthetic or semi-synthetic datasets are commonly used since the treatment assignment mechanism and outcome function are known and hence counterfactual outcomes can be generated. Note that this does not change the fact that only factual outcomes are accessible during training and methods need to address the selection bias. In this section, we discuss our experiment setting and evaluate the performance of ACFR and state-of-the-art methods. The code is available at https://github.com/CTEErepo/ACFR.

4.1 EXPERIMENTAL SETUP

Semi-synthetic data generation: We used TCGA (Network et al., 2013) and News (Johansson et al., 2016) semi-synthetic datasets (covariates are real-world data but treatment and outcome are generated synthetically). TCGA dataset consists of gene expression measurements of the 4000 most variable genes for 9659 cancer patients. The News dataset which was introduced as a benchmark in Johansson et al. (2016) consists of 3477 word counts for 5000 randomly sampled news items from the NY times corpus. For each dataset, we first normalized each covariate and then scaled every sample to have a norm 1. We then split the datasets with $68/12/20$ ratio into training, validation, and test sets. We followed treatment and outcome generating process of Bica et al. (2020b), summarized in Table 1. α in treatment function determines the selection bias level (α is set 2 in all experiments unless otherwise stated), and v_1, v_2 and v_3 are vectors that their elements are sampled from normal distribution $\mathcal{N}(0, 1)$, and then became normalized. Using the functions in Table 1, we assigned the treatment and factual outcome for all samples in the training and validation sets. All methods are then trained on the training set, and the validation set has been used for hyperparameter selection. Same as Bica et al. (2020b), potential outcomes for a unit are generated using the outcome function given the unit's covariates and 65 grids in the range $[0, 1]$ as an approximation of the treatment range.

Baselines: We compare ACFR against Varying Coefficient Network (VCNet) (Nie et al., 2021), Dose-Response Network (DRNet) (Schwab et al., 2020), two variants of counterfactual regression (CFR) (Shalit et al., 2017), CFR-Wass and CFR-HSIC, and multi-layer perceptron (MLP). MLP is a feed-forward network that simply takes covariates and continuous treatment and predicts the outcome. It is included to indicate the difficulty level of the problem on each dataset. CFR-Wass and CFR-HSIC originally proposed for binary treatments are adjusted to continuous treatments by considering multiple hypothesis heads on top of the representation each corresponding to a treatment range. The difference between adjusted CFRs and DRNet is that CFR-Wass and CFR-HSIC reduce selection bias of the representation by Wasserstein (Villani, 2008) and Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2007) distributional distance respectively. In contrast, DRNet does not reduce the selection bias but takes the treatment value as input of hypothesis heads. The implementation details and the hyperparameter search space of methods are detailed in Appendix A.1.

Metrics: Having $\mu(x, t)$ as the ground-truth outcome of the unit with covariate x under treatment t and $f(x, t)$ as the predicted outcome, we report the performance of methods in terms of the two following metrics defined in Schwab et al. (2020). The Mean Integrated Squared Error (MISE) is the squared error of the predicted outcome averaged over all treatment values and all units. The Policy Error (PE) measures the squared error of the estimated optimal policy averaged over all units.

$$\text{MISE} = \frac{1}{N}\Sigma_{i=1}^{N}\int_0^1 [\mu(x_i, t) - f(x_i, t)]^2 dt \qquad \text{PE}^8 = \frac{1}{N}\Sigma_{i=1}^{N}[\mu(x_i, t_i^*) - \mu(x_i, \hat{t}_i^*)]^2$$

[8] $t_i^* = \arg\max_t \mu(x_i, t)$ and $\hat{t}_i^* = \arg\max_t f(x_i, t)$.

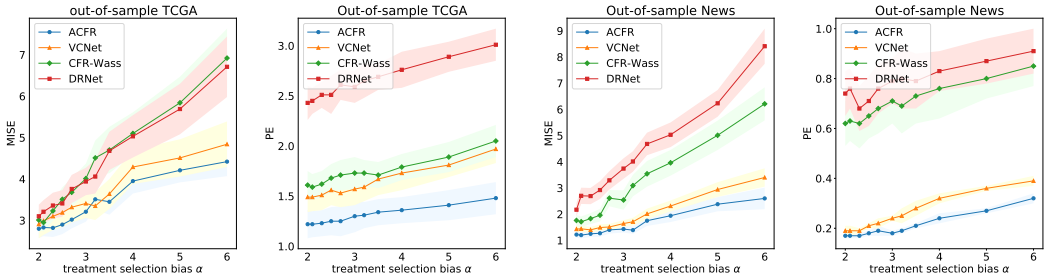

Figure 3: Performance of 4 methods with varying selection bias levels.

4.2 Baseline comparison

We performed two sets of experiments for potential outcome prediction, called out-of-sample prediction and within-sample prediction. The out-of-sample experiment shows the ability of models in predicting the potential outcomes for units in the held-out test set, and the within-sample experiment shows the ability for units in the training set. For all baselines, we reported the mean and the standard deviation of MISE and PE in the format of mean$\pm$std over 20 realizations of each dataset in Table 2. In the out-of-sample setting, ACFR achieved the best MISE and PE on both datasets. The same applies to the within-sample setting with the exception of MISE for the TCGA dataset on which VCNet outperformed ACFR. The PE of our method on the TCGA dataset and its MISE on the News dataset are significantly better than those of all contenders. We believe that ACFR outperformed VCNet in most of the scenarios because of the representation learning procedure of VCNet. VCNet extracts its representation based on the sufficiency theorem of the generalized propensity score (Hirano & Imbens, 2004) which holds for the average-level treatment effect but not for the individual-level treatment effect. Shi et al. (2019) stated that the representation learned based on this theorem does not necessarily lead to an accurate counterfactual predictor. Also, the superiority of ACFR compared to the CFR variants indicates the benefits of the adversarial approach compared to IPM metrics when balancing the representation for continuous treatments. Moreover, the number of parameters in the hypothesis network of ACFR is smaller than for all baselines. However, a training iteration of ACFR is more time-consuming and requires more samples since it has to train the treatment-predictor network as well.

Table 2: Results on News and TCGA datasets for out-of-sample and within-sample settings.

		Out-of-sample		**Within-sample**	
Dataset	Method	MISE	PE	MISE	PE
TCGA	MLP	4.12 ± 0.33	2.81 ± 0.17	3.37 ± 0.21	2.43 ± 0.18
	VCNet	2.91 ± 0.34	1.49 ± 0.15	$\mathbf{2.40 \pm 0.20}$	1.18 ± 0.15
	DRNet	3.10 ± 0.28	2.43 ± 0.17	2.48 ± 0.24	1.82 ± 0.15
	CFR-HSIC	3.02 ± 0.25	1.55 ± 0.18	2.41 ± 0.21	1.34 ± 0.12
	CFR-Wass	3.01 ± 0.26	1.61 ± 0.19	2.45 ± 0.20	1.36 ± 0.13
	ACFR	$\mathbf{2.80 \pm 0.22}$	$\mathbf{1.22 \pm 0.14}$	2.42 ± 0.17	$\mathbf{1.04 \pm 0.08}$
News	MLP	2.31 ± 0.2	1.30 ± 0.02	2.46 ± 0.22	1.28 ± 0.02
	VCNet	1.43 ± 0.12	0.19 ± 0.02	1.17 ± 0.07	0.16 ± 0.02
	DRNet	2.17 ± 0.25	0.74 ± 0.09	1.99 ± 0.25	0.61 ± 0.07
	CFR-HSIC	1.83 ± 0.21	0.65 ± 0.08	1.52 ± 0.31	0.51 ± 0.07
	CFR-Wass	1.76 ± 0.21	0.62 ± 0.08	1.45 ± 0.22	0.52 ± 0.07
	ACFR	$\mathbf{1.22 \pm 0.12}$	$\mathbf{0.17 \pm 0.01}$	$\mathbf{0.92 \pm 0.10}$	$\mathbf{0.16 \pm 0.01}$

4.3 Selection bias robustness

In this section, we study the robustness of 4 methods (ACFR, VCNet, DRNet, and CFR-Wass) against selection bias. As mentioned earlier, the α parameter of Beta distribution in the treatment generating function controls the amount of selection bias. As α increases the selection bias of the observational dataset increases and consequently, we expect the error of methods to increase as

well. As shown in Figure 3, ACFR has a consistent performance compared to the baselines on both datasets, and its gap with the contenders at the strong selection bias level (α = 6) is significant. In order to obtain a better understanding of the contribution of ACFR's components in its robust performance, we performed source of gain analysis. For each selection bias level, we compared the error of the MLP network trained to minimize L_{fo}, ACFR network without contribution loss term trained to minimize $L_{fo} - \gamma_1 L_{adv}$, and complete ACFR trained to minimize $L_{fo} - \gamma_1 L_{adv} + \gamma_2 L_{con}$. Figure 5 in Appendix A.3 indicates the contribution of each loss term to ACFR performance.

5 Related Works

Continous treatment effect estimation. Among the previous works that studied continuous treatments, Varying Coefficient Network (VCNet) (Nie et al., 2021) and Dose-Response Network (DRNet) (Schwab et al., 2020) are the most related works to ours. VCNet was proposed for continuous average treatment effect estimation and its main contribution is that instead of taking t as input, it gives more influence to the value of the treatment through a dynamic neural network parameterized by a spline basis of t. Despite substantially improving ADRF estimation on News and IHDP datasets, VCNet is not as accurate in individual-level treatment effect estimation tasks. DRNet, proposed to estimate the continuous ITE, discretizes treatment range into K intervals and employs a hierarchical neural network with shared layers for all samples and specific head layers each shared only among samples having received treatment within the same interval. Apart from the parameter complexity that these methods introduce to the network, their performance might be domain-dependent. For instance, under severe covariate shift among samples, some head layers in DRNet may receive a limited number of samples which is not sufficient for training a neural network. Also, the choice of the best spline functions in the VCNet architecture depends on the dataset, and the authors did not discuss a systematic way for choosing them.

Adversarial balanced representation for treatment effect estimation. Learning a balanced representation via an adversarial procedure has been studied for the categorical treatment effect estimation problem. Du et al. (2021) and Berrevoets (2020) employed a binary treatment-predictor (classifier) to balance the distributions of treated and control groups in the latent space. Bica et al. (2020a) extended the approach to multiple time-varying treatment setting and proved that the global minimum of the proposed minimax game between encoder and treatment-predictor is achieved if and only if distributions of treatment groups are aligned. Our work differs from previous adversarially balanced methods. We considered continuous treatment which requires treatment-predictor to regress the assigned treatment. We justified the adversarial balancing approach by the mutual information minimization $I(T, Z; \phi)$ term between treatment and representation variables. Finally, unlike the above-mentioned works, we provided the analysis of the adversarially balanced approach with regard to the extent of selection bias removal and how ACFR deals with confounding variables.

The expanded related works can also be found in Appendix A.2.

6 Discussion

This paper presented the method ACFR for predicting potential outcomes under a continuous-valued treatment. Based on a novel information-theoretic objective for the continuous treatment effect estimation problem, ACFR learns a latent representation of the covariates that reduces the selection bias and prevents the outcome prediction from overfitting to the representation. ACFR outperformed the state-of-the-art methods on two semi-synthetic datasets, and its performance was shown to be robust against varying levels of selection bias. This paper also suggests various directions for future research to extend the ACFR framework and its applications. For instance, the adversarial treatment-predictor can be designed to predict higher-order moments as well (e.g. variance), which theoretically guarantees further selection bias reduction (Wang et al., 2020). Also, more sophisticated attribution methods can be employed to obtain a better estimate of the association between the latent dimensions and the assigned treatment. From the application perspective, ACFR can be extended to consider other forms of treatment such as a tuple of a categorical treatment and its associated continuous dosage(Schwab et al., 2020), which is achievable by having a treatment-predictor for each treatment type.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. 2016. doi: 10.48550/ARXIV.1612.00410. URL https://arxiv.org/abs/1612.00410.

Alexis Bellot, Anish Dhir, and Giulia Prando. Generalization bounds and algorithms for estimating conditional average treatment effect of dosage, 2022. URL https://arxiv.org/abs/2205.14692.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Jeroen Berrevoets. Organite: Optimal transplant donor organ offering using an individual treatment effect. In *NeurIPS*, 2020.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. 2020a. doi: 10.48550/ARXIV.2002.04083. URL https://arxiv.org/abs/2002.04083.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks, 2020b.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), mar 2010. doi: 10.1214/09-aoas285. URL https://doi.org/10.1214%2F09-aoas285.

Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4):1713–1738, may 2021. doi: 10.1007/s10618-021-00759-3. URL https://doi.org/10.1007%2Fs10618-021-00759-3.

Farzan Farnia and David Tse. A minimax approach to supervised learning, 2016. URL https://arxiv.org/abs/1606.02206.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. 2015. doi: 10.48550/ARXIV.1505.07818. URL https://arxiv.org/abs/1505.07818.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *NIPS*, 2007.

Negar Hassanpour and Russ Greiner. Counterfactual regression with importance sampling weights. pp. 5880–5887, 08 2019a. doi: 10.24963/ijcai.2019/815.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019b.

Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, March 2009. doi: 10.1257/jel.47.1.5. URL https://www.aeaweb.org/articles?id=10.1257/jel.47.1.5.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference, 2016. URL https://arxiv.org/abs/1605.03661.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv: Statistics Theory*, 2017.

Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, and Fabio Vandin. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013. ISSN 1061-4036. doi: 10.1038/ng.2764.

Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Invited Commentary: Understanding Bias Amplification. *American Journal of Epidemiology*, 174(11):1223–1227, 10 2011. ISSN 0002-9262. doi: 10.1093/aje/kwr352. URL https://doi.org/10.1093/aje/kwr352.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818. URL https://doi.org/10.1080/01621459.1994.10476818.

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, 1983.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5612–5619, 2020.

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

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

Ricardo Silva. Observational-interventional priors for dose-response learning, 2016. URL https://arxiv.org/abs/1605.01573.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013. URL https://arxiv.org/abs/1312.6034.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle, 2015. URL https://arxiv.org/abs/1503.02406.

Cédric Villani. *Optimal transport – Old and new*, volume 338, pp. xxii+973. 01 2008. doi: 10.1007/978-3-540-71050-9.

Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. *arXiv preprint arXiv:2007.01807*, 2020.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *NeurIPS*, 2018.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Trans. Knowl. Discov. Data*, 15(5):74:1–74:46, 2021. URL https://doi.org/10.1145/3444944.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning, 2019. URL https://arxiv.org/abs/1911.02685.

A Appendix

A.1 Hyperparameters and Implementation details

We re-implemented DRNet because it considers a different form of treatment, a tuple consisting of a categorical treatment and associated continuous dosage. Since we were interested in DRNet capability only in the continuous dosage setting, the same architecture was not applicable. The number of hypothesis heads of this method is chosen from $[5, 7, 10]$. The CFRs are also re-implemented because they were originally proposed for the binary treatment setting. The hyperparameter balancing the IPM loss and factual loss for CFRS is selected from $[0.01, 0.05, 0.1, 0.5, 1, 5, 10]$ range. Also, the number of hypothesis heads on top of the representation is chosen from $[5, 7, 10]$.

We fine-tuned VCNet because its performance on the individual level is of important. VCNet needs two parameters to construct its dynamic hypothesis network, called degree and knots. We used $[2, 3, 4]$ and $[\{1/3, 2/3\}, \{1/4, 2/4, 3/4\}]$ ranges for degree and knots parameters respectively. For the hyperparameter balancing the loss term of factual outcome prediction and the loss term of propensity score density, we searched in $[0.01, 0.05, 0.1, 0.5, 1, 5, 10]$ range. The number of output nodes in its density network is also chosen from $[10, 15, 30]$. Note that, to ensure fairness the same spline functions used in the architecture of VCNet are given to other methods as inputs. For example, with the degree, $= 2$ and knots $= [1/3, 2/3]$, the spline functions are $[1, t, t^2, (t-1/3)^2, (t-2/3)^2]$, and these functions are also the input of the hypothesis heads of other methods.

The hyperparameters for ACFR's adversarial loss L_{adv} and contribution loss L_{con} are chosen from $[0.05, 0.1, 0.2, 0.5, 1, 5, 10]$ and $[0.1, 0.2, 0.5]$ lists respectively. The hyperparameters shared among all methods (e.g. learning rate, dropout) are tuned using the search ranges shown in the Table below.

Hyper-parameter	**search range**
learning rate	$[5 * 10^{-5}, 10^{-4}, 5 * 10^{-4}, 10^{-3}, 5 * 10^{-3}, 10^{-2}]$
Batch size	[32, 64, 128]
Dropout	[0, 0.1, 0.2]
Maximum epoch	[100, 300, 600]
L_2 Regularization coefficient	$[10^{-4}, 5 * 10^{-4}, 10^{-3}]$
Number of layers in encoder	[1, 2, 3]
Number of layers in hypothesis	[1, 2]
Number of nodes per layer	[50, 100, 200]

Figure 4 shows the effect of two hyperparameters of ACFR loss function (γ_1 and γ_2) on the performance. For both datasets, we can extract a pattern from this experiment. In the case that γ_1 is very small, L_{adv} loss does not sufficiently impact the network, and thus even instrumental variables may remain in the representation space. Therefore, with more contribution of L_{con} the model achieves better performance. In the case that γ_1 is large, the representation might lose the necessary information for outcome prediction, and thus substantial impact of L_{con} leads to worse performance.

A.2 Expanded related works

Adversarial Domain Adaptation. Adversarial domain adaptation approaches are established on the assumption that a generalizable latent representation contains almost no discriminative information about the domain of the samples (Zhuang et al., 2019). In the standard setting of transferring knowledge from one (or multiple) source domain(s) to one (or multiple) target domain(s), studies such as Ganin et al. (2015), encourage the encoder toward extracting a domain-invariant representation

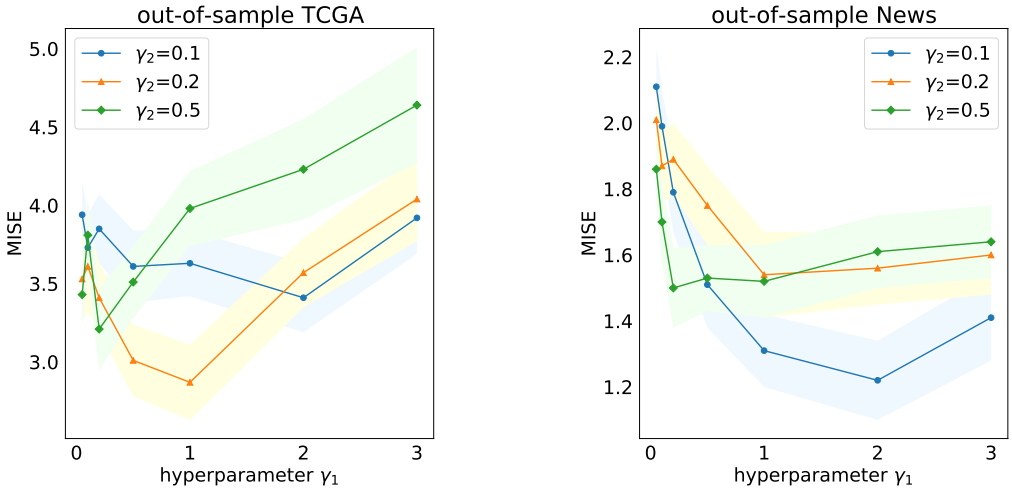

Figure 4: Effect of loss function hyperparameters on ACFR's performance.

through an adversarial game between a domain-classifier and the encoder. Recently, continuously indexed domain adaptation (CIDA) (Wang et al., 2020) generalized the setting to a possibly infinite number of domains. i.e. domains that are indexed continuously, replacing the domain-classifier with a domain-regressor that predicts the expected index of a domain given the representation. We borrowed the idea of the adversarial game between feature extractor and treatment-regressor from Wang et al. (2020) and applied it to a new problem in the context of counterfactual prediction.

A.3 Source of gain analysis

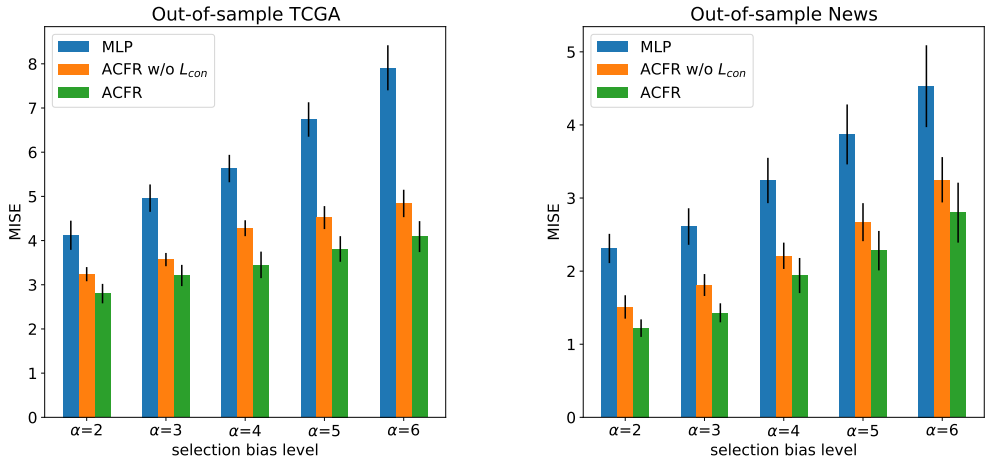

Figure 5: Source of gain analysis in varying selection bias level.