ACE: Adaptively Similarity-preserved Representation Learning for Individual Treatment Effect Estimation

Liuyi Yao[1], Sheng Li[2], Yaliang Li[3], Mengdi Huai[4], Jing Gao[1], Aidong Zhang[4]
[1]*University at Buffalo, {liuyiyao, jing}@buffalo.edu*
[2]University of Georgia, sheng.li@uga.edu
[3]Alibaba Group, yaliang.li@alibaba-inc.com
[4]University of Virginia, {mh6ck,aidong}@virginia.edu

***Abstract*—Treatment effect estimation refers to the estimation of causal effects, which benefits decision-making process across various domains, but it is a challenging problem in real practice. The estimation of causal effects from observational data at the individual level faces two major challenges, i.e., treatment selection bias and missing counterfactuals. Existing methods tackle the selection bias problem by learning a balanced representation and infer the missing counterfactuals based on the learned representation. However, most existing methods learn the representation in a global manner and ignore the local similarity information, which is essential for an accurate estimation of causal effects. Motivated by the above observations, we propose a novel representation learning method, which adaptively extracts fine-grained similarity information from the original feature space and minimizes the distance between different treatment groups as well as the similarity loss during the representation learning procedure. Experiments on three public datasets demonstrate that the proposed method achieves the best performance in causal effect estimation among all the compared methods and is robust to the treatment selection bias.**

***Keywords*-treatment effect estimation; similarity preserving; representation learning**

I. Introduction

Causal effect estimation is an essential task across many domains, such as business [1], [2], sociology science [3], bioinformatics [4], and healthcare [5]. It provides a powerful tool to support decision-making. For example, in the healthcare domain, treatment effect estimation can answer questions like "If a diabetics patient had used another oral antidiabetic medicine, would she/he be better?" to provide better therapies. In the social domain, it can answer questions like "if she/he had participated in the job training, would she/he get a job?" to help people to have a better career. In these questions, the first clause describes the treatment, and the second describes the counterfactual outcomes. Treatment effect estimation can answer all the above questions by estimating the causal effects that measure the expected differences between the outcomes of different treatments (i.e., settings or interventions).

In the Big Data era, a huge amount of data are accumulated, which can be used to conduct treatment effect estimation. For most of the data, treatment assignment is not explicitly controlled, and such data are known as the observational data [6]. Due to its easy access and low cost, the estimation of causal effects from the observational data at either the population-level or individual-level has been widely adopted [2], [7], [8]. However, there are two challenges in practice when estimating the causal effect at the individual level: *missing counterfactuals* and *treatment selection bias*.

The missing counterfactual challenge comes from the fact that an individual can only accept one treatment, so the outcomes of other treatments (i.e., counterfactual outcomes) are always unknown [9]. However, treatment effect estimation requires comparing the outcomes of an individual under different treatments. A possible solution is to infer missing counterfactuals from the observations of other individuals, and the underlying principle is that similar individuals with the same treatment should have similar outcomes.

The second challenge, treatment selection bias, is brought by the fact that individuals have their own preferences for treatment selection. Such selection bias increases the difficulty of the aforementioned counterfactual inference: We need to estimate an individual's counterfactual from another group in which people usually have preferences different from the preference of this particular individual.

To tackle the above two challenges, existing methods [10]–[12] project individuals into a balanced representation space, where different treatment groups are close to each other, and then an outcome prediction model is trained to estimate the counterfactuals. Most of these methods balance the distributions of different groups from a global view and ignore the local similarity information, and thus the relative similarity information between units might be omitted when learning the representation. In [12], the similarity preserved individual treatment effect estimation (SITE) method is proposed to retain the similarity information when learning the latent representation. However, SITE only considers the similarity of extreme cases (i.e., the coarse-grained similarity information), and the causal effect estimation based on such cases may not have enough improvement. Meanwhile, SITE also requires that the underlying data are spherically distributed when calculating the group distance, which might be unrealistic in the high dimensional data.

Motivated by the above observations, we propose an **A**daptively similarity-preserved representation learning method for **C**ausal **E**ffect estimation (**ACE**). Through the deep

2374-8486/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDM.2019.00186

representation network, ACE maps the individuals from the original space to the representation space. Then, the outcomes are estimated based on the learned representations. The most important component of ACE is the Balancing & Adaptive-Similarity preserving (BAS) regularization applied on the representation space. BAS regularization not only balances the control/treated group but also adaptively preserves the important similarity information from the original space. For units located in the regions where most of the units are from the same treatment group, it is important to preserve the similarity information when learning the new representation; while for the units in the mixture region of different treatment groups, the preserving strength of pairwise similarity can be made weaker when they are mapped to the representation space. In general, the representation learnt by the proposed ACE method has more overlapping between control/treated group and preserves the fine-grained similarity information, and corresponding the causal effect estimation can be greatly improved. Experiments on three public datasets demonstrate the effectiveness of our method.

II. Methodology

A. Preliminary

Let $\mathbf{X}$ denote all the feature variables and let W_i denote the binary treatment assignment on unit i, i.e., $W_i = 0$ or 1. The unit i is in the treated group if $W_i = 1$, and belongs to the control group if $W_i = 0$. Before the treatment assignment, any outcome $Y_1^{(i)}$ (treated) or $Y_0^{(i)}$ (control), is a *potential outcome*. After the intervention, the outcome $Y_{W_i}^{(i)}$ is the *observed outcome* or *factual outcome*, and the other treatment's outcome is the *counterfactual outcome*.

Treatment effect can be estimated at either the population-level or individual-level. We mainly focus on the Individual treatment effect (ITE) estimation in this paper. The ITE for unit i is defined as[1]: $ITE_i = Y_1^{(i)} - Y_0^{(i)}$, where $Y_1^{(i)}$ and $Y_0^{(i)}$ are treated and control outcome of i-th unit.

The success of the potential outcome framework is based on the four assumptions [5], [13]: **Stable Unit Treatment Value Assumption (SUTVA)**, **Consistency**, **Ignorability** and **Positivity** [14]. These assumptions ensure the identification of the ITE.

B. Overview

When estimating the individual treatment effect, Shalit et. al. [11] and Ahmed et. al. [9] prove that the bound of ITE estimation error comprises two parts: the divergence between the control/treated group and the outcome prediction loss. In the light of the theoretical results, our proposed ACE method imposes the BAS regularization to decrease the discrepancy of control/treated group and reduce the outcome prediction error by adaptively preserving the similarity information when learning the representation. Fig. 1 shows the framework of SCE, which contains two procedures: (1) Representation learning procedure which learns the balanced and similarity preserved representation; (2) Outcome prediction procedure which estimates all the potential outcomes using the learned representations. The following sections introduce the two procedures in detail.

[1]In some literature, ITE is also known as conditional average treatment (CATE), which is defined as $CATE = \mathbb{E}[Y_1 - Y_0|\mathbf{X} = \mathbf{x}]$.

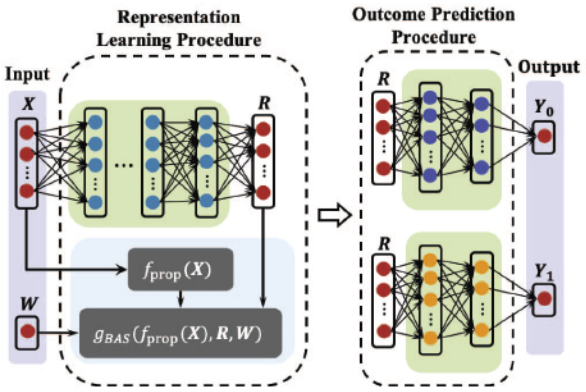


Fig. 1: Framework of ACE. The covariate $\mathbf{X}$ is fed into the representation network to get the latent representation $\mathbf{R}$. Meanwhile, the propensity score $f_{prop}(X)$ is calculated and then fed into the Balancing & Adaptive-Similarity preserving (BAS) regularization denoted as $g_{BAS}(*)$. After the representation learning procedure, two potential outcomes $\hat{Y}_0$ and $\hat{Y}_1$ are finally obtained through the outcome prediction procedure.

C. Representation Learning Procedure

In the representation learning procedure, ACE first learns the representation via the standard feed-forward neural network: $\mathbf{R} = f_{rep}(\mathbf{X}; \Theta_{rep})$, where $\mathbf{R}$ is the latent representation, and f_{rep} denotes the neural network with Θ_{rep} as its parameters. To decrease the ITE estimation error, the BAS regularization is applied to the representation layer. The details of BAS regularization are illustrated in the following section.

1) BAS Regularization Overview: As mentioned previously, existing similarity preserved work might be inadequate to capture the control/treated group discrepancy and preserve the important similarity information as much as possible, when only taking the selected triplets into consideration. To address this issue, ACE utilizes a new strategy, called Balancing & Adaptive-similarity preserving regularization (BAS) regularization, which can overcome the shortcomings of the existing work. The BAS regularization contains two components: (1) Distribution distance minimization. (2) Adaptive pairwise similarity preserving when mapping units from the original space to the representation space. The next two sections will describe the two components precisely.

2) Group Distance Minimization: In the representation space, the distance between different treatment groups should be minimized. Similar to the metrics used in [11], we adopt the integral probability metric (IPM) [15], [16] to measure the distances between different treatment groups. Then, the distribution distance minimization term $\mathcal{L}_d$ is defined as:

$$\mathcal{L}_d = IPM(\mathbf{R}_{I_c}, \mathbf{R}_{I_t}), \tag{1}$$

where $I_c = \{i : W_i = 0\}$ and $I_t = \{i : W_i = t\}$ are the index set of control and treated group; $\mathbf{R}_{I_c}$ and $\mathbf{R}_{I_t}$ are the represen-

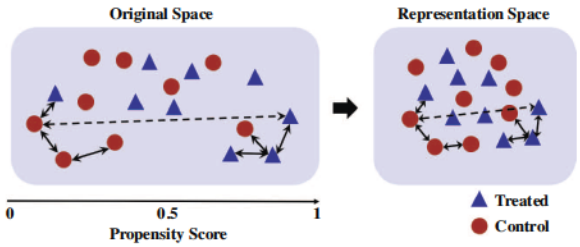


Fig. 2: A 2-D example of adaptive similarity preserving.

tations of treated and control groups, respectively. The adopted distance metric is capable to measure the control/treated group discrepancy more precisely.

3) Adaptive Similarity Preserving: Local similarity information is essential for counterfactual inference, as similar units tend to have similar outcomes. In contrast to existing similarity preserved ITE estimation method, we propose a novel strategy that preserves the fine-grained similarity which adjusts the strength of the pairwise similarity preservation according to the data distributions. Therefore, we call it *adaptive similarity preserving*.

Similarity Preserving Strength. Usually the treated and control groups are distributed in the same covariate space with partial overlap. In the intermediate region with sufficient overlap, both the control and treated units are relatively dense, which means a certain level of similarity change would not affect the counterfactual inference. Thus, the similarity preserving strength can be made weaker. While, for the regions where the treated and control units are incredibly unbalanced, the local similarity relationship will be changed dramatically in the representation space after distribution distance minimization, which further incurs an unreliable estimation of the counterfactual outcome. Therefore, preserving the local similarity information in these regions will be critical. A 2-D toy example is shown in Fig. 2 to illustrate the effect of above adaptive similarity preserving. The details of the adaptive similarity preserving are explained as follows.

Similarity Preserving Loss. Motivated by the dimensionality reduction methods, stochastic neighbor embedding (SNE) and t-SNE [17], we measure the similarity loss by K-L divergence during representation learning. By minimizing the K-L divergence, the similarity information extracted from the original covariate space is preserved as much as possible in the representation space. The proposed similarity preserving regularization is formulated as:

$$\mathcal{L}_s(\mathcal{P},\mathcal{Q}) = -\sum_{i,j}\mathcal{P}_{i,j}\log\frac{\mathcal{Q}_{i,j}}{\mathcal{P}_{i,j}}, \tag{2}$$

where $\mathcal{P}$ denotes the joint probability of $\mathbf{x}_i$ and $\mathbf{x}_j$:$\mathcal{P}_{i,j} = \frac{\exp(S(\mathbf{x}_i,\mathbf{x}_j))}{\sum_{k\neq l}\exp(S(\mathbf{x}_k,\mathbf{x}_l))}$ with $S(\cdot,\cdot)$ being the similarity function; And $\mathcal{Q}$ denotes the joint probability of $\mathbf{R}_i$ and $\mathbf{R}_j$, which is calculated as: $\mathcal{Q}_{i,j} = \frac{\exp\left(-\|\mathbf{R}_i-\mathbf{R}_j\|^2\right)}{\sum_{k\neq l}\exp(-\|\mathbf{R}_k-\mathbf{R}_l\|^2)}$.

Similarity Score Function. The most important part in calculating the similarity preserving loss is the similarity score function $S(\cdot,\cdot)$, which reflects the similarity preserving strength in the original covariate space. Motivated by [12], the definition of $S(\cdot,\cdot)$ is:

$$\begin{aligned} S(\mathbf{x}_i,\mathbf{x}_j) = \; & 0.75\left|\tfrac{f_{prop}(\mathbf{x}_i)+f_{prop}(\mathbf{x}_j)}{2} - 0.5\right| \\ & -0.5\left|f_{prop}(\mathbf{x}_i) - f_{prop}(\mathbf{x}_j)\right| + 0.5, \end{aligned} \tag{3}$$

where $f_{prop}(\cdot)$ is the pre-trained propensity score function. The similarity function calculation is based on the propensity score. The propensity score is the probability that the unit is treated conditioned on the covariates [18]. Based on the propensity score, our similarity function measures the similarity in two aspects: (1) The first term measures the *similarity preserving strength*, which is the deviation of the paired units to the intermediate region. The larger the deviation is, the larger the preserving strength is, and the higher the similarity score is; (2) The second term measures *the relative distance within the pair*, which can be viewed as the original similarity. The larger the relative distance is, the smaller the similarity score. By considering both the deviation as well as the relative distance, the similarity function $S(\cdot,\cdot)$ integrates the original similarity and the similarity preserving strength together.

4) BAS Regularization Summary: The BAS regularization is formulated as:

$$g_{BAS}(f_{prop}(\mathbf{X}),\mathbf{R},\mathbf{W}) = \alpha\mathcal{L}_d + \gamma\mathcal{L}_s, \tag{4}$$

where $\mathcal{L}_d$ and $\mathcal{L}_s$ are defined in Eqn. (1) and Eqn. (2), respectively. Overall, the BAS regularization enjoys the following benefits: (1) With the help of integral probability metric, the control/treated group discrepancy can be better measured; (2) BAS explores all the pairwise similarity and adaptively preserves the important similarity information.

D. Outcome Prediction Procedure

Based on the learned representation, the outcome prediction procedure infers the two potential outcome. As suggested by [9], it is better to use separate model to infer control/treated outcomes: $\hat{Y}_0 = f_c(R;\Theta_c)$, $\hat{Y}_1 = f_t(R;\Theta_t)$, where f_c and f_t are the neural networks, parameterized by Θ_c and Θ_t respectively, to predict the control and treated outcome.

In all, the factual outcome prediction loss can be calculated as:

$$\mathcal{L}_f = \sum_{i\in I_c} b_i L(Y_F^{(i)}, f_c(R_i;\Theta_c)) + \sum_{i\in I_t} b_i L(Y_F^{(i)}, f_t(R_i;\Theta_t)), \tag{5}$$

where I_c (I_t) is the index set of all control (treated) units; $L(\cdot,\cdot)$ denotes the loss measure function. For continuous outcomes, the square loss is adopted, and for categorical outcomes, the cross entropy loss is adopted. b_i is the re-weighting term and $b_i = \frac{N}{2\sum_{j=1}^{N}W_j}W_i + \frac{N}{2(N-\sum_{j=1}^{N}W_j)}(1-W_i)$. Note that in the observational dataset, the size of the control group is usually much larger than that of the treated group, so re-weighting each unit in the factual loss is needed.

E. Objective Function

Combining BAS regularization and the factual loss in estimating the observed outcomes, Eqn. (6) gives us the final loss function.

$$\mathcal{L} = \mathcal{L}_f + \alpha\mathcal{L}_d + \gamma\mathcal{L}_s + \lambda(||\Theta_{rep}^{-bias}||_2 + ||\Theta_c^{-bias}||_2 + ||\Theta_t^{-bias}||_2), \tag{6}$$

where $\mathcal{L}_f$ is the factual loss shown in Eqn. (5). $\mathcal{L}_d$ and $\mathcal{L}_s$ forms the BAS regularization, and are the same as Eqn. (1) and Eqn. (2) respectively. The last term is the parameter regularization, and Θ_*^{-bias} denotes the parameters excluding the bias term. α, γ and λ are three trade-off parameters.

By minimizing the total loss $\mathcal{L}$, the proposed model estimates two potential outcomes as well as the counterfactual outcomes upon the representation space, where the distributions of different groups are adaptively balanced.

Optimization. The networks $f_{rep}(\cdot)$, $f_c(\cdot)$ and $f_t(\cdot)$ are all feed-forward neural networks with ELU [19] as the activation function. We adopt the Adam optimizer [20] to optimize the objective function.

III. Experiment

A. Experimental Setting

1) Dataset: The datasets we adopt are the same as [12], which are three public datasets IHDP, Jobs, and Twins. On IHDP and Twins datasets, we average over 10 realizations with $61/27/10$ ratio of train/validation/test splits. And on Jobs dataset, because of the extremely low treated/control ratio, we conduct the experiment on 10 train/validation/test splits with $56/24/20$ split ratio, as suggested in [11].

2) Performance Metric: On IHDP and Twins dataset, the expected Precision in Estimation of Heterogeneous Effect (PEHE) [21] is adopted. The lower the $\mathcal{E}_{\text{PEHE}}$ is, the better the method is. On Jobs dataset, only the observed outcomes are available and the ground truth of ITE is unavailable. We adopt the policy risk [11] to measure the expected loss when taking the treatment as the ITE estimator suggests. Policy risk reflects how good the ITE estimation can guide the decision. The lower the policy risk is, the better the ITE estimation model can support the decision making.

3) Baselines: We compare the proposed method with the following three groups of baselines: Regression based methods: Least square Regression with the treatment as feature (**OLS/LR$_1$**), separate linear regressors for each treatment group (**OLS/LR$_2$**); Nearest neighbor matching based methods: Hilbert-Schmidt Independence Criterion based Nearest Neighbor Matching (**HSIC-NNM**) [22], Propensity score match with logistic regression (**PSM**) [18], k-nearest neighbor (**k-NN**) [23]; Tree based method: Causal Forest (**C. Forest**) [24]; Representation learning based methods: Balancing neural network (**BNN**) [10], counterfactual regression with MMD metric (**CFR-MMD**) [11], counterfactual regression with Wasserstein metric (**CFR-WASS**) [11], Treatment-Agnostic Representation Network (**TARNet**) [11] and Similarity Preserved Individual Treatment Effect Estimation(**SITE**) [12].

B. Result Analysis

1) Performance Comparison: The performance of ACE and baselines are summarized in Table I. The proposed method achieves the best results on both IHDP and Twins datasets. On Jobs dataset, ACE has the best performance in the out-of-sample case, and achieves similar result with the best baseline CFR-MMD in the within-sample case. The results demonstrate that jointly minimizing the group distance and preserving the fine-grained pairwise similarity information during the representation learning can benefit ITE estimation.

Among the representation learning based models, CFR-MMD, CFR-WASS, and SITE are competitive baselines. CFR-MMD and CFR-WASS are similar in that they both minimize the distribution distance in the representation space, and train separate outcome prediction models for different treatments. Different from CFR-WASS and CFR-MMD, additionally, SITE preserves the similarity information among the selected triplet pairs in each mini batch. As similarity information is helpful for outcome inference, in most of the cases, SITE performs better than CFR-MMD and CFR-WASS. In comparison with SITE, our proposed ACE method has the superior result, because ACE utilizes the BAS regularization to calculate the group discrepancy more accurately and fully retain the fine-grained important similarity information when learning the representations. Specially, on IHDP dataset, ACE performs 23.5% and 17.5% better than the best baseline SITE in within-sample and out-of-sample case, respectively.

2) The Effect of BAS Regularization Components.: BAS regularization contains two components: distance minimization and adaptive similarity preservation. To analyze the effect of these two parts, we compare ACE with its two variants: ACE without distance minimization component (ACE w/o B) and ACE without similarity preservation component (ACE w/o S). Fig. 3 shows the performance of ACE and its variants on the three datasets. It is observed from the figure that, except the within-sample case of Jobs dataset, ACE performs much better than its variants in most of the cases. Overall, when propensity score is relative accurate, BAS regularization can greatly enhance the ITE estimation.

C. Experiment on Treatment Selection Bias

In the problem of estimating causal effect from observational data, selection bias is one major challenge. To validate the performance of ACE under different levels of selection bias, we conduct the following experiments on IHDP and Twins datasets.

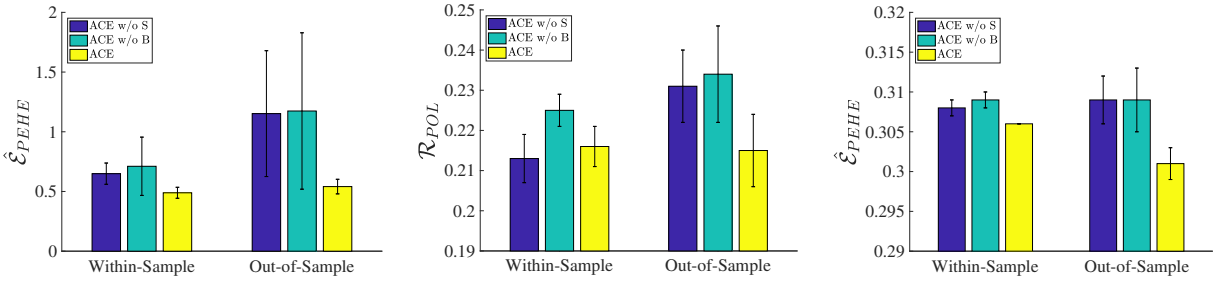
1) Treatment Selection Bias Creation: Depending on the way to vary selection bias, we have the following two cases:

Case 1: the selection bias is varied based on the propensity scores. On IHDP datasets, following the settings in [11], with probability q, we remove the control units that have propensity score closest to 1. Removing the control unit close to 1 creates less overlap between the control and treated groups. Thus, the higher the q is, the larger the selection bias is. We vary the removing probability q from 0.5 to 1.

Case 2: the selection bias is varied based on the variables. On Twins dataset, following the settings in [25], the selection bias is varied based on the variable `GESTAT10`, which is highly correlated with the outcome. The treatment is assigned as follows: $t_i|\mathbf{x}_i \sim Bern(Sigmoid(\mathbf{w}_{-g}^{'}\mathbf{x}_{i,-g} + w_g(x_{i,g}/10 + 0.1) + n))$,$\mathbf{w}_{-g} \sim \mathcal{U}(-0.1, 0.1)^{39\times 1}$,$w_g \sim \mathcal{N}(\mu_g, 0.1)$, $n \sim \mathcal{N}(0, 0.1)$ where $\mathbf{x}_{i,-g}$ denotes the pre-treatment covariates except the `GESTAT10`, and $x_{i,g}$ is the `GESTAT10`. The μ_g controls the weight of `GESTAT10` in the treatment assignment procedure. By varying μ_g from 0 to 5, different levels of selection bias are simulated.

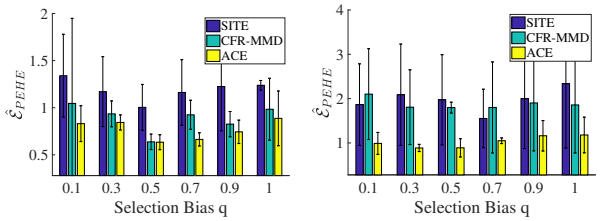
TABLE I: Performance Comparison.

Method	**IHDP** ($\mathcal{E}_{\text{PEHE}}$)		**Jobs** ($\mathcal{R}_{\text{pol}}$)		**Twins** ($\hat{\mathcal{E}}_{\text{PEHE}}$)	
	Within-Sample	Out-of-Sample	Within-Sample	Out-of-Sample	Within-Sample	Out-of-Sample
OLS/LR$_1$	10.761 ± 4.350	7.345 ± 2.914	0.297 ± 0.010	0.307 ± 0.084	0.308 ± 0.001	0.309 ± 0.012
OLS/LR$_2$	10.280 ± 3.794	5.245 ± 0.986	0.295 ± 0.006	0.297 ± 0.084	0.313 ± 0.002	0.312 ± 0.020
HSIC-NNM	2.439 ± 0.445	2.401 ± 0.367	0.291 ± 0.019	0.311 ± 0.069	0.602 ± 0.010	0.606 ± 0.028
PSM	7.188 ± 2.679	7.290 ± 3.389	0.292 ± 0.019	0.307 ± 0.053	0.607 ± 0.015	0.597 ± 0.021
k-NN	4.432 ± 2.345	4.303 ± 2.077	0.230 ± 0.016	0.262 ± 0.038	0.534 ± 0.008	0.573 ± 0.022
C. Forest	4.732 ± 2.974	4.095 ± 2.528	0.232 ± 0.018	0.224 ± 0.034	$\mathbf{0.306 \pm 0.000}$	0.305 ± 0.003
BNN	3.827 ± 2.044	4.874 ± 2.850	0.232 ± 0.008	0.240 ± 0.012	0.307 ± 0.001	0.309 ± 0.004
TARNet	0.729 ± 0.088	1.342 ± 0.597	0.228 ± 0.004	0.234 ± 0.012	0.314 ± 0.001	0.313 ± 0.002
CFR-MMD	0.663 ± 0.068	1.202 ± 0.550	$\mathbf{0.213 \pm 0.006}$	0.231 ± 0.009	0.312 ± 0.001	0.316 ± 0.003
CFR-WASS	0.649 ± 0.089	1.152 ± 0.527	0.225 ± 0.004	0.225 ± 0.010	0.308 ± 0.001	0.309 ± 0.003
SITE	0.604 ± 0.093	0.656 ± 0.108	0.224 ± 0.004	0.219 ± 0.009	0.309 ± 0.002	0.311 ± 0.004
ACE (Ours)	$\mathbf{0.489 \pm 0.046}$	$\mathbf{0.541 \pm 0.061}$	0.216 ± 0.005	$\mathbf{0.215 \pm 0.009}$	$\mathbf{0.306 \pm 0.000}$	$\mathbf{0.301 \pm 0.002}$

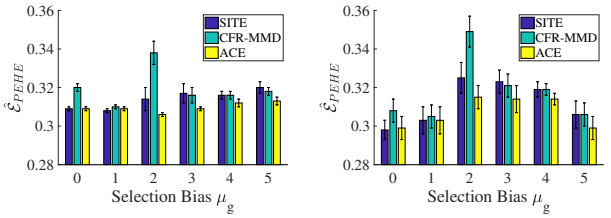


(a) IHDP (b) Jobs (c) Twins

Fig. 3: The effect of BAS regularization.



(a) Case 1: IHDP Within-Sample (b) Case 1: IHDP Out-of-Sample



(c) Case 2: Twins Within-Sample (d) Case 2: Twins Out-of-Sample

Fig. 4: Results on Datasets with different Selection Bias

2) Result Analysis: Fig. 4 reports the performance of the proposed method as well as the baseline methods (SITE, CFR-MMD) on IHDP and Twins datasets.

In Case 1, as shown in Fig. 4(a) and Fig. 4(b), in both within-sample and out-of-sample settings, the proposed method always performs the best for different selection bias. The observed results indicate that our proposed method is robust to different levels of selection bias. In Case 2, as shown in Fig. 4(c) and Fig. 4(d), it is observed that the performance of different methods varies a lot. The CFR-MMD and SITE methods are sensitive to the selection bias level. And the performance of ACE is much more stable under different levels of selection bias.

IV. Related Work

The existing methods of estimating the individual causal effect can be divided into five categories. (1) Regression-based models, such as double robust estimator [26], [27] and balancing linear regression (BLR) [10]. (2) Tree-based models, such as Bayesian additive regression trees (BART) [28], random forest [24], [29]. (3) Nearest neighbor based methods, such as k-NN [23], propensity score matching [18], and nearest neighbor matching through HSIC criteria [22]. (4) Multi-task learning based methods, such as multi-task neural network [30] and multi-task Gaussian process [31]. (5) Deep representation learning based methods. Feed-forward neural network and variational autoencoder have been adopted to learn the representation and estimate the counterfactuals [10]–[12], [25], [32].

The fifth category of deep representation learning methods usually perform better than other categories, as demonstrated by extensive evaluations. Our method ACE belongs to this category. In this category, except SITE [12], most of the methods ignore the similarity information when learning the

representation. Compared with existing similarity preserved method SITE, which only includes the selected triplets' information, ACE designs a more powerful regularization to achieve the following two expectations: (1) Precisely minimizing the discrepancy of control/treatment group, in order to make the two groups overlap as much as possible; (2) Adaptively retaining most of the important pairwise similarity according to the data location in the original space, which is the fine-grained similarity. With the help of the designed regularization, ACE achieves the state-of-the-art performance on causal effect estimation.

V. Conclusion

Estimating causal effect at the individual level is the base of causal inference. In this paper, we present a new approach for causal effect estimation by adaptively preserving similarity in representation learning. Different from the existing similarity-preserving based work, the proposed method ACE imposes the BAS regularization to fully explore the fine-grained similarity information in the original space and retain as much important similarity information as possible during the representation learning procedure. Extensive experiments on three benchmark datasets show that ACE consistently outperforms the state-of-the-art methods, which demonstrates the effectiveness of ACE in estimating the causal effect. Further experiments on the datasets with different levels of selection bias confirm that compared with existing methods, the BAS regularization makes ACE more robust to the selection bias.

Acknowledgements

This work was supported in part by the US National Science Foundation under grants NSF-IIS 1747614 and IIS-1514204. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] P. Wang, W. Sun, D. Yin, J. Yang, and Y. Chang, "Robust tree-based causal inference for complex ad effectiveness analysis," in *Proc. of International Conference on Web Search and Data Mining (WSDM'15)*, pp. 67–76, 2015.

[2] S. Li, N. Vlassis, J. Kawale, and Y. Fu, "Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI'16)*, pp. 3768–3774, 2016.

[3] M. Gangl, "Causal inference in sociological research," *Annual review of sociology*, vol. 36, pp. 21–47, 2010.

[4] W. Zhang, T. D. Le, L. Liu, Z.-H. Zhou, and J. Li, "Mining heterogeneous causal effects for personalized cancer treatment," *Bioinformatics*, vol. 33, no. 15, pp. 2372–2378, 2017.

[5] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[6] P. R. Rosenbaum, "Observational studies," in *Observational studies*, pp. 1–17, Springer, 2002.

[7] M. J. van der Laan and M. L. Petersen, "Causal effect models for realistic individualized treatment and intention to treat rules," *The International Journal of Biostatistics*, vol. 3, no. 1, 2007.

[8] K. Kuang, P. Cui, B. Li, M. Jiang, and S. Yang, "Estimating treatment effect in the wild via differentiated confounder balancing," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*, pp. 265–274, 2017.

[9] A. Alaa and M. van der Schaar, "Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design," in *Proc. of the International Conference on Machine Learning (ICML'18)*, pp. 129–138, 2018.

[10] F. D. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *Proc. of the International Conference on Machine Learning (ICML'16)*, pp. 3020–3029, 2016.

[11] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *Proc. of the International Conference on Machine Learning (ICML'17)*, pp. 3076–3085, 2017.

[12] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," in *Advances in Neural Information Processing Systems (NeurIPS'18)*, pp. 2638–2648, 2018.

[13] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[14] A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon, "Overlap in observational studies with high-dimensional covariates," *arXiv preprint arXiv:1711.02582*, 2017.

[15] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, vol. 29, no. 2, pp. 429–443, 1997.

[16] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet, *et al.*, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, vol. 6, pp. 1550–1599, 2012.

[17] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[18] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[19] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[22] Y. Chang and J. G. Dy, "Informative subspace learning for counterfactual inference," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI'17)*, pp. 1770–1776, 2017.

[23] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, "Nonparametric tests for treatment effect heterogeneity," *The Review of Economics and Statistics*, vol. 90, no. 3, pp. 389–405, 2008.

[24] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, no. just-accepted, 2017.

[25] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," *arXiv preprint arXiv:1705.08821*, 2017.

[26] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian, "Doubly robust estimation of causal effects," *American journal of epidemiology*, vol. 173, no. 7, pp. 761–767, 2011.

[27] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," in *Proc. of the International Conference on Machine Learning (ICML'11)*, pp. 1097–1104, 2011.

[28] H. A. Chipman, E. I. George, R. E. McCulloch, *et al.*, "Bart: Bayesian additive regression trees," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266–298, 2010.

[29] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7353–7360, 2016.

[30] M. v. d. S. Jinsung Yoon, James Jordan, "GANITE: Estimation of individualized treatment effects using generative adversarial nets," in *Proc. of International Conference on Learning Representations (ICLR'18)*, 2018.

[31] A. M. Alaa and M. van der Schaar, "Bayesian inference of individualized treatment effects using multi-task gaussian processes," in *Advances in Neural Information Processing Systems (NIPS'17)*, pp. 3427–3435, 2017.

[32] L. Yao, S. Li, Y. Li, H. Xue, J. Gao, and A. Zhang, "On the estimation of treatment effect with text covariates," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI'19)*, pp. 4106–4113, 2019.