Supplementary Material for Cycle-Balanced Representation Learning For Counterfactual Inference

1 Datasets

The following statements describe the three datasets in detail.

IHDP. The Infant Health and Development Program is a randomized clinical trial designed to evaluate the efficacy of a comprehensive early intervention, aiming to reduce the health problems of low birth weight, premature infants. Following [1] and keeping the same as early works [2, 3], we adopt the setting 'A' in [4] to simulate potential outcomes and make 100 replications. The treatment and control groups are imbalanced because a subset of the treatment groups is removed. This dataset includes 747 units (608 for control and 139 for treatment) and 25 covariates. These 25 covariates represent a child's weight, gender, and some information about parents like maternal age, education and employment.

Jobs. The Jobs dataset is created by [5, 6] for evaluating the effect of job training on income and employment status. It includes a randomized study from the National Supported Work program and observational data. The Lalonde randomized controlled experiment comprises 297 treated samples and 425 control samples. And the PSID observational group includes 2490 control records [6]. Each sample consists of 8 covariates that represent age, education, previous earnings, etc. The outcome is employment status.

Twins. The Twins dataset is created from all twins birth in the USA between 1989-1991 [7], and we only pay attention to the twins weighing less than 2kg and without missing features [8]. Each sample in the dataset records 30 covariates related to the parents, the pregnancy and the birth. We use the treatment $\mathcal{T} = 1$ as being the heavier twin, and $\mathcal{T} = 0$ is expressed as the lighter twin. The outcome is mortality after one year. The final dataset includes 11,400 pairs of twins. In order to simulate the selection bias, we follow the procedure in [3, 9]. We selectively choose one of the twins as the observation and hide the other: $t_i|x_i \sim Bern(Sigmoid(w^{\mathrm{T}}x + n))$, where $w^T \sim \mathcal{U}((-0.1, 0.1)^{30\times 1})$ and $n \sim \mathcal{N}(0, 01)$.

2 Notations

The notations in the paper are summarized in Table 1.

Table 1: Notations.

Symbol	Description
x_i	covariates of the i-th unit
t_i	treatment assignment for i-th unit
y_i or y_i^F	observed/factual outcome for i-th unit
y_i^{CF}	counterfactual outcome for i-th unit
x_t	raw data in treatment group
x_c	raw data in control group
$\hat{y}_i^F$	estimated factual outcome for i-th unit
$\hat{y}_i^{CF}$	estimated counterfactual outcome for i-th unit
n	number of units
p	dimension of raw data

3 Training and Optimization

This detailed training procedure of Algorithm 1 is corresponding to Section 3.4 in the paper.

Algorithm 1 : CBRE: Cycle-Balanced Representation Learning For Counterfactual Inference

Input: Factual sample (x_1, t_1, y_1),...,(x_n, t_n, y_n), representation network ϕ with parameter Θ_ϕ, discriminator f_D with parameter Θ_D, two decoders ψ_t and ψ_c with corresponding parameters Θ_{ψ_t} and Θ_{ψ_c}, outcome prediction network h with Θ_h; loss functions are $\mathcal{L}_D, \mathcal{L}_{rec}, \mathcal{L}_{cyc}, \mathcal{L}_p$.
Compute $u = \frac{1}{n}\sum_{i=1}^{n} t_i$ and $\omega_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}$
$\{\Theta_\phi, \Theta_D, \Theta_{\psi_t}, \Theta_{\psi_c}, \Theta_h\}$ = *Random Initialize*()
repeat
  Sample mini-batch $\{i_1, i_2, ..., i_m\} \subset \{1, 2, .., n\}$ and get batch dataset $\{\mathcal{X}_b, \mathcal{T}_b, \mathcal{Y}_b\}$
  Compute latent representations $z = \phi_{(\mathcal{X}_b)}$
  Draw noise $\boldsymbol{v} \sim \mathcal{N}(0, 1)$
  Set $\Theta_D \leftarrow \text{Adam}(\mathcal{L}_D(z, \boldsymbol{v}), \Theta_D)$
  Set $\Theta_\phi, \Theta_{\psi_t}, \Theta_{\psi_c} \leftarrow \text{Adam}(\mathcal{L}_{rec}(\mathcal{X}_b, \mathcal{T}_b), \Theta_\phi, \Theta_{\psi_t}, \Theta_{\psi_c})$
  Set $\Theta_\phi, \Theta_{\psi_t}, \Theta_{\psi_c} \leftarrow \text{Adam}(\mathcal{L}_{cyc}(\mathcal{X}_b, \mathcal{T}_b), \Theta_\phi, \Theta_{\psi_t}, \Theta_{\psi_c})$
  Set $\Theta_\phi, \Theta_h \leftarrow \text{Adam}(\mathcal{L}_p(\mathcal{X}_b, \mathcal{T}_b, \mathcal{Y}_b), \Theta_\phi, \Theta_h)$
until convergence
return $\{\Theta_\phi, \Theta_D, \Theta_{\psi_t}, \Theta_{\psi_c}, \Theta_h\}$

4 Hyper-parameter Optimization

Following Section 4.6 in the paper, we report the optimal hyper-parameters in Table 2.

Table 2: Optimal Hyper-parameter for three datasets.

Hyper-parameters	Datasets		
	IHDP	Jobs	Twins
α	0.5	0.5	1.0
β	1.0	1.0	1.0
γ	1.0	1.0	1.0
λ	1e-4	1e-4	1e-4
δ	10.0	10.0	10.0
Optimizer	Adam	Adam	Adam
Batch Size	80	130	200
Depth of encoder	5	5	5
Depth of discriminator	3	3	3
Depth of decoder ψ_t	5	5	5
Depth of decoder ψ_c	5	5	5
Depth of predictor	3	3	3
Dimension of encoder	200	200	200
Dimension of discriminator	200	200	200
Dimension of decoder ψ_t	200	200	200
Dimension of decoder ψ_c	200	200	200
Dimension of predictor	100	100	100
Learing rate	1e-3	1e-3	1e-3

References

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

[2] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.
[3] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.
[4] Vincent Dorie. non-parmetric causal inference models, 2016.
[5] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.
[6] Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353, 2005.
[7] Alexis Hannart, J Pearl, FEL Otto, P Naveau, and M Ghil. Causal counterfactual theory for the attribution of weather and climate-related events. *Bulletin of the American Meteorological Society*, 97(1):99–110, 2016.
[8] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.
[9] 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*, pages 1–26, 2021.