Transfer Learning for Individual Treatment Effect Estimation (Supplementary material)

Ahmed Aloui*[1] **Juncheng Dong***[1] **Cat P. Le**[1] **Vahid Tarokh**[1]

[1] Department of Electrical and Computer Engineering, Duke University

1 REPRODUCIBILITY STATEMENT

The supplementary material includes the implementation codes for our proposed framework, TARNet, and CITA.

2 CAUSAL INFERENCE: AN EXAMPLE

Let $X \in \mathcal{X}$ be the features (e.g., age, height, weight), the treatment assignment $A \in \{0, 1\}$ be the indicator representing if the subject received vaccine 0 or 1. The mortality outcome is denoted by $Y \in \mathcal{Y}$.

The main challenge of causal inference arises from the absence of counterfactual observations. We do not observe the outcomes of individuals upon receiving treatment 1 if they have received treatment 0 and vice versa. The subjects who received vaccine 1 may differ significantly from those who received treatment 0. This issue is called selection bias. For instance, older people are more likely to receive the treatment than young people). Thus, estimating the counterfactual effects is challenging due to the unbalance between the treatment groups.

Let $\hat{f}(x, a)$ be a hypothesis modeling the outcome for an individual x if he/she received treatment a. The factual loss is defined as follows:

$$\epsilon_F(\hat{f}) = \int_{\mathcal{X}\times\{C,B\}\times\mathcal{Y}} l_{\hat{f}}(x, a, y)\ p(x, a, y) dxdady \tag{1}$$

By Bayes rule, we can write the factual loss as

$$\begin{aligned}
&\epsilon_F(\hat{f}) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} l_{\hat{f}}(x, a = 0, y)\ p(x, y|A = 0)p(A = 0)dxdy + \\
&\int_{\mathcal{X}\times\mathcal{Y}} l_{\hat{f}}(x, a = 1, y)\ p(x, y|A = 1)p(A = 1)dxdy \\
&= p(A = 0) \int_{\mathcal{X}\times\mathcal{Y}} l_{\hat{f}}(x, a = 0, y)\ p(x, y|A = 0)dxdy + \\
&(1 - p(A = 0)) \int_{\mathcal{X}\times\mathcal{Y}} l_{\hat{f}}(x, a = 1, y)\ p(x, y|A = 1)dxdy \\
&= p(A = 0)\epsilon_F^{A=0}(\hat{f}) + (1 - p(A = 0))\, \epsilon_F^{A=0}(\hat{f})
\end{aligned}$$

We define the factual loss for the group who received vaccine 0 as follows:

*Equal Contribution.

Table 1: The settings to generate IHDP datasets

Dataset	μ	ω
IHDP (*Base*)	(0.6, 0.1, 0.1, 0.1, 0.1)	4
IHDP 1	(0.61, 0.09, 0.1, 0.1, 0.1)	4.1
IHDP 2	(0.62, 0.08, 0.1, 0.1, 0.1)	4.2
IHDP 3	(0.63, 0.07, 0.1, 0.1, 0.1)	4.3
IHDP 4	(0.64, 0.06, 0.1, 0.1, 0.1)	4.4
IHDP 5	(0.65, 0.05, 0.1, 0.1, 0.1)	4.5
IHDP 6	(0.66, 0.04, 0.1, 0.1, 0.1)	4.6
IHDP 7	(0.67, 0.03, 0.1, 0.1, 0.1)	4.7
IHDP 8	(0.68, 0.02, 0.1, 0.1, 0.1)	4.8
IHDP 9	(0.69, 0.01, 0.1, 0.1, 0.1)	4.9

$$\epsilon_F^{A=0}(\hat{f}) = \int_{\mathcal{X}\times\mathcal{Y}} l_{\hat{f}}(x, a=0, y)\; p(x, y|A=0)dxdy \tag{2}$$

Similarly, the factual loss for the group who received vaccine 1 is described as:

$$\epsilon_F^{A=1}(\hat{f}) = \int_{\mathcal{X}\times\mathcal{Y}} l_{\hat{f}}(x, a=1, y)\; p(x, y|A=1)dxdy \tag{3}$$

Consider a parallel universe where the treatment assignments are flipped (i.e., those who received vaccine 1 receive vaccine 0 and vice versa). The performance of our hypothesis $\hat{f}$ in this scenario is the counterfactual loss, defined as follows:

$$\epsilon_{CF}(\hat{f}) = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} l_{\hat{f}}(x, a, y)\; p(x, 1-a, y)dxdady \tag{4}$$

3 DATASETS AND EXPERIMENTS DESCRIPTIONS

3.1 DATASETS

IHDP The IHDP dataset was first introduced by Hill [2011] based on real covariates available from the Infant Health and Development Program (IHDP), studying the effect of development programs on children. The features in this dataset come from a Randomized Control Trial. The potential outcomes were simulated using Setting B. The dataset consists of 747 individuals (e.g., 139 in the treatment group and 608 in the control group), each with 25 features. The potential outcomes are generated as follows:

$$Y_0 \sim \mathcal{N}(\exp(\beta^T \cdot (X + W)), 1)$$

and

$$Y_1 \sim \mathcal{N}(\beta^T (X + W) - \omega, 1)$$

where W has the same dimension as X with all entries equal 0.5 and $\omega = 4$. The regression coefficient β, a vector of length 25, is randomly sampled from a categorical distribution with the support $(0, 0.1, 0.2, 0.3, 0.4)$ and the respective probabilities $\mu = (0.6, 0.1, 0.1, 0.1, 0.1)$. The dataset generated according to these parameters is referred to as the *base* dataset.

Additionally, we generate 9 additional datasets by introducing 9 new settings. These settings, which are constructed by varying μ and ω, are shown in Table 1. Each of these generated datasets consists of 747 individuals (e.g., 139 in the treatment group and 608 in the control group).

Jobs The Jobs dataset [LaLonde, 1986] consists of 619 observations. In this experiment, the causal inference task aims to learn the effect of participation in a specific professional training program on landing a job in the following three years. Here, we generate a family of related datasets by randomly reverting the original treatment assignments (i.e., $0 \leftrightarrow 1$) with the probability $p \in \{0 = 0/9, 1/9, 2/9, 3/9, 4/9, 5/9, \cdots, 9/9 = 1\}$. The dataset corresponding to $p = 0$ is considered the original dataset, and the dataset with $p = 1$ has all treatment assignments reversed. We select the original Jobs dataset, introduced in [LaLonde, 1986] as the *base* dataset for our experiments.

Twins The Twins dataset Louizos et al. [2017] is based on the collected birthday data of twins born in the United States from 1989 to 1991. It is assumed that twins share significant parts of their features. Consider the scenario where one of the twins was born heavier than the other as the treatment assignment. The outcome is whether the baby died in infancy (i.e., mortality). Here, the twins are divided into two groups: the treatment and the control groups. The treatment group consists of heavier babies from the twins. On the other hand, the control group consists of lighter babies from the twins. All given observations from this dataset are considered factual.

We first construct a *base* dataset by selecting a set of 2000 pairs of twins from the original dataset [Louizos et al., 2017]. Each individual is assigned to the treatment group according to a Bernoulli experiment with the probability of $q = 0.75$. In an analogous manner to that of the Jobs dataset, we generate a family of related datasets by randomly reverting the treatment assignments of the *base* dataset (i.e., $0 \leftrightarrow 1$) with corresponding probabilities $p \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, \cdots, 1\}$. For instance, to generate dataset $i = 1, 2, \cdots, 11$, we revert the individual treatment assignments in the base dataset using the Bernoulli experiment with the probability of $p_i = (i-1)/10$. In particular, $p = 0$ corresponds to the original dataset, while $p = 1$ corresponds to all treatment assignments reverted.

RKHS In this experiment, we generate 100 Reproducing Kernel Hilbert Space (RKHS) datasets, each having 2000 data points. Next, we generate the treatment and the control populations $X_1, X_0 \in \mathbb{R}^4$ respectively from Gaussian distributions $\mathcal{N}(\mu_1, I_4)$ and $\mathcal{N}(\mu_0, I_4)$ for each dataset. We sample $\mu_1 \in \mathbb{R}^4$ and $\mu_0 \in \mathbb{R}^4$ respectively according to Gaussian distributions $\mathcal{N}(\boldsymbol{e}, I_4)$ and $\mathcal{N}(-\boldsymbol{e}, I_4)$ where $\boldsymbol{e} = [1, 1, 1, 1]^T$.

Subsequently, we generate the potential outcome functions f_0 and f_1 with a Radial Basis Function (RBF) kernel $K(\cdot, \cdot)$, described as follows:

Let $\gamma_0, \gamma_1 \in \mathbb{R}^4$ be two vectors sampled from $\mathcal{N}(7\boldsymbol{e}, I_4)$ and $\mathcal{N}(9\boldsymbol{e}, I_4)$, respectively. Let $\lambda \in \mathbb{N}$ be sampled uniformly from $\{10, 11, \ldots, 99, 100\}$. For $j \in \{0, 1\}$:

1. We sample $m_j \in \mathbb{N}$ according to the Poisson distribution with parameter λ (i.e., Pois)
2. For every $i \in \{1, \ldots, m_j\}$, we sample x_j^i according to $\mathcal{N}(\gamma_j, I_4)$
3. The potential outcome functions $f_j, j = 0, 1$ are constructed as $f_j(\cdot) = \sum_{i=1}^{m_j} K(x_j^i, \cdot)$

Given the potential outcome functions $f_j, j \in \{0, 1\}$, the corresponding potential outcomes Y_0 and Y_1 are generated by:

$$Y_0(x) = f_0(x), \text{ for every } x \in \mathbb{R}^4,$$

and

$$Y_1(x) = f_1(x), \text{ for every } x \in \mathbb{R}^4.$$

We will refer to the first constructed dataset above as the *base* dataset. Here, all the generated potential outcome functions are in the same RKHS.

Heat (Physics) Consider a hot object left to cool off over time in a room with temperature $T(0)$. A person will likely suffer a burn if he/she touches the object at time u.

The causal inference task of interest is the effect of room temperature $T(0)$ on the probability of suffering a burn. This family consists of 20 datasets; each includes 4000 observations (e.g., 2000 in the control group and 2000 in the treatment group). The treatment in our setting is $a = 1$ when $T(0) = 5$, and $a = 0$ when $T(0) = 25$. The touching times of the treatment and control groups are sampled from two Chi-squared distributions $\chi^2(5)$ and $\chi^2(2)$, respectively, to introduce artificial bias.

From the solution to Newton's Heat Equation [Winterton, 1999], the underlying causal structure is governed by the following equation:

$$T(u) = C \cdot \exp(-ku) + T(0)$$

where $T(u)$ is the temperature at time u and C, k are constants. Let $T_0 = 25, C = 75$ for the control groups and $T_0 = 5, C = 95$ for the treatment groups in the datasets. We choose 20 values of $k = \{0.5, \cdots, 2\}$ uniformly spaced in $[0.5, 2]$. For each value of k, we generate a new dataset. The dataset corresponding to $k = 0.5$ is referred to as the *base* dataset.

Let $T^0(u)$ and $T^1(u)$ denote the temperature at time u for the control and treatment groups, respectively. The potential outcomes $Y_0(u)$ and $Y_1(u)$ corresponding to the probability of suffering a burn at time t for the control and treatment groups are described as follows:

$$Y_j(u) = \max\left(\frac{1}{75}(T^j(u) - 25), 0\right)$$

Movement (Physics) Consider a free-falling object encountering air resistance. Opening the parachute can change the air resistance and control the descent velocity. The causal inference task of interest is the effect of the air resistance (e.g., with $a = 1$ or without parachute $a = 0$) on the object's velocity at different times.

In this experiment, the family of datasets is generated, consisting of 12 datasets. Each dataset includes 4000 observations (e.g., 2000 in the treatment group and 2000 in the control group). The covariate is the time u. The outcome is the velocity at time u. The times of the treatment and control groups are sampled from two Chi-squared distributions $\chi^2(2)$ and $\chi^2(5)$, respectively, to create artificial bias.

The underlying causal structure is governed by an ordinary differential equation (ODE) with the following analytical solution describing the velocity of a person at time u:

$$v(u) = \frac{g}{C} + (v(0) - \frac{g}{C})e^{-Cu} \tag{5}$$

where $g = 10$ is the earth's gravitational constant, $C = k/m$, and m, k are the mass and the air resistance constant, respectively. We assume that $v(0) = 0$ corresponds to a free-falling object without initial velocity.

For the control group, $m = k = C = 1$ and the potential outcome is calculated as $Y_0(u) = v(u) = 10 - e^{-u}$. We use different sets of (m, k) to generate the treatment groups for each dataset. The values of (m, k) used in this experiment are as follows: $(5, 1)$, $(5, 5)$, $(5, 10)$, $(5, 20)$, $(10, 5)$, $(10, 10)$, $(10, 20)$, $(20, 5)$, $(20, 10)$, $(20, 20)$, $(50, 10)$, $(50, 20)$. The potential outcome function $Y_1(u)$ is calculated from Equation 5 with the values of m, k shown above. We choose the dataset corresponding to $(m, k) = (5, 1)$ as the *base* dataset.

3.1.1 Details of Experiments

In this paper, we first create a number of causal inference tasks from the above families of datasets. For each family of datasets (e.g., IHDP, Jobs, Twins), the **base** task is created from its *base* dataset. Similarly, we construct the other tasks from the remaining datasets in that family. In order to study the effects of transfer learning on causal inference, we define the source tasks and the target tasks as follows:

- In the first experiment in Section 6.3, we choose the *base* task to be the source task and the other tasks to be the target tasks.
- In the second experiment in Section 6.4, we choose the *base* task to be the target task and the other tasks to be the source tasks.

4 PROOFS OF THEOREMS

Theorem 4.1. Let $\hat{f}^S$ be a model trained on a source task, then

$$\epsilon_F^T(\hat{f}^S) + u\epsilon_{CF}^{T,a=0}(\hat{f}^S) \leq \varepsilon_{PEHE}^T(\hat{f}^S)$$

where $u = p_F^T(a = 1)$.

***Proof of Theorem 4.1*.** We have:

$$
\begin{aligned}
&\varepsilon_{PEHE}(\hat{f}^S) \\
&= \int_{\mathcal{X}} \left[(\hat{f}^S(x,1) - \hat{f}^S(x,0)) - (f^T(x,1) - f^T(x,0))\right]^2 \\
&\quad p_F^T(x)dx \\
&= \int_{\mathcal{X}} \left[(\hat{f}^S(x,1) - f^T(x,1)) - (f^T(x,0) - \hat{f}^S(x,0))\right]^2 \\
&\quad p_F^T(x)dx \\
&= \int_{\mathcal{X}} (\hat{f}^S(x,1) - f^T(x,1))^2 p_F(x)dx \\
&\quad + \int_{\mathcal{X}} (\hat{f}^S(x,0) - f^T(x,0))^2 p_F^T(x)dx \\
&\quad - 2\int_{\mathcal{X}} (\hat{f}^S(x,1) - f^T(x,1))(f^T(x,0) - \hat{f}^S(x,0)) \\
&\quad p_F^T(x)dx
\end{aligned}
\tag{6}
$$

First, we have the following properties of the factual and counterfactual distributions:

$$
\begin{aligned}
&1.\ \forall x \in \mathcal{X},\ p_F(x) = p_{CF}(x) \\
&2.\ \forall x \in \mathcal{X}, \forall a \in \{0,1\},\ p_F(x,a) = p_{CF}(x, 1-a)
\end{aligned}
$$

Applying these properties, the first term of Equation (6) can be expressed as:

$$
\begin{aligned}
&\int_{\mathcal{X}} (\hat{f}^S(x,0) - f^T(x,0))^2 p_F^T(x)dx \\
&= u\int_{\mathcal{X}} (\hat{f}^S(x,0) - f^T(x,0))^2 p_F^T(x|a=1)dx \\
&\quad + (1-u)\int_{\mathcal{X}} (\hat{f}^S(x,0) - f^T(x,0))^2 p_F^T(x|a=0)dx \\
&= u\int_{\mathcal{X}} (\hat{f}^S(x,0) - f^T(x,0))^2 p_{CF}^T(x|a=0)dx \\
&\quad + (1-u)\int_{\mathcal{X}} (\hat{f}^S(x,0) - f^T(x,0))^2 p_F^T(x|a=0)dx \\
&= u\epsilon_{CF}^{T,a=0}(\hat{f}^S) + (1-u)\ \epsilon_F^{T,a=0}(\hat{f}^S)
\end{aligned}
$$

Similarly, the second term of Equation (6) can be expressed as:

$$
\begin{aligned}
&\int_{\mathcal{X}} (\hat{f}^S(x,1) - f^T(x,1))^2 p_F^T(x)dx \\
&= (1-u)\epsilon_{CF}^{T,a=1}(\hat{f}^S) + u\ \epsilon_F^{T,a=1}(\hat{f}^S)
\end{aligned}
$$

The potential outcome is independent given the features $Y_1 \perp\!\!\!\perp Y_0 | X$ due to its unconfoundedness. Hence, the third term of Equation (6) can be expressed as:

$$
\begin{aligned}
&\mathbb{E}\big[(\hat{f}^S(X,1) - f^T(X,1))(f^T(X,0) - \hat{f}^S(X,0))\big] \\
&= \mathbb{E}_x\left[\mathbb{E}\left[\hat{f}^S(x,1) - Y_1^T)(Y_0^T - \hat{f}^S(x,0))|X=x\right]\right] \\
&= 0
\end{aligned}
$$

The factual and counterfactual losses of the treatment and control groups are positive. Thus, we have:

$$
\begin{aligned}
&u\epsilon_F^{T,a=1}(\hat{f}^S) + (1-u)\epsilon_F^{T,a=0}(\hat{f}^S) + u\epsilon_{CF}^{T,a=0}(\hat{f}^S) \\
&= \epsilon_F^T(\hat{f}^S) + u\epsilon_{CF}^{T,a=0}(\hat{f}^S) \\
&\leq \varepsilon_{PEHE}^T(\hat{f}^S)
\end{aligned}
$$

□

Theorem 4.2. For any hypothesis $\hat{f}$, we have:

$$
\begin{aligned}
\epsilon_{CF}^T(\hat{f}) \leq & \epsilon_F^S(\hat{f}) + V(p_F^T, p_F^S) + V(p_F^T, p_{CF}^T) \\
& + \mathbb{E}_{p_F^S}[|f^S(x,t) - f^T(x,t)|]
\end{aligned} \tag{7}
$$

and

$$
\begin{aligned}
\varepsilon_{PEHE}^T(\hat{f}) \leq & 4\epsilon_F^S(\hat{f}) + 4V(p_F^T, p_F^S) + 2V(p_F^T, p_{CF}^T) \\
& + 4\mathbb{E}_{p_F^S}[|f^S(x,a) - f^T(x,a)|]
\end{aligned} \tag{8}
$$

Proof of Theorem 4.2. Adapting the first theorem in Ben-David et al. [2010] to our setting, we have the following two inequalities:

$$\epsilon_{CF}^T(\hat{f}) \leq \epsilon_F^T(\hat{f}) + V(p_F^T, p_{CF}^T)$$

and

$$\epsilon_F^T(\hat{f}) \leq \epsilon_F^S(\hat{f}) + V(p_F^T, p_F^S) + \mathbb{E}_{p_F^S}[|f^S(x,a) - f^T(x,a)|]$$

Therefore, we have:

$$
\begin{aligned}
\epsilon_{CF}^T(\hat{f}) \leq & \epsilon_F^S(\hat{f}) + V(p_F^T, p_F^S) + V(p_F^T, p_{CF}^T) \\
& + \mathbb{E}_{p_F^S}[|f^S(x,a) - f^T(x,a)|]
\end{aligned}
$$

From Shalit et al. [2017], we have:

$$\varepsilon_{PEHE}^T(\hat{f}) \leq 2\epsilon_F^T(\hat{f}) + 2\epsilon_{CF}^T(\hat{f})$$

Therefore, we have:

$$
\begin{aligned}
\varepsilon_{PEHE}^T(\hat{f}) \leq & 4\epsilon_F^S(\hat{f}) + 4V(p_F^T, p_F^S) + 2V(p_F^T, p_{CF}^T) \\
& + 4\mathbb{E}_{p_F^S}[|f^S(x,a) - f^T(x,a)|]
\end{aligned}
$$

□

Theorem 4.3. Suppose that the function class G is stable under addition and multiplication and $\hat{f}, f^T \in G$, then

$$
\begin{aligned}
\epsilon_{CF}^T(\hat{f}) \leq & \epsilon_F^S(\hat{f}) + \underset{G}{\text{IPM}}(p_F^T, p_F^S) + \underset{G}{\text{IPM}}(p_F^T, p_{CF}^T) \\
& + \mathbb{E}_{p_F^S}[|f^S(x,a) - f^T(x,a)|]
\end{aligned} \tag{9}
$$

and

$$
\begin{aligned}
\varepsilon_{PEHE}^T(\hat{f}) \leq & 4\epsilon_F^S(\hat{f}) + 4\underset{G}{\text{IPM}}(p_F^T, p_F^S) + 2\underset{G}{\text{IPM}}(p_F^T, p_{CF}^T) \\
& + 4\mathbb{E}_{p_F^S}[|f^S(x,a) - f^T(x,a)|]
\end{aligned} \tag{10}
$$

Proof of Theorem 4.3. we have that:

$$
\begin{aligned}
\epsilon_{CF}^T(\hat{f}) \leq \epsilon_F^T(\hat{f}) + \| & \int (f^T(x,a) - \hat{f}(x,a))^2 \\
& (p_F^T(x,a) - p_{CF}^T(x,a)) da dx \| \\
\leq & \epsilon_F^T(\hat{f}) + \sup_{g \in G} \| \int g(x,a) \\
& (p_F^T(x,a) - p_{CF}^T(x,a)) da dx \|
\end{aligned}
$$

Hence, we have:

$$\epsilon^T_{CF}(\hat{f}) \leq \epsilon^T_F(\hat{f}) + \underset{G}{\text{IPM}}(p^T_F, p^T_{CF})$$

Similarly, we have:

$$\begin{aligned}
&\epsilon^T_F(\hat{f}) \\
&\leq \epsilon^S_F(\hat{f}) + \mathbb{E}_{p^S_F}[|f^S(x,a) - f^T(x,a)|] \\
&+ \| \int (f^S(x,a) - \hat{f}(x,a))^2 (p^S_F(x,a) - p^S_F(x,a)) dadx \| \\
&\leq \epsilon^T_F(\hat{f}) + \mathbb{E}_{p^S_F}[|f^S(x,a) - f^T(x,a)|] + \underset{G}{\text{IPM}}(p^T_F, p^S_F)
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
&\epsilon^T_F(\hat{f}) \\
&\leq \epsilon^S_F(\hat{f}) + \mathbb{E}_{p^S_F}[|f^S(x,a) - f^T(x,a)|] + \underset{G}{\text{IPM}}(p^T_F, p^S_F)
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}
\epsilon^T_{CF}(\hat{f}) \leq &\epsilon^S_F(\hat{f}) + \underset{G}{\text{IPM}}(p^T_F, p^S_F) + \underset{G}{\text{IPM}}(p^T_F, p^T_{CF}) \\
&+ \mathbb{E}_{p^S_F}[|f^S(x,a) - f^T(x,a)|]
\end{aligned}$$

From Shalit et al. [2017], we have:

$$\varepsilon^T_{PEHE}(\hat{f}) \leq 2\epsilon^T_F(\hat{f}) + 2\epsilon^T_{CF}(\hat{f})$$

Therefore, we have:

$$\begin{aligned}
\varepsilon^T_{PEHE}(\hat{f}) \leq &4\epsilon^S_F(\hat{f}) + 4\underset{G}{\text{IPM}}(p^T_F, p^S_F) + 2\underset{G}{\text{IPM}}(p^T_F, p^T_{CF}) \\
&+ 4\mathbb{E}_{p^S_F}[|f^S(x,a) - f^T(x,a)|]
\end{aligned}$$

□

Next, we will use the following results from Shalit et al. [2017] for causal inference. For $x \in \mathcal{X}, a \in \{0,1\}$, with notation simplicity, we define:

$$L^T_{\Phi,h}(x,a) = \int_Y l_{\Phi,h}(x,a,y) P(Y^T_a = y|x) dy.$$

Theorem 4.1 (Bounding The Counterfactual Loss). *Let Φ be an invertible representation with inverse Ψ. Let $p^{a=i}_{\Phi} = p_{\phi}(r|a = i), a \in \{0,1\}$ Let $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ be a hypothesis. Assume that for $a = 0,1$, the function $r \mapsto L_{\Phi,h}(\Psi(r), a) \in G$ then:*

$$\begin{aligned}
&\epsilon_{CF}(\Phi,h) \leq \\
&(1-u)\epsilon^{a=1}_F(\Phi,h) + a\epsilon^{a=0}_F(\Phi,h) + \\
&\underset{G}{\mathit{IPM}}\left(p^{a=1}_{\Phi}, p^{a=0}_{\Phi}\right).
\end{aligned} \tag{11}$$

Theorem 4.2 (Bounding the ϵ_{PEHE}). *The Expected Precision in Estimating Heterogeneous Treatment Effect ϵ_{PEHE} satisfies*

$$\begin{aligned}
&\varepsilon_{PEHE}(\Phi,h) \\
&\leq 2\left(\epsilon_{CF}(\Phi,h) + \epsilon_F(\Phi,h)\right) \\
&\leq 2\left(\epsilon^{a=0}_F(\Phi,h) + \epsilon^{a=1}_F(\Phi,h) + \underset{G}{\mathit{IPM}}\left(p^{a=1}_{\Phi}, p^{a=0}_{\Phi}\right)\right)
\end{aligned} \tag{12}$$

In the next section, the performance of target task $\epsilon^{T,a=0}_F(\Phi,h)$ is related to that of a source task $\epsilon^{S,a=0}_F(\Phi,h)$. Without loss of generality, we present the proof for the case when $a = 0$.

First, we make the following assumptions:

- **A1**: Φ is injective (Thus, $\Psi = \Phi^{-1}$ exists on Im(Φ)).

- **A2**: There exists a real function space G on $\mathrm{Im}(\Phi)$ such that the function $r \mapsto \ell^T_{\Phi,h}(\Psi(r), a, y) \in G$.
- **A3**: There exists a function class G' on $\mathcal{Y}$ such that $y \mapsto \ell_{\Phi,h}(x, a, y) \in G'$.

The measure of the fundamental difference between two causal inference tasks is defined as follows:

$$\gamma^* = \mathbb{E}_{x \sim P(X^S)} \left[\operatorname*{IPM}_{G'}(P(Y^S_a|x), P(Y^T_a|x)) \right]$$

Lemma 4.3. *Suppose that Assumptions 1-3 hold. The factual losses of any model* (Φ, h) *on source and target task satisfy for every* $a \in \{0, 1\}$

$$\begin{aligned}&\epsilon^{T,a}_F(\Phi, h) \leq \\ &\epsilon^{S,a}_F(\Phi, h) + \operatorname*{\mathit{IPM}}_{G}(P(\Phi(X^T_a)), P(\Phi(X^S_a))) + \gamma^*\end{aligned}$$

Proof of Lemma 4.3.

$$\begin{aligned}
&\epsilon^{T,a=0}_F(\Phi, h) - \epsilon^{S,a=0}_F(\Phi, h) \\
&= \int_{\mathcal{X}} L^T_{\Phi,h}(x, 0)P(X^T_0 = x) - L^S_{\Phi,h}(x, 0)P(X^S_0 = x)dx \\
&= \int_{\mathcal{X}} L^T_{\Phi,h}(x, 0)P(X^T_0 = x) - L^T_{\Phi,h}(x, 0)P(X^S_0 = x) \\
&+ L^T_{\Phi,h}(x, 0)P(X^S_0 = x) - L^S_{\Phi,h}(x, 0)P(X^S_0 = x)dx \\
&= \underbrace{\int_{\mathcal{X}} L^T_{\Phi,h}(x, 0)P(X^T_0 = x) - L^T_{\Phi,h}(x, 0)P(X^S_0 = x)dx}_{\Gamma} \\
&+ \underbrace{\int_{\mathcal{X}} \left(L^T_{\Phi,h}(x, 0) - L^S_{\Phi,h}(x, 0)\right) P(X^S_0 = x)dx}_{\Theta}
\end{aligned}$$

To bound Θ, we use the following inequality:

$$\begin{aligned}
&L^T_{\Phi,h}(x, t) - L^S_{\Phi,h}(x, t) \\
&= \int_Y \ell_{\Phi,h}(x, a, y) \left(P(Y^T_a = y|x) - P(Y^S_a = y|x)\right) dy \\
&\leq \max_{f \in G'} \left| \int_Y f(y)P(Y^T_a = y|x) - P(Y^S_a = y|x)dy \right| \\
&= \operatorname*{IPM}_{G'}\big(P(Y^T_a = y|x), P(Y^S_a = y|x)\big)
\end{aligned}$$

From the above inequality, we have:

$$\begin{aligned}
\Theta &= \int_{\mathcal{X}} \left(L^T_{\Phi,h}(x, 0) - L^S_{\Phi,h}(x, 0)\right) P(X^S_0 = x)dx \\
&\leq \mathbb{E}_{x \sim P(X^S)} \left[\operatorname*{IPM}_{G'}(P(Y^S_a|x), P(Y^T_a|x)) \right] \\
&= \gamma^*
\end{aligned}$$

To bound Γ, we use the change of variable formula:

$$\begin{aligned}\Gamma = &\int_{\mathcal{X}} L^T_{\Phi,h}(x,0)P(X^T_0 = x) - \\ &L^T_{\Phi,h}(x,0)P(X^S_0 = x)dx \\ = &\int_{\mathcal{R}} L^T_{\Phi,h}\big(\Psi(r),0\big)P\big(\Phi(X^T_0) = r\big) - \\ &L^T_{\Phi,h}\big(\Psi(r),0\big)P\big(\Phi(X^S_0) = r\big)dr \\ \le &\max_{g\in G}\Bigg|\int g(r)\Big(P\big(\Phi(X^T_0) = r\big) - \\ &P\big(\Phi(X^S_0) = r\big)\Big)dr\Bigg| \\ = &\,\underset{G}{\mathrm{IPM}}\Big(P\big(\Phi(X^T_0)\big), P\big(\Phi(X^S_0)\big)\Big)\end{aligned}$$

Combining the above upper bounds for Γ and Θ, we have:

$$\begin{aligned}&\epsilon^{T,a=0}_F(\Phi,h) - \epsilon^{S,a=0}_F(\Phi,h) \\ &\le \underset{G}{\mathrm{IPM}}\Big(P\big(\Phi(X^T_0)\big), P\big(\Phi(X^S_0)\big)\Big) + \gamma^*\end{aligned}$$

Thus, we conclude that:

$$\begin{aligned}&\epsilon^{T,a=0}_F(\Phi,h) \\ &\le \epsilon^{S,a=0}_F(\Phi,h) + \underset{G}{\mathrm{IPM}}\Big(P\big(\Phi(X^T_0)\big), P\big(\Phi(X^S_0)\big)\Big) + \gamma^*\end{aligned}$$

□

Lemma 4.4. Suppose that Assumptions A1, A2, A3 hold. Then the counterfactual loss of any model (Φ, h) on the target task satisfy:

$$\begin{aligned}\epsilon^T_{CF}(\Phi,h) \le &\epsilon^{S,a=1}_F(\Phi,h) + \epsilon^{S,a=0}_F(\Phi,h) \\ &+ \underset{G}{\mathrm{IPM}}(P(\Phi(X^T_1)), P(\Phi(X^S_1))) \\ &+ \underset{G}{\mathrm{IPM}}(P(\Phi(X^T_0)), P(\Phi(X^S_0))) \\ &+ \underset{G}{\mathrm{IPM}}(P(\Phi(X^T_0)), P(\Phi(X^T_1))) + 2\gamma^*\end{aligned}$$

where

$$\gamma^* = \underset{x\sim P(X^S)}{\mathbb{E}}\Big[\underset{G'}{\mathrm{IPM}}(P(Y^S_a|x), P(Y^T_a|x))\Big] \tag{13}$$

measures the fundamental difference between two causal inference tasks.

Proof of Lemma 4.4. Theorem 4.1 is applied to establish an upper bound for the counterfactual loss of the target task. Subsequently, we apply Lemma 4.3.

$$\begin{aligned}&\epsilon^T_{CF}(\Phi,h) \\ &\le \epsilon^{T,a=1}_F(\Phi,h) + \epsilon^{T,a=0}_F(\Phi,h) + \underset{G}{\mathrm{IPM}}\big(\Phi(X^T_0), \Phi(X^T_1)\big)\end{aligned}$$

Therefore,

$$\begin{aligned}\epsilon^T_{CF}(\Phi,h) \le &\epsilon^{S,a=1}_F(\Phi,h) + \epsilon^{S,a=0}_F(\Phi,h) + 2\gamma^* \\ &+ \underset{G}{\mathrm{IPM}}\Big(P\big(\Phi(X^T_1)\big), P\big(\Phi(X^S_1)\big)\Big) \\ &+ \underset{G}{\mathrm{IPM}}\Big(P\big(\Phi(X^T_0)\big), P\big(\Phi(X^S_0)\big)\Big) \\ &+ \underset{G}{\mathrm{IPM}}\Big(P\big(\Phi(X^T_0)\big), P\big(\Phi(X^T_1)\big)\Big)\end{aligned}$$

□

Theorem 4.5. (Transferability of Causal Knowledge) Suppose that Assumptions A1, A2, A3 hold. The performance of source model on target task, i.e. $\varepsilon^{T}_{PEHE}(\Phi, h)$, is upper bounded by:

$$\begin{aligned}\varepsilon^{T}_{PEHE}(\Phi, h) \leq & 2(\epsilon_F^{S,a=1}(\Phi, h) + \epsilon_F^{S,a=0}(\Phi, h) \\ & + \underset{G}{\text{IPM}}(P(\Phi(X_1^T)), P(\Phi(X_1^S))) \\ & + \underset{G}{\text{IPM}}(P(\Phi(X_0^T)), P(\Phi(X_0^S))) \\ & + \underset{G}{\text{IPM}}(P(\Phi(X_0^T)), P(\Phi(X_1^T)) + 2\gamma^*)\end{aligned}$$

Proof of Theorem 4.5. By applying Theorem 4.2, we get

$$\begin{aligned}& \varepsilon^{T}_{PEHE}(\Phi, h) \\ & \leq 2\Big(\epsilon_F^{T,a=0}(\Phi, h) + \epsilon_F^{T,a=1}(\Phi, h) \\ & + \underset{G}{\text{IPM}}\left(P\left(\Phi(X_0^T)\right), P\left(\Phi(X_1^T)\right)\right)\Big)\end{aligned}$$

After applying Lemma 4.3 to the first and second terms of the above equation, we have:

$$\begin{aligned}\varepsilon^{T}_{PEHE}(\Phi, h) \leq & \ 2\ (\epsilon_F^{S,a=1}(\Phi, h) + \epsilon_F^{S,a=0}(\Phi, h) \\ & + \underset{G}{\text{IPM}}(P(\Phi(X_1^T)), P(\Phi(X_1^S))) \\ & + \underset{G}{\text{IPM}}(P(\Phi(X_0^T)), P(\Phi(X_0^S))) \\ & + \underset{G}{\text{IPM}}(P(\Phi(X_0^T)), P(\Phi(X_1^T)) + 2\gamma^*)\end{aligned}$$

□

5 BASELINE: DATA BUNDLING

In many causal inference scenarios, we only have access to the trained model, and the corresponding data is unavailable. This situation could be the case in medical applications due to privacy reasons. Consequently, bundling the datasets of source tasks with the target task is not feasible. In contrast, the data may be available for some specific applications. In this case, we create another baseline referred to as data bundling.

In data bundling, we create the bundled dataset by combining the datasets of source tasks and the target task. Here, we compare our approach with data bundling for the IHDP and the Movement(Physics) datasets. For data bundling, we report the model's best performance (i.e., ε_{PEHE}) achieved by hyper-parameter search. For our approach, we only report the model's performance with the lowest training error. This setup gives more advantage to the data bundling baseline. The results are illustrated in Figure 1. Even with the aforementioned advantage, the data bundling method achieves poorer performance than our approach. This is due to data imbalance, lack of precision in determining similarity from propensity score, and **differences in outcome functions**.

6 CAUSAL INFERENCE TASK AFFINITY

Let $\mathcal{P}_{N_\theta}(T, D^{te}) \in [0, 1]$ be a function that measures the performance of a given model N_θ parameterized by $\theta \in \mathbb{R}^d$ on the test set D^{te} of the causal task T.

Definition 6.1 (ε-approximation Network). A model N_θ is called an ε-approximation network for a task-dataset pair (T, D) if it is trained using the training data D^{tr} such that $\mathcal{P}_{N_\theta}(T, D^{te}) \geq 1 - \varepsilon$, for a given $0 < \varepsilon < 1$.

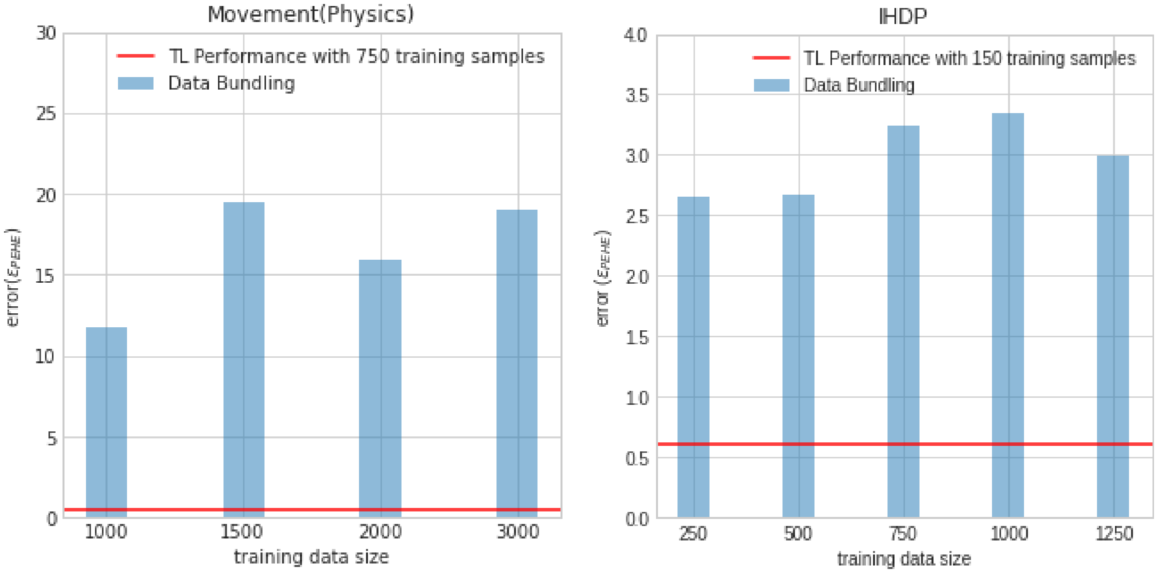

Figure 1: Performance comparison between data bundling and our approach. Our approach (red horizontal line) significantly outperforms data bundling. An increase in the size of training data doesn't improve the performance of data bundling.

Definition 6.2 (Fisher Information Matrix)**.** For a neural network N_{θ_s} with weights θ_s trained on data D_s, a given test dataset D_t and the negative log-likelihood loss function $L(\theta, D)$, the Fisher Information matrix is defined as:

$$F_{s,t} = \mathbb{E}_{D\sim D_t}\Big[\nabla_\theta L(\theta_s, D)\nabla_\theta L(\theta_s, D)^T\Big] \tag{14}$$

$$= -\mathbb{E}_{D\sim D_t}\Big[\mathbf{H}\big(L(\theta_s, D)\big)\Big], \tag{15}$$

where $\mathbf{H}$ is the Hessian matrix, i.e., $\mathbf{H}\big(L(\theta, D)\big) = \nabla^2_\theta L(\theta, D)$, and expectation is taken w.r.t the data. It is proven that the Fisher Information Matrix is asymptotically well-defined [Le et al., 2022]. In practice, we approximate the above with the empirical Fisher Information matrix:

$$\hat{F}_{s,t} = \frac{1}{|D_t|}\sum_{x\in D_t} \nabla_\theta L(\theta_s, x)\nabla_\theta L(\theta_s, x)^T. \tag{16}$$

Here, the empirical Fisher Information Matrix is positive semi-definite because it is the summation of positive semi-definite terms, regardless of the number of samples.

6.1 TASK AFFINITY BETWEEN COUNTERFACTUAL TASKS

In the following section, we denote the task-dataset pair $a = (T_a, D_a)$ by $a_F = (T_{a_F}, D_{a_F})$ where D_{a_F} is sampled from the factual distribution. Similarly, $a_{CF} = (T_{a_{CF}}, D_{a_{CF}})$ denotes the counterfactual task-dataset pair, where $D_{a_{CF}}$ is sampled from the counterfactual distribution. We refer to (T_{a_F}, D_{a_F}) and $(T_{a_{CF}}, D_{a_{CF}})$ as the corresponding factual and counterfactual tasks.

The following theorem proves that the order of proximity of tasks is preserved even if we observe the counterfactual tasks instead. In other words, a task, which is more similar to the target task when measured using factual data, remains more similar to the target task even when measured using counterfactual data.

Theorem 6.3. *Let $\mathbb{T}$ be the set of tasks and let $a_F = (T_{a_F}, D_{a_F})$, $b_F = (T_{b_F}, D_{b_F})$, and $c_F = (T_{c_F}, D_{c_F})$ be three factual tasks and $a_{CF} = (T_{a_{CF}}, D_{a_{CF}})$, $b_{CF} = (T_{b_{CF}}, D_{b_{CF}})$, and $c_{CF} = (T_{c_{CF}}, D_{c_{CF}})$ their corresponding counterfactual tasks.*

Suppose that there exists a class of neural networks (well-trained causal inference neural networks) $\mathcal{N} = \{N_\theta\}_{\theta\in\Theta}$ for *which:*

$$\forall a, b, c \in \mathbb{T},\ d[a, b] \leq d[a, c] + d[c, b] \tag{17}$$

and the task affinity between the factual and the counterfactual can be arbitrarily small, described as follows:

$$\forall \epsilon > 0, \exists N_\theta \in \mathcal{N},\ d[a_F, a_{CF}] < \epsilon \tag{18}$$

We have the following result:

$$d[a_F, b_F] \leq d[a_F, c_F] \implies d[a_{CF}, b_{CF}] \leq d[a_{CF}, c_{CF}] \tag{19}$$

Proof of Theorem 6.3. Suppose $d[a_F, b_F] \leq d[a_F, c_F]$. For every $\epsilon > 0$, we have:

$$\begin{aligned} d[a_{CF}, b_{CF}] &\leq d[a_{CF}, a_F] + d[a_F, b_F] + d[b_F, b_{CF}] \\ &\leq \epsilon + d[a_F, c_F] + \epsilon \\ &\leq d[a_F, a_{CF}] + d[a_{CF}, c_{CF}] + d[c_F, c_{CF}] \\ &\ + 2\epsilon \\ &\leq d[a_{CF}, c_{CF}] + 4\epsilon \end{aligned}$$

Therefore, $d[a_{CF}, b_{CF}] \leq d[a_{CF}, c_{CF}]$ as $\epsilon \to 0$. □

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

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

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

Cat P. Le, Mohammadreza Soltani, Juncheng Dong, and Vahid Tarokh. Fisher task distance and its application in neural architecture search. *IEEE Access*, 10:47235–47249, 2022. doi: 10.1109/ACCESS.2022.3171741.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

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.

RHS Winterton. Newton's law of cooling. *Contemporary Physics*, 40(3):205–212, 1999.