CAUSAL KNOWLEDGE TRANSFER FROM TASK AFFINITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent developments in deep representation models through counterfactual balancing have led to a promising framework for estimating Individual Treatment Effects (ITEs). While Randomized Control Trials are vital to the understanding of causal effects, they are sometimes infeasible, costly, or unethical to conduct. Here, we focus on transferring the causal knowledge acquired in prior experiments to new scenarios where only limited data is available. We first provide regret bounds on the counterfactual loss and ITE error of the target task indicating the transferability of causal knowledge. We also observe that the absolute values of ITEs are invariant under the action of the symmetric group on the labels of treatments. Given this invariance, we propose a symmetrized task distance for calculating the similarity of a target scenario with those encountered before. The aforementioned task distance is then used to transfer causal knowledge from the closest of all the available previously learned tasks to the target scenario. Empirical studies are provided for various datasets demonstrating that the proposed symmetrized task distance is strongly related to the estimation of the counterfactual loss. Our results indicate that transferring causal knowledge reduces the amount of required data by up to 95% when compared to training from scratch.

1 INTRODUCTION

One of the most remarkable characteristics of humans is their ability to transfer causal knowledge learned in a scenario to other *similar* situations. It is highly desirable for neural networks to have the same ability because of their numerous potential applications. For instance, mutations of old viruses often necessitate the development of new vaccines for treatment. To study the effect of new vaccine candidates, researchers need to collect data from randomized control trials, which is time-consuming and expensive (Kaur & Gupta, 2020). If the mutated viruses can be related to old ones by a measure of similarity, then the effects of vaccine candidates can be quickly calculated based on this similarity with a small amount of data collected for the new scenario. In other words, transfer learning methods can help the research on the effects of various treatments (e.g., applications in medicines, personal training, social policy) progress much faster (Ebbehoj et al., 2022).

Recently, there has been significant progress in transfer learning, especially in computer vision and natural language processing applications (Wang & Deng, 2018; Alyafeai et al., 2020; Pan & Yang, 2010; Zhuang et al., 2021). While this is very promising, a challenge for transferring causal knowledge arises from statistical learning models' vulnerability to non-causal correlations. For example, camels and horses often exist in images with different background colors, and a classifier may learn to use these colors to classify these objects (Arjovsky et al., 2019; Geirhos et al., 2019; Beery et al., 2018). A more critical challenge for transferring causal knowledge is that, in practice, the performance of the trained model for estimating ITEs can never be computed. This is because counterfactual data can never be collected as shown in Figure 1. This problem is known in the literature as *the fundamental problem of causal inference* (Rubin, 1974) and (Holland, 1986). For example, to compute the effect of vaccination on an individual at some given time, she/he must be both vaccinated and not be given the vaccine, which is obviously impossible. This contrasts with conventional supervised learning problems, where practitioners often use a separate validation set to estimate the true accuracy.

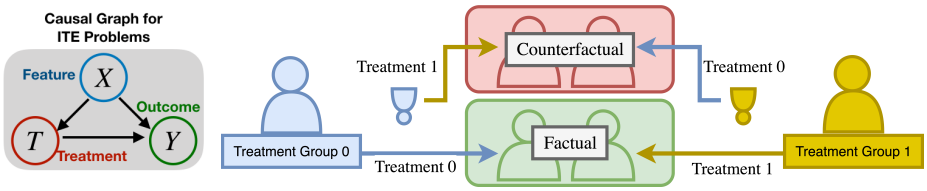

Figure 1: Inaccessibility to counterfactual data (which requires a "flipped" world where alternative treatments were assigned) makes transferring causal knowledge more challenging, since we do not have access to the validation dataset to estimate test performance to avoid the negative transfer.

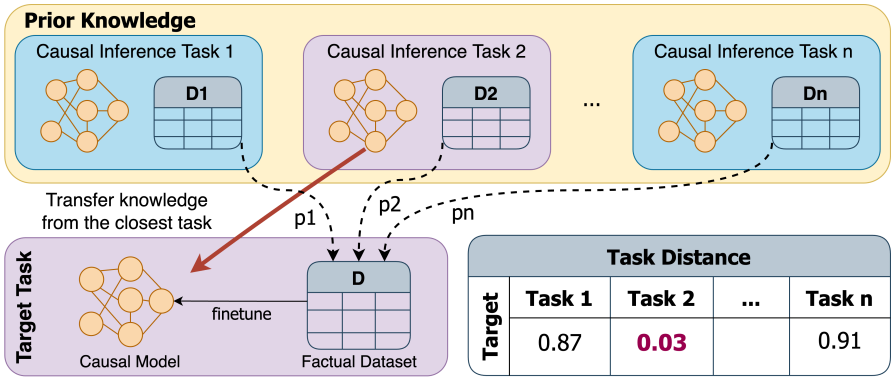

Figure 2: Overview of transfer learning in causal inference. The task distance is used to identify the closest task(s) from the set of prior tasks. The models and datasets from the relevant prior tasks are transferred to the target task.

The aforementioned challenges imply that much attention must be paid to selecting the appropriate source model to transfer from in causal knowledge transfer. Additionally, the similarity of scenarios must be calculated using a distance related to variations of counterfactual loss between scenarios. This motivates our work in this paper, where we propose a task distance between causal inference scenarios. The task distance is then used for transferring causal knowledge, as shown in Figure 2. Our contributions can be summarized as follows:

1. For causal transfer learning scenarios, we establish new (to the best of our knowledge) regret bounds for the learning of counterfactual outcomes and ITEs for target tasks. These bounds prove the feasibility of transferring causal knowledge.
2. We observe a special property (symmetry) of causal inference tasks. Specifically, the absolute value of ITEs must be invariant to relabeling the treatment groups under the action of the symmetric group. Subsequently, we propose an intuitively appealing symmetrized Fisher task distance for which this property holds. While we construct the proposed task distance to satisfy this property mathematically, we also provide empirical evidence that it successfully lends itself to this symmetry in Section 5.3.
3. We provide both theoretical (e.g., Theorem 4) and empirical evidence (e.g., Figure 3) supporting the relevance of the symmetrized Fisher task distance to transferring causal knowledge. Through extensive experiments, we demonstrate that the proposed task affinity is highly correlated with the loss in estimating counterfactuals (not measurable in practice).
4. We present a representative set of causal inference datasets *suitable for studying causal knowledge transfer*. Some of these are well-established datasets in the literature, while others are derived from known causal relations in social sciences, physics, and mathematics.
5. We provide empirical evidence based on the above datasets that our methods can compute the ITEs for the target task with significantly fewer (up to 95% reduction) data points compared to the case where transfer learning is not performed.

2 MATHEMATICAL BACKGROUND

We first establish the notation and briefly review the required mathematical background.

2.1 CAUSAL INFERENCE

Let $X \in \mathcal{X} \subset R^d$ be the covariates (features), $T \in \{0, \ldots, M\}$ be the treatment, and $Y \in \mathcal{Y} \subset \mathbb{R}$ be the factual (observed) outcome. For every $j \in \{0, \ldots, M\}$ we define Y_j to be the Potential Outcome that would have been observed if only treatment $T = j$, $j \in \{0, 1, \cdots, M\}$ was assigned. For example, in the medical context, X is the individual information (e.g. weight, heart rate, etc), T is the treatment assignment (e.g., $t = 0$ when the individual didn't receive a vaccine, and $t = 1$ where he/she did), Y is the outcome (e.g mortality data). A *causal inference dataset* is given by a set of factual observations $D_F = \{(x_i, t_i), y_i\}_{i=1}^N$, where N is the number of samples.

We present our results for $M = 1$ (binary case) in the sequel. However, our approach immediately applies to any positive integer $M < \infty$. In the binary case, the individuals who received $t = 0$ (respectively t=1) are denoted by the control group (respectively the treatment group).

Definition 1 (ITE)**.** *The Individual Treatment Effect also referred to as the Conditional Average Treatment Effect (CATE) (Imbens & Rubin, 2015), is defined as:*

$$\forall x \in \mathcal{X},\ \tau(x) = \mathbb{E}[Y_1 - Y_0 | X = x] \tag{1}$$

We assume that our data generation process respects overlap (i.e. $0 < p(t = 1|x) < 1$ for all $x \in \mathcal{X}$) and conditional unconfoundedness (i.e. $(Y^1, Y^0) \perp\!\!\!\perp T|X$) (Robins, 1987). These assumptions are sufficient conditions for the ITE to be identifiable (Imbens, 2004). We also assume that a true underlying function $f(x, t)$ describes the causal relationship. By definition $\tau(x) = f(x, 1) - f(x, 0)$. Let $\hat{f}(x, t)$ denote a *hypothesis* that estimates the true function $f(x, t)$. Thus, the ITE function can then be estimated as $\hat{\tau}(x) = \hat{f}(x, 0) - \hat{f}(x, 0)$. We let $l_{\hat{f}}(x, t, y)$ denote a loss function that quantifies the performance of $\hat{f}(\cdot, \cdot)$. A possible example is $l_{\hat{f}}(x, t, y) = (y - \hat{f}(x, t))^2$ (L^2 loss) .

Definition 2 (Factual Loss)**.** *For a hypothesis $\hat{f}$ and a corresponding loss function $l_{\hat{f}}$ we define the factual and counterfactual losses respectively as*

$$\epsilon_F(\hat{f}) = \int_{\mathcal{X} \times \{0,1\} \times \mathcal{Y}} l_{\hat{f}}(x, t, y)\ p(x, t, y) dx dt dy \tag{2}$$

We also define the factual loss for the treatment ($t = 1$) and control ($t = 0$) groups respectively as:

$$\epsilon_F^{t=1}(\hat{f}) = \int_{\mathcal{X} \times \mathcal{Y}} l_{\hat{f}}(x, 1, y)\ p(x, y|t = 1) dx dy \tag{3}$$

and

$$\epsilon_F^{t=0}(\hat{f}) = \int_{\mathcal{X} \times \mathcal{Y}} l_{\hat{f}}(x, 0, y)\ p(x, y|t = 0) dx dy \tag{4}$$

Definition 3 (Counterfactual Loss)**.** *The counterfactual loss is defined as (Shalit et al., 2016)*

$$\epsilon_{CF}(\hat{f}) = \int_{\mathcal{X} \times \{0,1\} \times \mathcal{Y}} l_{\hat{f}}(x, t, y)\ p(x, 1 - t, y) dx dt dy \tag{5}$$

Intuitively, the counterfactual loss corresponds to the expected loss value in a parallel universe where the roles of the control and treatment groups are exchanged.

Definition 4. *We define the Expected Precision in Estimating Heterogeneous Treatment Effect (PEHE) (Hill, 2011) as*

$$\varepsilon_{PEHE}(\hat{f}) = \int_{\mathcal{X}} \left(\hat{\tau}(x) - \tau(x)\right)^2 p(x) dx. \tag{6}$$

The value ε_{PEHE} is often used as the performance metric for estimation of ITEs as in (Shalit et al., 2016),(Hill, 2011), and (Johansson et al., 2016). Small factual and counterfactual losses are sufficient conditions for causal models to have good performance (i.e., low ε_{PEHE}) (Shalit et al.,

2016). Intuitively, this measures if a model has a good performance in predicting the effect both when the treatment is administered or not. Lower ε_{PEHE} also implies that the model is good for predicting the ITEs. We note that the above measures of performance are not directly accessible in causal inference scenarios, because the calculation of the ground truth ITE values requires access to counterfactual values. In this light, we may resort to selecting a hypothesis that optimizes an upper bound instead, such as the one given in the following section (see Equation 8).

2.2 TARNET AND COUNTERFACTUAL REGRESSION

TARNet (Shalit et al., 2016) has proven to be a successful framework for counterfactual balancing to estimate ITEs. It is defined as a pair of functions (Φ, h) where $\Phi : \mathbb{R}^d \to \mathbb{R}^l$ is a representation function of the features and $h : \mathbb{R}^l \times \{0, 1\} \to \mathbb{R}$ is a function learning the two potential outcomes functions in the representation space. The *hypothesis* learning the true causal function is: $\hat{f}(x, t) = h(\Phi(x), t)$. We denote the loss function $l_{\hat{f}}$ by $l_{(\Phi,h)}$. TARNet uses integral probability metric (IPM) defined as

$$\mathrm{IPM}_{\mathrm{G}}(p, q) := \sup_{g \in \mathrm{G}} \left| \int_{\mathcal{S}} g(s)(p(s) - q(s))ds \right|, \tag{7}$$

where the supremum is taken over a given class of functions G to measure the distance between distributions. It is a consequence of Kantorovich-Rubinstein duality Villani (2009) that IPM reduces to 1-Wassertein distance when G is the set of 1-Lipschtiz functions as is the case in our numerical experiments.

TARNet (Shalit et al., 2016) estimates the counterfactual outcomes by minimizing:

$$\mathcal{L}(\Phi, h) = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot l_{(\Phi,h)}(x_i, t_i, y_i) + \alpha \cdot \mathrm{IPM}_{\mathrm{G}} \left(\{\Phi(x_i)\}_{i:t_i=0}, \{\Phi(x_i)\}_{i:t_i=1} \right) \tag{8}$$

where $w_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}$, and $u = \frac{1}{N} \sum_{i=1}^{N} t_i$. The parameter α is referred to as the *balancing weight* since it controls the trade-off between the similarity of the representations in the latent domain, and the performance of the model on the factual data.

3 TRANSERABILITY OF CAUSAL KNOWLEDGE

In this section, we use superscripts Ta and Sr to denote quantities related to target and source task respectively. Suppose that we have a model (Φ^{Sr}, h^{Sr}) trained on a source causal inference task. We apply the source model to a different target task. For notational simplicity, we denote $P(\Phi(X)|T = t)$ by $P(\Phi(X_t))$ for $t \in \{0, 1\}$. We are interested in the performance of a well-trained source model when applied to a target task, i.e.

$$\epsilon^{Ta}_{PEHE}(\Phi^{Sr}, h^{Sr}) = \int_{x \in \mathcal{X}} \Big(\tau^{Ta}(x) - [h^{Sr}(\Phi^{Sr}(x), 1) - h^{Sr}(\Phi^{Sr}(x), 0)] \Big)^2 P(X^{Ta} = x)dx$$

where τ^{Ta} is the individual treatment effect function of the target, Φ is the representation learning function, and h is the potential outcomes hypothesis. While it is difficult to estimate, this error can have an upper bound that only involves obtainable quantities if we make reasonable assumptions about the relationship between the source and target task (defined in the Assumption 4 below). We make the following assumptions throughout this section:

1. **Assumption 1**: The loss function is non-negative, i.e. $\ell^{Ta}_{\Phi,h}(x, t, y) \geq 0$ for all $(x, t, y) \in (\mathcal{X} \times \{0, 1\} \times \mathcal{Y})$,
2. **Assumption 2**: Φ is injective (thus $\Psi = \Phi^{-1}$ exists on $\mathrm{Im}(\Phi)$) (We borrow this assumption from (Shalit et al., 2016)),
3. **Assumption 3**: There exists a real function space G on $\mathcal{R} = \mathrm{Im}(\Phi)$ and a constant B^{Ta}_{Φ} such that the function $r \mapsto \frac{1}{B^{Ta}_{\Phi}} \cdot \ell^{Ta}_{\Phi,h}(\Psi(r), t, y) \in G$.
4. **Assumption 4**: **Causal Knowledge Transferability Assumption**: There exists a function class G' on $\mathcal{Y}$ such that $y \mapsto l_{\Phi,h}(x, t, y) \in G'$ and $IPM_{G'}(P(Y^{Sr}_t|x), P(Y^{Ta}_t|x)) \leq \delta$ for $t \in \{0, 1\}$.

Note that the causal knowledge transferability assumption implies that the outcome distributions (causal effects) of treatment t in source and target tasks need to be similar in order for transfer learning to be beneficial.

Our main Theorem guarantees that causal knowledge can be transferred and is proved using two Lemmas that are stated below. These lemmas provide upper bounds on the factual and counterfactual losses for transferring causal knowledge and may be by themselves of independent interest. The proofs of these Lemmas and that of the Theorem are provided in the Appendix 8.5.

Lemma 1. *(Factual Loss of Source Model on Target Task) Suppose that Assumptions 1-4 hold. The factual losses of any model (Φ, h) on source and target task satisfy:*

$$\forall t \in \{0,1\},\ \epsilon_F^{Ta,t}(\Phi,h) \leq \epsilon_F^{Sr,t}(\Phi,h) + B_\Phi^{Ta} \cdot IPM_G(P(\Phi(X_t^{Ta})), P(\Phi(X_t^{Sr}))) + \delta$$

Lemma 2. *(Counterfactual Loss of Source Model on Target Task) Suppose that Assumptions 1-4 hold. The counterfactual losses of any model (Φ, h) on source and target task satisfy:*

$$\begin{aligned}\epsilon_{CF}^{Ta}(\Phi,h) \leq & \epsilon_F^{Sr,t=1}(\Phi,h) + \epsilon_F^{Sr,t=0}(\Phi,h^S) \\ & + B_\Phi^{Ta} \cdot IPM_G(P(\Phi(X_1^{Ta})), P(\Phi(X_1^{Sr}))) \\ & + B_\Phi^{Ta} \cdot IPM_G(P(\Phi(X_0^{Ta})), P(\Phi(X_0^{Sr}))) \\ & + B_\Phi^{Ta} \cdot IPM_G(P(\Phi(X_0^{Ta})), P(\Phi(X_1^{Ta}))) + 2\delta\end{aligned}$$

The above lemmas quantify the relationship between causality and transfer learning. In particular Lemma 2 bounds the inherently non-observable counterfactual loss by tractable quantities.

Theorem 1. *(Transferability of Causal Knowledge) Suppose that Assumptions 1-4 hold. The performance of source model on target task, i.e. $\epsilon_{PEHE}^{Ta}(\Phi^{Sr}, h^{Sr})$, is upper bounded by:*

$$\begin{aligned}\epsilon_{PEHE}^{Ta}(\Phi^{Sr},h^{Sr}) \leq & 2(\epsilon_F^{Sr,t=1}(\Phi^{Sr},h^{Sr}) + \epsilon_F^{Sr,t=0}(\Phi^{Sr},h^{Sr}) \\ & + B_{\Phi^{Sr}}^{Ta} \cdot IPM_G(P(\Phi^{Sr}(X_1^{Ta})), P(\Phi^{Sr}(X_1^{Sr}))) \\ & + B_{\Phi^{Sr}}^{Ta} \cdot IPM_G(P(\Phi^{Sr}(X_0^{Ta})), P(\Phi^{Sr}(X_0^{Sr}))) \\ & + B_{\Phi^{Sr}}^{Ta} \cdot IPM_G(P(\Phi^{Sr}(X_0^{Ta})), P(\Phi^{Sr}(X_1^{Ta})) + 2\delta)\end{aligned}$$

Theorem 1 implies that good performance on the target task is guaranteed if (1) the source model has a small factual loss (e.g., the first and second term in the upper bound) and (2) the distributions of the control and the treatment group features are similar in the latent domain (the rest three terms in the upper bound). This upper bound provides us with a sufficient condition for transfer learning in causal inference scenarios, indicating the transferability of causal knowledge. Please note that these regret bounds can be applied to any transfer learning framework that involves a pair of tasks.

4 Symmetrized Task Affinity for Causal Inference Tasks

While these regret bounds indicate the transferability of causal knowledge between any pair of causal inference tasks, they don't provide a constructive way to choose the best source task to transfer from, when multiple source tasks exist. The order of performance of different models and that of their upper bounds are not necessarily the same. Hence, we propose a label-invariant task affinity that finds the closest source task. Moreover, this task affinity satisfies the symmetry property (see section 4.3) of causal inference tasks. Our new task affinity is built on the Fisher task distance (FTD). We first give a brief introduction to FTD, then we propose a symmetrized Fisher task distance for causal inference tasks.

4.1 Task Representation

The ordered pair of a causal task $\mathcal{T}$ and its dataset $D = (X, T)$ will be denoted by $(\mathcal{T}, D)$, where dataset D itself consists of pair of covariates and their assigned treatments.

We will mathematically formalize a *sufficiently well-trained* deep network representing a causal task-dataset pair $(\mathcal{T}, D)$ in the Appendix 8.7. From now on, we assume that all the previously trained models are sufficiently well-trained networks.

4.2 FISHER TASK DISTANCE

Here, we recall the definition of the Fisher Information matrix for a neural network, and well-defined Fisher task distance (Achille et al., 2019; Le et al., 2021b; 2022b).

Definition 5 (Fisher Information Matrix). *For a neural network N_{θ_a} with weights θ_a trained on data D_a, a given test dataset D_b and the negative log-likelihood loss function $L(\theta, D)$, the Fisher Information matrix is defined as:*

$$F_{a,b} = \mathbb{E}_{D\sim D_b}\Big[\nabla_\theta L(\theta_a, D)\nabla_\theta L(\theta_a, D)^T\Big] = -\mathbb{E}_{D\sim D_b}\Big[\mathbf{H}\big(L(\theta_a, D)\big)\Big], \tag{9}$$

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 can proved that Fisher Information Matrix is asymptotically well-defined (Le et al., 2022b). In practice, we approximate the above with the empirical Fisher Information matrix:

$$\hat{F}_{a,b} = \frac{1}{|D_b|}\sum_{d\in D_b} \nabla_\theta L(\theta_a, d)\nabla_\theta L(\theta_a, d)^T. \tag{10}$$

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. For completeness, we next review the task affinity score (Le et al., 2021b).

Definition 6 (Task Affinity Score (TAS)). *Let $(\mathcal{T}_a, D_a)$ and $(\mathcal{T}_b, D_b)$ respectively denote the source and target task-dataset pairs. Let $D_a = D_a^{tr} \cup D_a^{te}$ (respectively $D_b = D_b^{tr} \cup D_b^{te}$) with D_a^{tr} (respectively D_b^{tr}) and D_a^{te} (respectively D_b^{te}) be the training and test sets of dataset D_a (respectively D_b), where the training for $\mathcal{T}_a$ is performed using the source representation network N_{θ_a}. Consider the Fisher information matrix $\mathbf{H}\big(L(\theta, D_a)\big)$ of N_{θ_a} with test data D_a^{te}. Let $F_{a,a}$ be the diagonal matrix of absolute values of elements of major diagonal of $\mathbf{H}\big(L(\theta, D_a)\big)$ normalized to have unit trace. Let $F_{a,b}$ be constructed in an analogous manner but using the training data D_b^{tr} (instead of D_a^{te}). The TAS from the source task $\mathcal{T}_a$ to the target task $\mathcal{T}_b$ is defined as:*

$$s[a, b] = \frac{1}{\sqrt{2}}\Big\|F_{a,a}^{1/2} - F_{a,b}^{1/2}\Big\|_F \tag{11}$$

It can be proved that $0 \leq \text{TAS} \leq 1$ where $\text{TAS} = 0$ denotes extreme similarity and $\text{TAS} = 1$ indicates extreme disimilarity. In Appendix 8.5, we prove under stringent assumptions that the order of TAS between candidate source tasks and the target task are preserved when a parallel universe experiment is performed in which the roles of the control and treatment groups are exchanged.

4.3 LABEL-INVARIANT TASK AFFINITY

Symmetry Property of Causal Inference Tasks Causal inference tasks can be considered as having multiple regression problems, one for each treatment group. Given a source task, if we alternate the treatment labels (i.e., 0 to 1 and 1 to 0), the treatment effect (i.e., $\mathbb{E}[Y_1 - Y_0|X]$) will be negated. Consequently, the unsymmetrized task distance (Le et al., 2022b) between the original task and the permuted task can be very large. However, the original model does not need to be retrained for transfer learning as we only need to permute the roles of output layers of the model to predict the individual treatment effects correctly for each group. In other words, the causal distance between these two permuted tasks must be zero. The following proposed label-invariant task affinity lends itself to this property of causal inference tasks.

Our causal inference tasks are represented by TARNet type networks. We also restrict to the case, where all causal tasks under consideration have the same number of treatment labels M (e.g., $M = 2$). Let $(\mathcal{T}_a, D_a)$ (respectively $(\mathcal{T}_b, D_b)$) with $D_a = (X_a, T_a, Y_a)$ (respectively $D_b = (X_b, T_b, Y_b)$) be the source (respectively target) causal inference tasks. Clearly $T_a, T_b \in \{0, 1, \ldots, M\}$.

Consider the symmetric group $\mathbb{S}_{M+1}$ consisting of all permutations of labels $\{0, 1, \ldots, M\}$. For $\sigma \in \mathbb{S}_{M+1}$, let $T_{\sigma(b)}$ denote the permutation of the target treatment labels under the action of σ. Let $d_\sigma = \frac{1}{\sqrt{2}}\Big\|F_{a,a}^{1/2} - F_{a,\sigma(b)}^{1/2}\Big\|_F$, then

$$s_{sym}[a, b] = \min_{\sigma\in\mathbb{S}_{M+1}} (d_\sigma)$$

is the *label-invariant task affinity distance* between causal tasks $\mathcal{T}_a$ and $\mathcal{T}_b$ (The pseudocode is provided in the Appendix 8.3). It follows from the above definition that the order of closeness of tasks under label-invariant task affinity closeness is robust to the architectural choice of the representation networks since task affinity distance has been shown to enjoy this property (Le et al., 2022b).

5 Experimental Results

We first describe the datasets we have used for our empirical studies. Subsequently, we present empirical results about quantifying the gains of transfer learning for causal inference, demonstrating the strong correlation between the proposed task distance and the counterfactual loss, and showing that the proposed task distance identifies the symmetries within causal inference tasks.

5.1 Causal Inference Datasets

We present a representative family of causal inference datasets suitable for studying causal knowledge transfer. Some of these are well-established datasets in the literature, while others are motivated by known causal structures in diverse areas such as social sciences, physics, health, and mathematics. Table 1 provides a brief description of the datasets used in our studies. A more detailed description is provided in Appendix 8.4.1. For each dataset, a number of corresponding causal inference tasks exist, which can be used to study transfer learning scenarios. Please note that we can only access the counterfactual data of the synthetic/semi-synthetic datasets (i.e., IHDP, RKHS, Movement, Heat). We are not in possession of the counterfactual data of the real-world datasets (i.e., Twins, Jobs).

Table 1: Causal inference datasets constructed for Transfer Learning Studies. Regression/Classification Problem (**REG/CLS**); Counterfactual Data Availability (**CF Avail**); Total Number of Tasks for the Dataset (**#Task**); Number of Samples in Each Task (**#Sample**).

Name	**Type**	**Task**	**CF Avail**	**Subject**	**#Task**	**#Sample**
IHDP	Semi-Synthetic	REG	YES	Health	100	747
Twins	Real-world	CLS	YES	Health	11	2000
Jobs	Real-world	CLS	NO	Social Sciences	10	619
RKHS	Synthetic	REG	YES	Mathematics	100	2000
Movement	Synthetic	REG	YES	Physics	12	4000
Heat	Synthetic	CLS	YES	Physics	20	4000

5.2 Comparison of Performance with/without Transfer Learning

Here we briefly discuss our experiments quantifying the impact of transferring causal knowledge on the size of required training data. In this experiment, we use Heat (Physics), Movement (Physics), IHDP, and RKHS datasets for which the counterfactual outcomes are available. We first fix a target causal inference task. For a wide range of balancing weights (α), we record the values of ε_{PEHE} for the training of the model from scratch while increasing the size of training datasets (at the end training process). In this process, the training datasets are slowly expanded such that smaller training sets are subsets of larger ones. We then report the minimum ε_{PEHE} achieved for each dataset size. For the Target task, we identify the closest source task and repeat the above process with a small amount of target task data. We then compare the performance with and without transfer learning to quantify the amount of data needed by transfer learning models to achieve the best possible performance without transferring causal knowledge. The results are summarized in Table 2, which demonstrates that transferring causal knowledge decreases the required amount of training data in this setting by a percentage between 75% and 95%.

5.3 Task Distance and Counterfactual Loss

Here, we show empirically the strong correlation between task distance (which only uses available data) and counterfactual loss (which is impossible to measure perfectly except for synthetic datasets). We show in Figure 3 that for different balancing weights α (see Equation 8), the correlation between

Table 2: The impact of causal knowledge transfer on the size of the required training dataset. Number of training data used without and with TL (**ORI/TL Size**); Minimum ε_{PEHE} achieved during training (not obtainable in practice because no validation data is available) (**W/O TL (Ideal)**); ε_{PEHE} of the **model without transfer learning** (model with minimum training loss)(**W/O TL (Prac)**); ε_{PEHE} of the model with transfer learning (model with minimum training loss)(**TL (Prac)**); Percentage of Reduction in Data provided by Causal Transfer Learning (**Gain**).

Dataset	ORI/TL Size	W/O TL(Ideal)	W/O TL (Prac)	TL (Prac)	Gain
IHDP	747/150	0.61	0.97	0.65	$> 80\%$
RKHS	2000/50	0.68	0.96	0.46	$> 95\%$
Movement	4000/750	0.021	0.025	0.011	$> 80\%$
Heat	4000/500	6.7e-6	1.4e-5	4.2e-6	$> 85\%$

task distance and counterfactual error on the IHDP, RKHS, Movement(Physics), and Heat(Physics) datasets for which counterfactuals are known. It is intuitively appealing and empirically observed that the task distance and the counterfactual loss have a strong correlation: the model of a source task has a smaller counterfactual loss on the target data if the target task is closer (in terms of the proposed task distance). Note that the points in Figure 3for different values of α (i.e., balancing weight) are extremely close. This shows that the proposed task affinity not only indicates counterfactual loss, but also is robust to change of hyper-parameters. This is a highly desirable property, especially in causal inference scenarios where no validation data can be accessed to cross-validate the hyper-parameters. Our numerical results for the Jobs and Twins datasets verify that the proposed task distance can capture the symmetries within causal inference problems. We flip treatment labels (0 and 1) with probability p (without any changes to the features and the outcomes) independently for each control and treatment data point. In Figure 4, we depict the trend of the symmetrized task distance between the original and the altered dataset by varying p, $p \in [0, 1]$. The symmetry of task distance is evident (with some deviation due to limited training data for calculating the task distance). The altered dataset with $p = 1$ is the closest to the original dataset (as it should be) since we have completely flipped the treatment assignments. The altered dataset with $p = 0.5$ is the furthest (as it should be), since we have randomly shuffled the control and the treatment groups. For all datasets, it can also be observed that the task distance trends are robust to variations in the balancing weight.

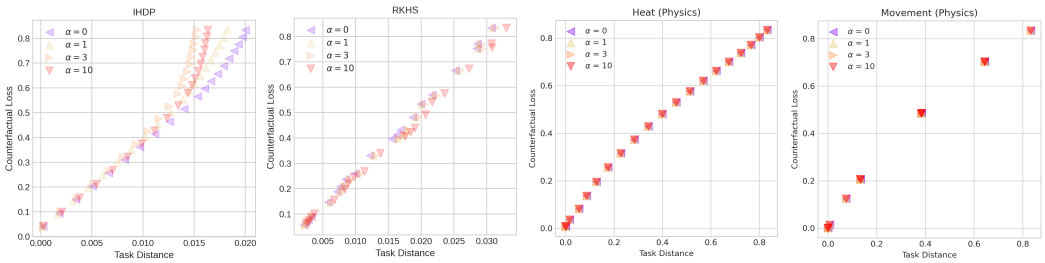

Figure 3: Task Distance vs. Counterfactual Error on causal inference datasets. The task distance shows a strong correlation with the (unmeasurable) counterfactual loss. Note that the various points for the Movement and Heat datasets are extremely close for different values of α.

6 Related Work

In the setting of transfer learning (Pan & Yang, 2010; Zhuang et al., 2021), prior learned models are used to increase the learning efficiency and decrease the required data. For instance, the parameters from a trained model may be used as initialization values for the target task. Many approaches in transfer learning (Thrun & Pratt, 1998; Blum & Mitchell, 1998; Silver & Bennett, 2008; Razavian et al., 2014; Finn et al., 2016; Fernando et al., 2017; Rusu et al., 2016) have been proposed, analyzed and applied in various machine learning applications. Transfer learning techniques inherently assume that prior knowledge in the selected source model helps learn a target task. In other words, these methods often do not consider the selection of the base task to perform knowledge transfer. Consequently, in some rare cases, transfer learning may even degrade the performance of

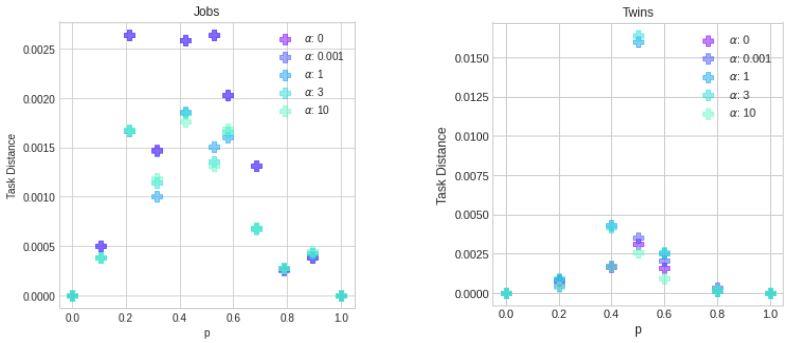

Figure 4: The symmetry of Task Distance. p (on the x-axis) denotes the probability of flipping treatment assignments of the original dataset. The altered dataset with $p = 1$ is the closest to the original task (as expected). The altered dataset with $p = 0.5$ is the furthest (as expected). Thus, the trend of points should resemble an inverted 'U', which is successfully captured by the task distance.

the model Standley et al. (2020). In order to avoid potential performance loss during knowledge transfer to a target task, *task affinity* (or task similarity) is considered as a selection method that identifies a group of closest base candidates from the set of the prior learned tasks. Task affinity has been recently investigated and applied to various domains, such as transfer learning (Zamir et al., 2018; Dwivedi & Roig, 2019; Achille et al., 2019; Wang et al., 2019), neural architecture search (Le et al., 2021a; 2022a; Le et al., 2021), few-shot learning (Pal & Balasubramanian, 2019; Le et al., 2022b), multi-task learning (Standley et al., 2020), and continual learning (Kirkpatrick et al., 2017; Chen et al., 2018). The related prior learned tasks are identified with similarity measures and then employed for knowledge transfer. Task affinity is inherently a non-commutative measure as it may be straightforward to transfer the knowledge from a more comprehensive task to a simpler task than the other way around (Le et al., 2021b).

While transfer learning and task affinity have been investigated in numerous application areas, their applications to causal inference have not been fully developed. Neyman-Rubin Causal Model and Pearl's Do-calculus are two popular frameworks for causal studies based on different perspectives. A central question in these frameworks is determining conditions for identifiability of causal quantities such as *Average* and *Individual Treatment Effects*. Past work considered estimators for Average Treatment Effect based on various methods such as Covariate Adjustment (a.k.a back-door adjustment) (Pearl, 2009; Rubin, 1978), weighting methods such as those utilizing propensity scores (Rosenbaum & Rubin, 1983), and Doubly Robust estimators (Funk et al., 2011). With the emergence of Machine Learning (ML) techniques, more recent approaches to causal inference include the applications of decision trees(Wager & Athey, 2015), Gaussian Processes (Alaa & van der Schaar, 2017) and Generative Modeling (Yoon et al., 2018) to ITE estimation. In particular, deep neural networks have successfully learned ITEs and estimated counterfactual outcomes by data balancing in the latent domain (Johansson et al., 2016; Shalit et al., 2016). It is important to note that the transportability of causal graphs is another closely related field that has been well-studied in the causality literature (Bareinboim & Pearl, 2021). It studies transferring knowledge of causal relationships in Pearl's do-calculus framework. In contrast, in this paper we are interested in transferring knowledge of Individual Treatment Effects from a source task to a target task in the Neyman-Rubin framework using representation learning.

7 Conclusion

In this paper, we provided theoretical analysis proving the transferability of causal knowledge and proposed a method for causal transfer learning based on a task affinity framework. To this end, we constructed a new task distance suitable for measuring the similarity of causal inference tasks. Given a new causal inference task, we transferred the causal knowledge from the closest available trained task. Extensive Simulations on a representative family of datasets provide empirical evidence demonstrating the gains of our method and the efficacy of the proposed symmetrized task distance. Reductions as much as 95% in the amount of required training data for new scenarios were observed.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task Embedding for Meta-Learning. *arXiv e-prints*, art. arXiv:1902.03545, Feb. 2019.

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6a508a60aa3bf9510ea6acb021c94b48-Paper.pdf.

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. A survey on transfer learning in natural language processing, 2020. URL https://arxiv.org/abs/2007.04239.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019. URL https://arxiv.org/abs/1907.02893.

Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):698–704, Sep. 2021. doi: 10.1609/aaai.v26i1.8232. URL https://ojs.aaai.org/index.php/AAAI/article/view/8232.

Sara Beery, Grant van Horn, and Pietro Perona. Recognition in terra incognita, 2018. URL https://arxiv.org/abs/1807.04975.

Avrim Blum and Tom. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT' 98*, 1998.

Shixing Chen, Caojin Zhang, and Ming Dong. Coupled end-to-end transfer learning with generalized fisher information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4329–4338, 2018.

Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12387–12396, 2019.

Andreas Ebbehoj, Mette Østergaard Thunbo, Ole Emil Andersen, Michala Vilstrup Glindtvad, and Adam Hulman. Transfer learning for non-image data in clinical research: A scoping review. *PLOS Digital Health*, 1(2):1–22, 02 2022. doi: 10.1371/journal.pdig.0000014. URL https://doi.org/10.1371/journal.pdig.0000014.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *CoRR*, abs/1701.08734, 2017. URL http://arxiv.org/abs/1701.08734.

Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Deep spatial autoencoders for visuomotor learning. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pp. 512–519. IEEE, 2016.

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *Am J Epidemiol*, 173(7):761–767, March 2011.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162. URL https://doi.org/10.1198/jcgs.2010.08162.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459. URL http://www.jstor.org/stable/2289064.

Guido W. Imbens. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics*, 86(1):4–29, 02 2004. ISSN 0034-6535. doi: 10.1162/003465304323023651. URL https://doi.org/10.1162/003465304323023651.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

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

Simran Preet Kaur and Vandana Gupta. Covid-19 vaccine: A comprehensive status report. *Virus Research*, 288:198114, 2020. ISSN 0168-1702. doi: https://doi.org/10.1016/j.virusres.2020.198114. URL https://www.sciencedirect.com/science/article/pii/S0168170220310212.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986. ISSN 00028282. URL http://www.jstor.org/stable/1806062.

Cat P. Le, Mohammadreza Soltani, Robert Ravier, and Vahid Tarokh. Improved Automated Machine Learning from Transfer Learning. *arXiv e-prints*, art. arXiv:2103.00241, Feb. 2021.

Cat P. Le, Mohammadreza Soltani, Robert Ravier, and Vahid Tarokh. Task-aware neural architecture search. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4090–4094, 2021a. doi: 10.1109/ICASSP39728.2021.9414412.

Cat P Le, Mohammadreza Soltani, Robert Ravier, and Vahid Tarokh. Task-aware neural architecture search. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4090–4094. IEEE, 2021b.

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, 2022a. doi: 10.1109/ACCESS.2022.3171741.

Cat Phuoc Le, Juncheng Dong, Mohammadreza Soltani, and Vahid Tarokh. Task affinity with maximum bipartite matching in few-shot learning. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=u2GZOiUTbt.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models, 2017. URL https://arxiv.org/abs/1705.08821.

Arghya Pal and Vineeth N Balasubramanian. Zero-shot task transfer, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '14, pp. 512–519, Washington, DC, USA, 2014. IEEE Computer Society. ISBN 978-1-4799-4308-1. doi: 10.1109/CVPRW.2014.131. URL http://dx.doi.org.stanford.idm.oclc.org/10.1109/CVPRW.2014.131.

James M. Robins. Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics With Applications*, 14:923–945, 1987.

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. ISSN 00063444. URL http://www.jstor.org/stable/2335942.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 1974. URL http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf.

Donald B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34 – 58, 1978. doi: 10.1214/aos/1176344064. URL https://doi.org/10.1214/aos/1176344064.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *CoRR*, abs/1606.04671, 2016. URL http://arxiv.org/abs/1606.04671.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, 2016. URL https://arxiv.org/abs/1606.03976.

Daniel L Silver and Kristin P Bennett. Guest editor's introduction: special issue on inductive transfer learning. *Machine Learning*, 73(3):215–220, 2008.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/standley20a.html.

Sebastian Thrun and Lorien Y. Pratt. Learning to learn: Introduction and overview. In *Learning to Learn*, 1998.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests, 2015. URL https://arxiv.org/abs/1510.04342.

Aria Y Wang, Leila Wehbe, and Michael J Tarr. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *BioRxiv*, pp. 708016, 2019.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, oct 2018. doi: 10.1016/j.neucom.2018.05.083. URL https://doi.org/10.1016%2Fj.neucom.2018.05.083.

R. H. S. Winterton. Newton's law of cooling. *Contemporary Physics*, 40(3):205–212, 1999. doi: 10.1080/001075199181549. URL https://doi.org/10.1080/001075199181549.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByKWUeWA-.

Amir R Zamir, Alexander Sax, William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76, 2021. doi: 10.1109/JPROC.2020.3004555.

8 APPENDIX

Here, we provide a simple example to help understand causal inference dataset, the pseudocode, the datasets description, the theorems, the proofs for those theorems, and other supplementary materials.

8.1 REPRODUCIBILITY STATEMENT

In the supplementary material, we have included our codes that implement TARNet and the proposed task distance.

8.2 CAUSAL INFERENCE: AN EXAMPLE

Let $X \in \mathcal{X}$ be the features (e.g., age, sex, etc..), the treatment variable $T \in \{A, B\}$ be the indicator representing if the subject received vaccine A or B, and $Y \in \mathcal{Y}$ indicates the mortality outcome.
The main challenge of causal inference arises from the absence of counterfactual observations. We do not observe the outcomes of individuals upon receiving treatment A if they have received treatment B (and vice versa). The subjects who received vaccine A may be significantly different from those who received treatment B. This is commonly called selection bias (e.g. elderly people are more likely to receive treatment A than young people). Thus estimating the counterfactual effects is challenging due to this unbalance between the treatment groups.
Let $\hat{f}(x, t)$ be a hypothesis modeling the outcome for an individual x if he received treatment t. The factual loss is defined as

$$\epsilon_F(\hat{f}) = \int_{\mathcal{X}\times\{A,B\}\times\mathcal{Y}} l_{\hat{f}}(x, t, y)\ p(x, t, y)dxdtdy \tag{12}$$

By Bayes rule, we can write the factual loss as

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

Where we define the factual loss for the group who received vaccine A to be

$$\epsilon_F^{t=A}(\hat{f}) = \int_{\mathcal{X}\times\mathcal{Y}} l_{\hat{f}}(x, A, y)\ p(x, y|t = A)dxdy \tag{13}$$

Respectively, the factual loss for the group who received vaccine B is

$$\epsilon_F^{t=B}(\hat{f}) = \int_{\mathcal{X}\times\mathcal{Y}} l_{\hat{f}}(x, B, y)\ p(x, y|t = B)dxdy \tag{14}$$

Let us now consider a parallel universe where the treatment assignments are flipped (those who received vaccine A receive vaccine B and vice versa). The performance of our hypothesis $\hat{f}$ in this parallel universe is the counterfactual loss, defined as:

$$\epsilon_{CF}(\hat{f}) = \int_{\mathcal{X}\times\{A,B\}\times\mathcal{Y}} l_{\hat{f}}(x, t, y)\ p(x, 1 - t, y)dxdtdy \tag{15}$$

8.3 PESUDOCODE FOR SYMMETRIZED TASK DISTANCE

The pseudocode for our proposed task affinity is given in Algorithm 1.

Algorithm 1: Label-Invariant Task Affinity Score for Causal Inference

Data: Source tasks: $S = \{(X_1, T_1, Y_1), \ldots, (X_m, T_m, Y_m)\}$, Target task: (X_t, T_t, Y_t)
Input: TARNet models $N_{\theta_1}, N_{\theta_2}, \ldots, N_{\theta_m}$
Output: TARNet model for the target task t

Function $\texttt{TAS}(X_a, T_a, X_b, T_b, N_{\theta_a})$**:**
  Compute $F_{a,a}$ using N_{θ_a} with X_a, T_a
  Compute $F_{a,b}$ using N_{θ_a} with X_b, T_b
  return $s[a,b] = \frac{1}{\sqrt{2}} \left\| F_{a,a}^{1/2} - F_{a,b}^{1/2} \right\|_F$

Function `Main`**:**
  ▷ Find the closest tasks in S
  for $i = 1, 2, \ldots, m$ **do**
    Train N_{θ_i} for source task i using (X_i, T_i, Y_i)
    Compute the distance from source task i to target task t:
      $s_i^+ = \texttt{TAS}(X_i, T_i, X_t, T_t, N_{\theta_i})$
    Compute the distance from source task i to target task t' where t''s treatments are inverted treatments of t:
      $s_i^- = \texttt{TAS}(X_i, T_i, X_t, 1 - T_t, N_{\theta_i})$
    Symmetrized distance: $s_{sym_i} = \min(s_i^+, s_i^-)$
  return closest tasks: $i^* = \underset{i}{\operatorname{argmin}}\ s_{sym_i}$
  ▷ Causal Knowledge Transfer
  Finetune $N_{\theta_i^*}$ with the target task's data (X_t, T_t, Y_t)
return $N_{\theta_i^*}$

8.4 DATASETS AND EXPERIMENTS DESCRIPTIONS

8.4.1 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, and the potential outcomes were simulated using Setting "B" in (Hill, 2011), hence the word semi-synthetic. The dataset consists of 747 individuals (139 in the treatment group and 608 in the control group), each with 25 features. Hill generated the potential outcomes with $Y_0 \sim \mathcal{N}(\exp(\beta^T \cdot (X+W)), 1)$, where W has the same dimension as X with all values $= 0.5$ and $Y_1 \sim \mathcal{N}(\beta^T (X+W) - \omega, 1)$ with $\omega = 4$. β is 25-element vector of regression coefficients randomly sampled from a categorical distribution with support $(0, 0.1, 0.2, 0.3, 0.4)$ and respective probabilities $\mu = (0.6, 0.1, 0.1, 0.1, 0.1)$. We refer to the dataset generated according to these parameters as the *base* dataset.

We retain the base dataset and introduce 9 new settings according to Table 3 by varying μ and ω. We also generate 10 new datasets for each setting, each consisting of 747 individuals (139 in the treatment group and 608 in the control group) by running the same process but with different random samples of the aforementioned Gaussian distribution.

Jobs The original Jobs dataset (LaLonde, 1986) has 619 observations. The causal inference task is to learn the effect of participation/lack of participation in a specific professional training program (corresponding to receiving a treatment $t = 1$) at a time on the success in landing a job in the following three years. We generate a family of related datasets by randomly reverting the treatment assignments of the original dataset with various probabilities $p \in [0, 1]$. Specifically, to generate a dataset, we first choose a probability value $p \in [0, 1]$, and then alter individuals (original) treatment assignment (i.e., $0 \leftrightarrow 1$) with probability p. We choose values

Table 3: 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

$p \in \{0 = 0/9, 1/9, 2/9, 3/9, 4/9, 5/9, \cdots, 9/9 = 1\}$. Clearly, $p = 0$ corresponds to the original dataset, and $p = 1$ corresponds to all reverted treatment assignments. We choose the original Jobs dataset (LaLonde, 1986) as the *base* dataset for our experiments, as discussed in Section 8.4.2.

Twins The Twins dataset was first introduced by Louizos et al. (2017) based on the collected data about twins' births in the United States from 1989 to 1991. It is assumed that twins share significant parts of their features. We consider whether one of the twins was born heavier than the other as the treatment assignment and if he/she died in infancy (mortality) as the outcome. We divide the twins into two groups: In the treatment (respectively control) group, we consider the outcome for the heavier (respectively lighter) twin as factual. In both groups, the outcome for the remaining twin is assumed to be counterfactual.

We first construct a *base* dataset by selecting a set of 2000 pairs of twins from the original dataset (Louizos et al., 2017). Then, each element is assigned to the treatment group according to a Bernoulli experiment with the probability of success $q = 0.75$.

Next, the *base* dataset is used to generate more datasets. 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 ($0 \leftrightarrow 1$) 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 let $p_i = (i-1)/10$ revert the individual treatment assignments in the base dataset Bernoulli experiment with probability of success p_i. Clearly, $p = 0$ corresponds to the original dataset while $p = 1$ corresponds to all treatment assignments reverted.

RKHS We generate 100 Reproducing Kernel Hilbert Space (RKHS) datasets, each having 2000 data points. For each dataset, we start by generating 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 $\mathcal{N}(\mu_0, I_4)$. 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 $\boldsymbol{e} = [1, 1, 1, 1]^T$ is the all ones vector.

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

Let $\gamma_0, \gamma_1 \in \mathbb{R}^4$ be two vectors sampled respectively from $\mathcal{N}(7\boldsymbol{e}, I_4)$ and $\mathcal{N}(9\boldsymbol{e}, I_4)$, and 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 Pois(λ) (e.g., the Poisson distribution with parameter λ),
2. For every $i \in \{1, \ldots, m_j\}$, we sample x_j^i according to $\mathcal{N}(\gamma_j, I_4)$, and
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 in the above as the *base* dataset.

Note that in the above, 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 is likely to 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 with 2000 in each control and treatment group. The treatment in our setting is $t = 1$ when $T_0 = 5$, and $t = 0$ when $T_0 = 25$.

The treatment and control groups touching times are respectively sampled from two Chi-squared distributions $\chi^2(5)$ and $\chi^2(2)$ (intentionally in order to create artificial bias).

From the solution to Newton's Heat Equation (Winterton, 1999) the underlying causal structure is governed by the equation

$$T(u) = C \cdot \exp(-ku) + T_0$$

where $T(u)$ is the temperature at time u and $C, k > 0$ are constants.

We let $T_0 = 25$ and $C = 75$ for all the control groups in the datasets. Similarly, we let $T_0 = 5$ and $C = 95$ for all 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 $T^1(u)$ respectively denote the temperature at time u for the control and treatment groups. The potential outcomes $Y_0(u)$ and $Y_1(u)$ corresponding to the probability of suffering a burn at time t for respectively the control and treatment groups are given by

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

for $j \in \{0, 1\}$.

Movement (Physics) Consider a falling person in the air encountering air resistance. Opening her/his parachute can change the air resistance and control its descent velocity. The causal inference task of interest is the effect of the air resistance (e.g., with $t = 1$ or without parachute $t = 0$) on the object's velocity at different times.

This family consists of 12 datasets. Each includes 4000 observations with 2000 in each treatment and control group. Here, the covariate is the time u, and the outcome is the velocity at time u. The treatment and control groups' times are respectively sampled from two Chi-squared distributions $\chi^2(2)$ and $\chi^2(5)$ (intentionally in order 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:

$$v(u) = \frac{g}{C} + (v_0 - \frac{g}{C})e^{-Cu} \tag{16}$$

where $C = \frac{k}{m}$, m and k are respectively the mass, and the air resistance constant, and $g = 10$ is the gravitational constant of earth. In the above $v_0 = v(0)$ is the initial velocity at time $u = 0$. We assume $v_0 = 0$, corresponding to a free fall without initial velocity.

For the control group, we assume $m = k = C = 1$ and the potential outcome is calculated as $Y_0(u) = v(u) = 10 - e^{-u}$ using the Equation 16. For the treatment groups, we vary m and k for different datasets with $(5, 1)$, $(5, 5)$, $(5, 10)$, $(5, 20)$, $(10, 5)$, $(10, 10)$, $(10, 20)$, $(20, 5)$, $(20, 10)$, $(20, 20)$, $(50, 10)$, $(50, 20)$. The potential outcomes $Y_1(u)$ is calculated from Equation 16. We have chosen the the dataset corresponding to $(m, k) = (5, 1)$ as the *base* dataset.

8.4.2 DETAILS OF EXPERIMENTS

In this paper, we first create various 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 5.3, 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 5.2, we choose the *base* task to be the target task and the other tasks to be the source tasks.

8.5 PROOF OF LEMMAS AND THEOREMS

We will use the following known results (Shalit et al., 2016) for causal inference. The proofs for these results are given in (Shalit et al., 2016) .

For $x \in \mathcal{X}, t \in \{0,1\}$, with notational simplicity, we define

$$L^{Ta}_{\Phi,h}(x,t) = \int_Y l_{h,\Phi}(x,t,y) P(Y^{Ta}_t = y|x) dy.$$

Theorem 2 (Bounding The Counterfactual Loss)**.** *Let Φ be an invertible representation with inverse Ψ. Let $p^{t=i}_\Phi = p_\phi(r|t=i), i \in \{0,1\}$ Let $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ be a hypothesis.*
Assume that there exists a constant $B_\Phi > 0$ such that for $t = 0, 1$, the function $g_{\Phi,h}(r,t) := \frac{1}{B_\Phi} \cdot L_{h,\Phi}(\Psi(r),t) \in G$. Here, we have

$$\epsilon_{CF}(h,\Phi) \leq (1-u)\epsilon^{t=1}_F(h,\Phi) + u\epsilon^{t=0}_F(h,\Phi) + B_\Phi \cdot IPM_{\mathrm{G}}\left(p^{t=1}_\Phi, p^{t=0}_\Phi\right). \tag{17}$$

Theorem 3 (Bounding the ϵ_{PEHE})**.** *The Expected Precision in Estimating Heterogeneous Treatment Effect ϵ_{PEHE} satisfies*

$$\begin{aligned} \epsilon_{PEHE}(h,\Phi) &\leq 2\left(\epsilon_{CF}(h,\Phi) + \epsilon_F(h,\Phi) - 2\sigma^2_Y\right) \\ &\leq 2\left(\epsilon^{t=0}_F(h,\Phi) + \epsilon^{t=1}_F(h,\Phi) + B_\Phi IPM_{\mathrm{G}}\left(p^{t=1}_\Phi, p^{t=0}_\Phi\right) - 2\sigma^2_Y\right). \end{aligned} \tag{18}$$

Next we relate the performance of target task $\epsilon^{Ta,t=0}_F(h,\Phi)$ to that of a source task $\epsilon^{Sr,t=0}_F(h,\Phi)$. Without loss of generality, we present the proof for the case $t = 0$.

We make the following assumptions throughout the sequel.

1. **Assumption 1**: The loss function is non-negative, i.e. $\ell^{Ta}_{\Phi,h}(x,t,y) \geq 0$ for all $(x,t,y) \in (\mathcal{X} \times \{0,1\} \times \mathcal{Y})$,
2. **Assumption 2**: Φ is injective (thus $\Psi = \Phi^{-1}$ exists on $\mathrm{Im}(\Phi)$) (Shalit et al., 2016),
3. **Assumption 3**: There exists a real function space G on $\mathcal{R} = \mathrm{Im}(\Phi)$ and a constant B^{Ta}_Φ such that the function $r \mapsto \frac{1}{B^{Ta}_\Phi} \cdot \ell^{Ta}_{\Phi,h}(\Psi(r),t,y) \in G$.
4. **Assumption 4**: **Causal Knowledge Transferability Assumption**: There exists a function class G' on $\mathcal{Y}$ such that $y \mapsto l_{\Phi,h}(x,t,y) \in G'$ and $IPM_{G'}(P(Y^{Sr}_t|x), P(Y^{Ta}_t|x)) \leq \delta$ for $t \in \{0,1\}$.

Proof of Lemma 1

$$
\begin{aligned}
&\epsilon_F^{Ta,t=0}(\Phi,h) - \epsilon_F^{Sr,t=0}(\Phi,h) \\
&= \int_{\mathcal{X}} L_{\Phi,h}^{Ta}(x,0)P(X_0^{Ta}=x) - L_{\Phi,h}^{Sr}(x,0)P(X_0^{Sr}=x)dx \\
&= \int_{\mathcal{X}} L_{\Phi,h}^{Ta}(x,0)P(X_0^{Ta}=x) - L_{\Phi,h}^{Ta}(x,0)P(X_0^{Sr}=x) + L_{\Phi,h}^{Ta}(x,0)P(X_0^{Sr}=x) \\
&\quad - L_{\Phi,h}^{Sr}(x,0)P(X_0^{Sr}=x)dx \\
&= \underbrace{\int_{\mathcal{X}} L_{\Phi,h}^{Ta}(x,0)P(X_0^{Ta}=x) - L_{\Phi,h}^{Ta}(x,0)P(X_0^{Sr}=x)dx}_{\Gamma} \\
&\qquad + \underbrace{\int_{\mathcal{X}} \left(L_{\Phi,h}^{Ta}(x,0) - L_{\Phi,h}^{Sr}(x,0)\right) P(X_0^{Sr}=x)dx}_{\Theta}
\end{aligned}
$$

We next upper bound Θ and Γ.
To bound Θ, we use the following inequality:

$$
\begin{aligned}
L_{\Phi,h}^{Ta}(x,t) - L_{\Phi,h}^{Sr}(x,t) &= \int_Y \ell_{\Phi,h}(x,t,y)\left(P(Y_t^{Ta}=y|x) - P(Y_t^{Sr}=y|x)\right)dy \\
&\leq max_{f\in G'}\left|\int_Y f(y)P(Y_t^{Ta}=y|x) - P(Y_t^{Sr}=y|x)dy\right| \\
&= IPM_{G'}\big(P(Y_t^{Ta}=y|x), P(Y_t^{Sr}=y|x)\big) \leq \delta
\end{aligned}
$$

With the above inequality:

$$
\begin{aligned}
\Theta &= \int_{\mathcal{X}} \left(L_{\Phi,h}^{Ta}(x,0) - L_{\Phi,h}^{Sr}(x,0)\right) P(X_0^{Sr}=x)dx \\
&\leq \int_{\mathcal{X}} \delta P(X_0^{Sr}=x)dx = \delta \int_{\mathcal{X}} P(X_0^{Sr}=x)dx = \delta
\end{aligned}
$$

To bound Γ, we use the change of variable formula

$$
\begin{aligned}
\Gamma &= \int_{\mathcal{R}} L_{\Phi,h}^{Ta}(x,0)P(X_0^{Ta}=x) - L_{\Phi,h}^{Ta}(x,0)P(X_0^{Sr}=x)dx \\
&= \int_{\mathcal{X}} L_{\Phi,h}^{Ta}\big(\Psi(r),0\big)P\big(\Phi(X_0^{Ta})=r\big) - L_{\Phi,h}^{Ta}\big(\Psi(r),0\big)P\big(\Phi(X_0^{Sr})=r\big)dr \\
&\leq B_\Phi^{Ta}\cdot max_{g\in G}\left|\int_{\mathcal{R}} g(r)\Big(P\big(\Phi(X_0^{Ta})=r\big) - P\big(\Phi(X_0^{Sr})=r\big)\Big)dr\right| \\
&= B_\Phi^{Ta}\cdot IPM_G\Big(P\big(\Phi(X_0^{Ta})\big), P\big(\Phi(X_0^{Sr})\big)\Big)
\end{aligned}
$$

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

$$
\epsilon_F^{Ta,t=0}(\Phi,h) - \epsilon_F^{Sr,t=0}(\Phi,h) \leq B_\Phi^{Ta}\cdot IPM_G\Big(P\big(\Phi(X_0^{Ta})\big), P\big(\Phi(X_0^{Sr})\big)\Big) + \delta.
$$

We conclude that

$$
\epsilon_F^{Ta,t=0}(\Phi,h) \leq \epsilon_F^{Sr,t=0}(\Phi,h) + B_\Phi^{Ta}\cdot IPM_G\Big(P\big(\Phi(X_0^{Ta})\big), P\big(\Phi(X_0^{Sr})\big)\Big) + \delta.
$$

This concludes the proof.

Proof of Lemma 2 We apply Theorem 2 to establish an upper bound for the counterfactual loss of the target task and subsequently apply Lemma 1 .

$$\epsilon_{CF}^{Ta}(h,\Phi) \leq \epsilon_F^{Ta,t=1}(h,\Phi) + \epsilon_F^{Ta,t=0}(h,\Phi) + B_\Phi^{Ta} IPM_{\mathrm{G}}\big(\Phi(X_0^{Ta}), \Phi(X_1^{Ta})\big)$$

Therefore,

$$\epsilon_{CF}^{Ta}(h,\Phi) \leq \epsilon_F^{Sr,t=1}(\Phi,h) + \epsilon_F^{Sr,t=0}(\Phi^{Sr},h^S) + 2\delta + B_{\Phi^{Sr}}^{T} \cdot IPM_{\mathrm{G}}\Big(P\big(\Phi(X_1^{Ta})\big), P\big(\Phi(X_1^{Sr})\big)\Big)$$
$$+ B_{\Phi^{Sr}}^{T} \cdot IPM_{\mathrm{G}}\Big(P\big(\Phi(X_0^{Ta})\big), P\big(\Phi(X_0^{Sr})\big)\Big) + B_{\Phi^{Sr}}^{T} \cdot IPM_{\mathrm{G}}\Big(P\big(\Phi^S(X_0^{Ta})\big), P\big(\Phi^{Sr}(X_1^{Ta})\big)\Big)$$

This concludes the proof.

Proof of Theorem 1 By applying Theorem 3, we get

$$\epsilon_{PEHE}^{Ta}(h,\Phi) \leq 2\left(\epsilon_F^{Ta,t=0}(h,\Phi) + \epsilon_F^{Ta,t=1}(h,\Phi) + B_\Phi^{Ta} IPM_{\mathrm{G}}\left(P(\Phi(X_0^{Ta})), P(\Phi(X_1^{Ta}))\right)\right)$$

After applying Lemma 1 to the first and second terms in the above:

$$\epsilon_{PEHE}^{Ta}(\Phi^S,h^S) \leq 2\Big(\epsilon_F^{Sr,t=0}(\Phi,h) + B_\Phi^{Ta} \cdot IPM_{\mathrm{G}}\big(P(X_0^{Ta}), P(X_0^{Sr})\big) + \delta + \epsilon_F^{Sr,t=1}(\Phi,h)$$
$$+ B_\Phi^{Ta} \cdot IPM_{\mathrm{G}}\big(P(X_1^{Ta}), P(X_1^{Sr})\big) + \delta + IPM_{\mathrm{G}}\Big(P\big(\Phi^S(X_0^{T})\big), P\big(\Phi^S(X_1^{T})\big)\Big)$$

Hence,

$$\epsilon_{PEHE}^{Ta}(\Phi^S,h^S) \leq 2\Big(\epsilon_F^{S,t=1}(\Phi^S,h^S) + \epsilon_F^{S,t=0}(\Phi^S,h^S) + B_{\Phi^{Sr}}^{T} \cdot IPM_{\mathrm{G}}\Big(P\big(\Phi(X_1^{Ta})\big), P\big(\Phi(X_1^{Sr})\big)\Big)$$
$$+ B_{\Phi^{Sr}}^{T} \cdot IPM_{\mathrm{G}}\Big(P\big(\Phi(X_0^{Ta})\big), P\big(\Phi(X_0^{Sr})\big)\Big) + B_{\Phi^{S}}^{T} \cdot IPM_{\mathrm{G}}\Big(P\big(\Phi^S(X_0^{T})\big), P\big(\Phi^S(X_1^{T})\big)\Big) + 2\delta\Big).$$

This concludes the proof.

8.6 Baseline: Data Bundling

8.6.1 Transfer Learning Scenario: Different Potential Outcomes

In many causal inference scenarios, we only have access to the trained model and the corresponding data is not available. For instance, in medical applications, this could be the case due to privacy reason. Consequently, bundling the datasets of source tasks is not feasible. In contrast, for some specific applications, the data may be available. 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 dataset of the target task. Below, we compare our approach with data bundling for the IHDP and the Movement(Physics) datasets. For data bunding, we report the model best performance, i.e. ε_{PEHE}, achieved by hyperparameter search. For our approach, we only report the performance of the model with the lowest training error. This gives more advantage to data bundling baseline. The results are summarized in Figure 5. Even with the aforementioned advantage, the data bundling method has poor performance. This may be due to data imbalance, lack of precision in determining similarity from propensity score, and **differences in outcome functions**.

8.6.2 Same Potential Outcomes, Different Propensity Scores

We compare the effectiveness of data bundling with that of transfer learning in scenarios where only the propensity scores are changing across tasks (i.e., same potential outcome functions). Figure 6 provides a summary of our findings. In this experiment, we generate synthetic source and target datasets, each having 1000 data points. For the source dataset, we generate the treatment X_1^{Sr} and the control X_0^{Sr} populations respectively from two Gaussian distributions $\mathcal{N}((0,0), I_2)$ and $\mathcal{N}((5,5), 2{\cdot}I_2)$. Subsequently, we build the target dataset by adding noise to the source task samples. Specifically, we add standard Gaussian noise to the i^{th} sample of the source task in order to generate the i^{th} sample of the target task, i.e., $x_i^{Ta} = x_i^{Sr} + \epsilon_i$ where x_i^{Ta} and x_i^{Sr} are respectively the i^{th} samples of the target task and the source task, and $\epsilon_i \sim \mathcal{N}((0,0), I_2)$ is the additive noise. For every sample $i \in \{1,..,1000\}$, we assign the treatment labels as follows:

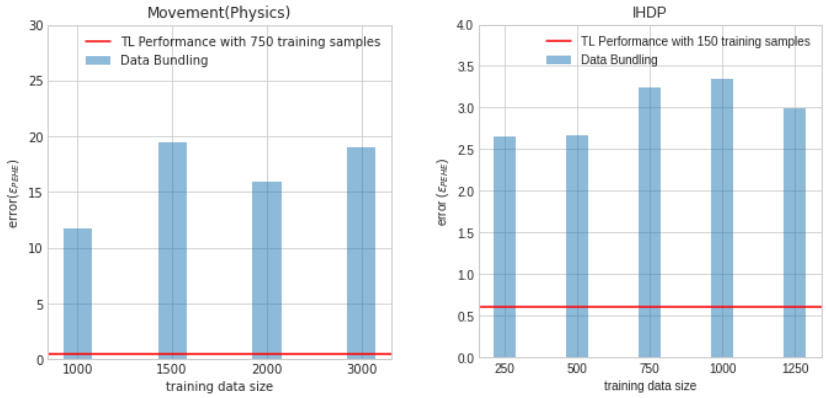

Figure 5: Performance comparison between data bundling and our approach. Our approach (red horizontal line) significantly outperforms data bundling. An increase in the size of training data doesn't improve the performance of data bundling.

- If the i^{th} sample of the source task is in the treatment group, then the corresponding i^{th} sample of the target task is also in the treatment group.
- If the i^{th} sample of the source task is in the control group, then the corresponding i^{th} sample of the target task is in the treatment group with probability $p = 0.1$, and in the control group with $p = 0.9$.

Subsequently, the output is defined based on the potential outcome functions f_0 and f_1 as follows:

$$f_0(x) = 0.5 \times e^{(-0.005 B_0^T x)},$$

and

$$f_1(x) = 0.5 \times e^{(-0.03 B_1^T x)} + 5,$$

where $B_0, B_1 \in \mathbb{R}^2$ and their components are respectively sampled from $\mathcal{N}(0, 1)$ and $\mathcal{N}(4, 1)$. Please note that these are only sampled once, and these parameters are shared between the source and the target tasks. Our experiments (see Figure 6) suggest that even when only the propensity scores are different, transfer learning has better performance than bundling the source and the target datasets together.

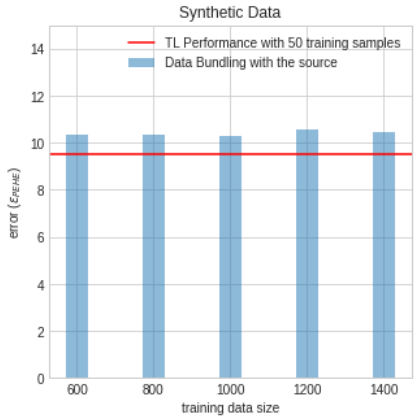

Figure 6: Performance comparison between data bundling and our approach. Here, only the propensity scores are changed across the two tasks. Our approach (red horizontal line) outperforms data bundling method, regardless of the size of training data.

8.7 TASK DISTANCE

Let $\mathcal{P}_{N_\theta}(\mathcal{T}, D^{te}) \in [0, 1]$ 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 $\mathcal{T}$.

Definition 7 (ε-approximation Network)**.** *A model N_θ is called an ε-approximation network for a task-dataset pair $(\mathcal{T}, D)$ if it is trained using the training data D^{tr} such that $\mathcal{P}_{N_\theta}(\mathcal{T}, D^{te}) \geq 1 - \varepsilon$, for a given $0 < \varepsilon < 1$.*

8.7.1 COMPARISON BETWEEN UNSYMMETRIZED AND SYMMETRIZED TASK DISTANCE

We compare the unsymmetrized and symmetrized task distances on the Jobs and the Twins dataset. Figure 7 shows that the proposed symmetrized task distance has successfully captured the symmetries within causal inference tasks. p (on the x-axis) denotes the probability of flipping treatment assignments of the original dataset. The altered datasets with $p = 1$ (i.e., the flipped dataset) and $p = 0$ (i.e., the original dataset) are the closest task to the original task (as expected). The altered dataset with $p = 0.5$ is the furthest dataset (as expected). Thus, the trend of the points is expected to resemble an inverted 'U'. We observe that the symmetrized task distance exhibits this trend. In contrast, the unsymmetrized task distance (in the right figures) fails to demonstrate this trend.

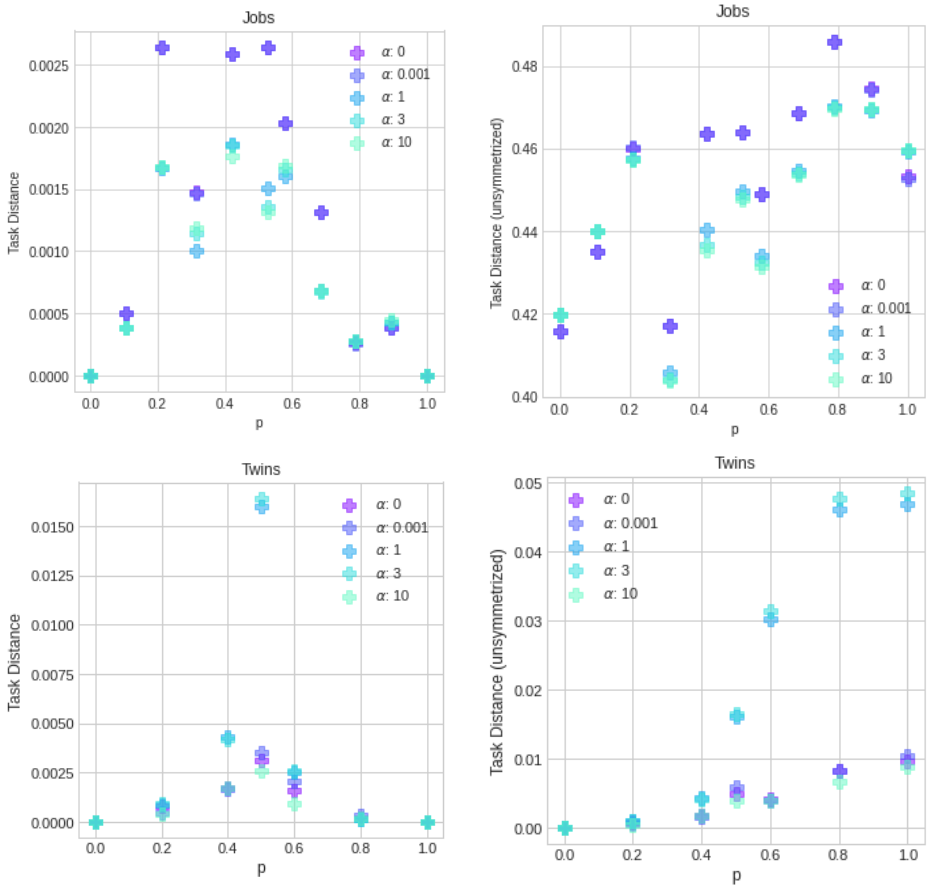

Figure 7: Comparison between unsymmtrized and symmetrized task distances. The left plots are symmetrized task distance; the right plots are unsymmetrized task distance. The inverted 'U' shape of the symmetrized task distance demonstrates that it has successfully identified the symmetry within causal inference tasks. This trend (i.e., the inverted 'U') is not observed on the unsymmetrized task distance.

8.7.2 TASK DISTANCE BETWEEN COUNTERFACTUAL TASKS

In the following section, we denote the pair $a = (\mathcal{T}_a, D_a)$ by $a_F = (\mathcal{T}_{a_F}, D_{a_F})$ (respectively $a_{CF} = (\mathcal{T}_{a_{CF}}, D_{a_{CF}})$) whenever D_a is sampled from the factual (respectively counterfactual) distribution. We refer to $(\mathcal{T}_{a_F}, D_{a_F})$ and $(\mathcal{T}_{a_{CF}}, D_{a_{CF}})$ as the corresponding factual and counterfactual tasks.

The following theorem proves that the order of proximity of tasks is preserved even if we go to a parallel universe where 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 4. *Let $\mathbb{T}$ be the set of tasks and let $a_F = (\mathcal{T}_{a_F}, D_{a_F})$, $b_F = (\mathcal{T}_{b_F}, D_{b_F})$, and $c_F = (\mathcal{T}_{c_F}, D_{c_F})$ be three factual tasks and $a_{CF} = (\mathcal{T}_{a_{CF}}, D_{a_{CF}})$, $b_{CF} = (\mathcal{T}_{b_{CF}}, D_{b_{CF}})$, and $c_{CF} = (\mathcal{T}_{c_{CF}}, D_{c_{CF}})$ their corresponding counterfactual tasks.*

Suppose that there exists a class of neural networks $\mathcal{N} = \{N_\theta\}_{\theta \in \Theta}$ for which:

$$\forall a, b, c \in \mathbb{T},\ s[a, b] \leq s[a, c] + s[c, b] \tag{19}$$

and the TAS between the factual and the counterfactual can be arbitrarily small

$$\forall \epsilon > 0, \exists N_\theta \in \mathcal{N},\ s[a_F, a_{CF}] < \epsilon \tag{20}$$

Then we have the following result,

$$s[a_F, b_F] \leq s[a_F, c_F] \implies s[a_{CF}, b_{CF}] \leq s[a_{CF}, c_{CF}] \tag{21}$$

Proof of Theorem 4
Suppose that $s[a_F, b_F] \leq s[a_F, c_F]$. Then for every $\epsilon > 0$ we have,

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

This is true for every $\epsilon > 0$, therefore $s[a_{CF}, b_{CF}] \leq s[a_{CF}, c_{CF}]$.
This concludes the proof.