CONTINUAL LIFELONG CAUSAL EFFECT INFERENCE WITH REAL WORLD EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

The era of real world evidence has witnessed an increasing availability of observational data, which much facilitates the development of causal effect inference. Although significant advances have been made to overcome the challenges in causal effect estimation, such as missing counterfactual outcomes and selection bias, they only focus on source-specific and stationary observational data. In this paper, we investigate a new research problem of causal effect inference from incrementally available observational data, and present three new evaluation criteria accordingly, including extensibility, adaptability, and accessibility. We propose a Continual Causal Effect Representation Learning method for estimating causal effect with observational data, which are incrementally available from non-stationary data distributions. Instead of having access to all seen observational data, our method only stores a limited subset of feature representations learned from previous data. Combining the selective and balanced representation learning, feature representation distillation, and feature transformation, our method achieves the continual causal effect estimation for new data without compromising the estimation capability for original data. Extensive experiments demonstrate the significance of continual causal effect inference and the effectiveness of our method.

1 INTRODUCTION

Causal effect inference is a critical research topic across many domains, such as statistics, computer science, public policy, and economics. Randomized controlled trials (RCT) are usually considered as the gold-standard for causal effect inference, which randomly assigns participants into a treatment or control group. As the RCT is conducted, the only expected difference between the treatment and control groups is the outcome variable being studied. However, in reality, randomized controlled trials are always time-consuming and expensive, and thus the study cannot involve many subjects, which may be not representative of the real-world population the intervention would eventually target. Nowadays, estimating causal effects from observational data has become an appealing research direction owing to a large amount of available data and low budget requirements, compared with RCT (Yao et al., 2020). Researchers have developed various strategies for causal effect inference with observational data, such as tree-based methods (Chipman et al., 2010; Wager & Athey, 2018), representation learning methods (Johansson et al., 2016; Li & Fu, 2017; Shalit et al., 2017; Chu et al., 2020), adapting Bayesian algorithms (Alaa & van der Schaar, 2017), generative adversarial nets (Yoon et al., 2018), variational autoencoders (Louizos et al., 2017) and so on.

Although significant advances have been made to overcome the challenges in causal effect estimation with observational data, such as missing counterfactual outcomes and selection bias between treatment and control groups, the existing methods only focus on source-specific and stationary observational data. Such learning strategies assume that all observational data are already available during the training phase and from the only one source. This assumption is unsubstantial in practice due to two reasons. The first one is based on the characteristics of observational data, which are incrementally available from non-stationary data distributions. For instance, the number of electronic medical records in one hospital is growing every day, or the electronic medical records for one disease may be from different hospitals or even different countries. This characteristic implies that one cannot have access to all observational data at one time point and from one single source. The second reason is based on the realistic consideration of accessibility. For example, when the new observational are available, if we want to refine the model previously trained by original data, maybe

the original training data are no longer accessible due to a variety of reasons, e.g., legacy data may be unrecorded, proprietary, too large to store, or subject to privacy constraint (Zhang et al., 2020). This practical concern of accessibility is ubiquitous in various academic and industrial applications. That's what it boiled down to: in the era of big data, we face the new challenges in causal inference with observational data: the **extensibility** for incrementally available observational data, the **adaptability** for extra domain adaptation problem except for the imbalance between treatment and control groups in one source, and the **accessibility** for a huge amount of data.

Existing causal effect inference methods, however, are unable to deal with the aforementioned new challenges, i.e., extensibility, adaptability, and accessibility. Although it is possible to adapt existing causal inference methods to address the new challenges, these adapted methods still have inevitable defects. Three straightforward adaptation strategies are described as follows. (1) If we directly apply the model previously trained based on original data to new observational data, the performance on new task will be very poor due to the domain shift issues among different data sources; (2) If we utilize newly available data to re-train the previously learned model, adapting changes in the data distribution, old knowledge will be completely or partially overwritten by the new one, which can result in severe performance degradation on old tasks. This is the well-known *catastrophic forgetting* problem (McCloskey & Cohen, 1989; French, 1999); (3) To overcome the catastrophic forgetting problem, we may rely on the storage of old data and combine the old and new data together, and then re-train the model from scratch. However, this strategy is memory inefficient and time-consuming, and it brings practical concerns such as copyright or privacy issues when storing data for a long time (Samet et al., 2013). Our empirical evaluations in Section 4 demonstrate that any of these three strategies in combination with the existing causal effect inference methods is deficient.

To address the above issues, we propose a **C**ontinual Causal **E**ffect **R**epresentation **L**earning method (CERL) for estimating causal effect with incrementally available observational data. Instead of having access to all previous observational data, we only store a limited subset of feature representations learned from previous data. Combining the selective and balanced representation learning, feature representation distillation, and feature transformation, our method preserves the knowledge learned from previous data and update the knowledge by leveraging new data, so that it can achieve the continual causal effect estimation for new data without compromising the estimation capability for previous data. To summarize, our main contributions include:

- Our work is the first to introduce the continual lifelong causal effect inference problem for the incrementally available observational data and three corresponding evaluation criteria, i.e., extensibility, adaptability, and accessibility.
- We propose a new framework for continual lifelong causal effect inference based on deep representation learning and continual learning.
- Extensive experiments demonstrate the deficiency of existing methods when facing the incrementally available observational data and our model's outstanding performance.

2 Background and Problem Statement

Suppose that the observational data contain n units from d different domains $\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_d\}$ with the same treatment options $\mathcal{T}$. Let $\mathcal{X}$ denote all observed variables, $\mathcal{Y}$ denote the outcomes in the observational data, and $\mathcal{T}$ denote the treatment options. Let $\mathcal{D}_{1:d} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_d\}$ be the set of combination of d domains. We define the d-th domain $\mathcal{D}_d$ as $\mathcal{D}_d = \{(x, y, t)|x \in \mathcal{X}_d, y \in \mathcal{Y}_d, t \in \mathcal{T}_d\}$, which contains n_d units. Each unit in the observational data received one of two or more treatments. Let t_i denote the treatment assignment for unit i; $i = 1, ..., n$. For binary treatments, $t_i = 1$ is for the treatment group, and $t_i = 0$ for the control group. The outcome for unit i is denoted by y_t^i when treatment t is applied to unit i; that is, y_1^i is the potential outcome of unit i in the treatment group and y_0^i is the potential outcome of unit i in the control group. For observational data, only one of the potential outcomes is observed. The observed outcome is called the factual outcome and the remaining unobserved potential outcomes are called counterfactual outcomes.

In this paper, we follow the potential outcome framework for estimating treatment effects (Rubin, 1974; Splawa-Neyman et al., 1990). The individual treatment effect (ITE) for unit i is the difference between the potential treated and control outcomes, and is defined as $\text{ITE}_i = y_1^i - y_0^i$. The average treatment effect (ATE) is the difference between the mean potential treated and control outcomes, which is defined as $\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}(y_1^i - y_0^i)$.

The success of the potential outcome framework is based on the following assumptions (Imbens & Rubin, 2015), which ensure that the treatment effect can be identified. **Stable Unit Treatment Value Assumption (SUTVA)**: The potential outcomes for any units do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. **Consistency**: The potential outcome of treatment t is equal to the observed outcome if the actual treatment received is t. **Positivity**: For any value of x, treatment assignment is not deterministic, i.e.,$P(\mathcal{T} = t|\mathcal{X} = x) > 0$, for all t and x. **Ignorability**: Given covariates, treatment assignment is independent to the potential outcomes, i.e., $(y_1, y_0) \perp\!\!\!\perp t|x$.

The goal of our work is to develop a novel continual causal inference framework, given new available observational data $\mathcal{D}_d$, to estimate the causal effect for newly available data $\mathcal{D}_d$ as well as the previous data $\mathcal{D}_{1:(d-1)}$ without having access to previous training data in $\mathcal{D}_{1:(d-1)}$.

3 The Proposed Framework

The availability of "real world evidence" is expected to facilitate the development of causal effect inference models for various academic and industrial applications. How to achieve continual learning from incrementally available observational data from non-stationary data domains is a new direction in causal effect inference. Rather than only focusing on handling the selection bias problem, we also need to take into comprehensive consideration three aspects of the model, i.e., the **extensibility** for incrementally available observational data, the **adaptability** for various data sources, and the **accessibility** for a huge amount of data.

We propose the **C**ontinual Causal **E**ffect **R**epresentation **L**earning method (CERL) for estimating causal effect with incrementally available observational data. Based on selective and balanced representation learning for treatment effect estimation, CERL incorporates feature representation distillation to preserve the knowledge learned from previous observational data. Besides, aiming at adapting the updated model to original and new data without having access to the original data, and solving the selection bias between treatment and control groups, we propose one representation transformation function, which maps partial original feature representations into new feature representation space and makes the global feature representation space balanced with respect to treatment and control groups. Therefore, CERL can achieve the continual causal effect estimation for new data and meanwhile preserve the estimation capability for previous data, without the aid of original data.

3.1 Model Architecture

To estimate the incrementally available observational data, the framework of CERL is mainly composed of two components: (1) the baseline causal effect learning model is only for the first available observational data, and thus we don't need to consider the domain shift issue among different data sources. This component is equivalent to the traditional causal effect estimation problem; (2) the continual causal effect learning model is for the sequentially available observational data, where we need to handle more complex issues, such as knowledge transfer, catastrophic forgetting, global representation balance, and memory constraint. We present the details of each component as follows.

3.1.1 The Baseline causal effect Learning Model

We first describe the baseline causal effect learning model for the initial observational dataset and then bring in subsequent datasets. For causal effect estimation in the initial dataset, it can be transformed into the traditional causal effect estimation problem. Motivated by the empirical success of deep representation learning for counterfactual inference (Shalit et al., 2017; Chu et al., 2020), we propose to learn the selective and balanced feature representations for treated and control units, and then infer the potential outcomes based on learned representation space.

Learning Selective and Balanced Representation. Firstly, we adopt a deep feature selection model that enables variable selection in one deep neural network, i.e., $g_{w_1} : \mathcal{X} \to \mathcal{R}$, where $\mathcal{X}$ denotes the original covariate space, $\mathcal{R}$ denotes the representation space, and w_1 are the learnable parameters in function g. The elastic net regularization term (Zou & Hastie, 2005) is adopted in our model, i.e., $L_{w_1} = \|w_1\|_2^2 + \|w_1\|_1$. Elastic net throughout the fully connected representation layers assigns larger weights to important features. This strategy can effectively filter out the irrelevant variables and highlight the important variables.

Due to the selection bias between treatment and control groups and among the sequential different data sources, the magnitudes of confounders may be significantly different. To effectively eliminate the imbalance caused by the significant difference in magnitudes between treatment and control groups and among different data sources, we propose to use cosine normalization in the last representation layer. In the multi-layer neural networks, we traditionally use dot products between the output vector of the previous layer and the incoming weight vector, and then input the products to the activation function. The result of dot product is unbounded. Cosine normalization uses cosine similarity instead of simple dot products in neural networks, which can bound the pre-activation between -1 and 1. The result could be even smaller when the dimension is high. As a result, the variance can be controlled within a very narrow range (Luo et al., 2018). Cosine normalization is defined as $r = \sigma(r_{norm}) = \sigma\big(\cos(w, x)\big) = \sigma(\frac{w \cdot x}{|w||x|})$, where r_{norm} is the normalized pre-activation, w is the incoming weight vector, x is the input vector, and σ is nonlinear activation function.

Motivated by Shalit et al. (2017), we adopt integral probability metrics (IPM) when learning the representation space to balance the treatment and control groups. The IPM measures the divergence between the representation distributions of treatment and control groups, so we want to minimize the IPM to make two distributions more similar. Let $P(g(x)|t = 1)$ and $Q(g(x)|t = 0)$ denote the empirical distributions of the representation vectors for the treatment and control groups, respectively. We adopt the IPM defined in the family of 1-Lipschitz functions, which leads to IPM being the Wasserstein distance (Sriperumbudur et al., 2012; Shalit et al., 2017). In particular, the IPM term with Wasserstein distance is defined as $\text{Wass}(P, Q) = \inf_{k \in \mathcal{K}} \int_{g(x)} \|k(g(x)) - g(x)\| P(g(x)) d(g(x))$, where γ denotes the hyper-parameter controlling the trade-off between $\text{Wass}(P, Q)$ and other terms in the final objective function. $\mathcal{K} = \{k | Q(k(g(x))) = P(g(x))\}$ defines the set of push-forward functions that transform the representation distribution of the treatment distribution P to that of the control Q and $g(x) \in \{g(x)_i\}_{i:t_i=1}$.

Inferring Potential Outcomes. We aim to learn a function $h_{\theta_1} : \mathcal{R} \times \mathcal{T} \rightarrow \mathcal{Y}$ that maps the representation vectors and treatment assignment to the corresponding observed outcomes, and it can be parameterized by deep neural networks. To overcome the risk of losing the influence of $\mathcal{T}$ on $\mathcal{R}$, $h_{\theta_1}(g_{w_1}(x), t)$ is partitioned into two separate tasks for treatment and control groups, respectively. Each unit is only updated in the task corresponding to its observed treatment. Let $\hat{y}_i = h_{\theta_1}(g_{w_1}(x), t)$ denote the inferred observed outcome of unit i corresponding to factual treatment t_i. We minimize the mean squared error in predicting factual outcomes: $L_Y = \frac{1}{n_1} \sum_{i=1}^{n_1} (\hat{y}_i - y_i)^2$.

Putting all the above together, the objective function of our baseline causal effect learning model is: $L = L_Y + \alpha Wass(P, Q) + \lambda L_{w_1}$, where α and λ denote the hyper-parameters controlling the trade-off among $Wass(P, Q)$, L_w, and L_Y in the objective function.

3.1.2 The Sustainability of Model Learning

By far, we have built the baseline model for causal effect estimation with observational data from a single source. To avoid catastrophic forgetting when learning new data, we propose to preserve a subset of lower-dimensional feature representations rather than all original covariates. We also can adjust the number of preserved feature representations according to the memory constraint.

After the completion of baseline model training, we store a subset of feature representations $\mathcal{R}_1 = \{g_{w_1}(x) | x \in \mathcal{D}_1\}$ and the corresponding $\{\mathcal{Y}, \mathcal{T}\} \in \mathcal{D}_1$ as memory $\mathcal{M}_1$. The size of stored representation vectors can be reduced to satisfy the pre-specified memory constraint by a herding algorithm (Welling, 2009; Rebuffi et al., 2017). The herding algorithm can create a representative set of samples from distribution and requires fewer samples to achieve a high approximation quality than random subsampling. We run the herding algorithm separately for treatment and control groups to store the same number of feature representations from treatment and control groups. At this point, we only store the memory set $\mathcal{M}_1$ and model g_{w_1}, without the original data ($\mathcal{D}_1$).

3.1.3 The Continual causal effect Learning Model

For now, we have stored memory $\mathcal{M}_1$ and baseline model. To continually estimate the causal effect for incrementally available observational data, we incorporate feature representation distillation and feature representation transformation to estimate causal effect for all seen data based on balanced global feature representation space. The framework of CERL is shown in Fig. 1.

Feature Representation Distillation. For next available data $\mathcal{D}_2 = \{(x, y, t)|x \in \mathcal{X}_2, y \in \mathcal{Y}_2, t \in \mathcal{T}_2\}$, we adopt the same selective representation learning $g_{w_2} : \mathcal{X}_2 \to \mathcal{R}_2$ with elastic net regularization (L_{w_2}) on new parameters w_2. Because we expect our model can estimate causal effect for both previous and new data, we want the new model to inherit some knowledge from previous model. In continual learning, knowledge distillation (Hinton et al., 2015; Li & Hoiem, 2017) is commonly adopted to alleviate the catastrophic forgetting, where knowledge is transferred from one network to another network by encouraging the outputs of the original and new network to be similar. However, for the continual causal effect estimation problem, we focus more on the feature representations, which are required to be balanced between treatment and control, and among different data domains. Inspired by Hou et al. (2019); Dhar et al. (2019); Iscen et al. (2020), we propose feature representation distillation to encourage the representation vector $\{g_{w_1}(x)|x \in \mathcal{D}_2\}$ based on baseline model to be similar to the representation vector $\{g_{w_2}(x)|x \in \mathcal{D}_2\}$ based on new model by Euclidean distance. This feature distillation can help prevent the learned representations from drifting too much in the new feature representation space. Because we apply the cosine normalization to feature representations and $\|A - B\|^2 = (A - B)^\intercal(A - B) = \|A\|^2 + \|B\|^2 - 2A^\intercal B = 2\big(1 - cos(A, B)\big)$, the feature representation distillation is defined as $L_{FD}(x) = 1 - cos\big(g_{w_1}(x), g_{w_2}(x)\big)$, where $x \in \mathcal{D}_2$.

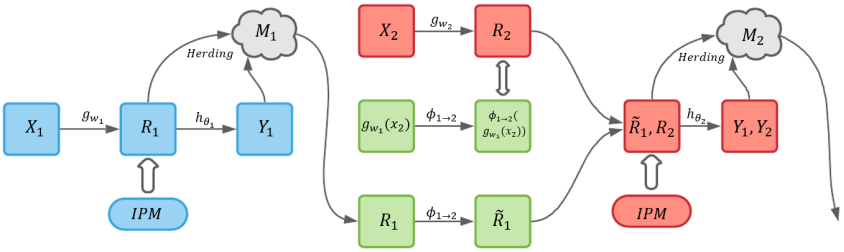

Figure 1: The blue part is baseline causal effect learning model for the first observational data. After baseline model training, store a subset of feature representations $\mathcal{R}_1$ into $\mathcal{M}_1$ by herding algorithm. The green part helps to map $\mathcal{R}_1$ to transformed feature representations $\tilde{\mathcal{R}}_1$ compatible with new feature representations space $\mathcal{R}_2$. Then the red part is used for continual causal effect estimation based on feature distillation and balanced global feature representation learning for $\tilde{\mathcal{R}}_1$ and $\mathcal{R}_2$.

Feature Representation Transformation. We have previous feature representations $\mathcal{R}_1$ stored in $\mathcal{M}_1$ and new feature representations $\mathcal{R}_2$ extracted from newly available data. $\mathcal{R}_1$ and $\mathcal{R}_2$ lie in different feature representation space and they are not compatible with each other because they are learned from different models. In addition, we cannot learn the feature representations of previous data from the new model g_{w_2}, as we no longer have access to previous data. Therefore, to balance the global feature representation space including previous and new representations between treatment and control groups, a feature transformation function is needed from previous feature representations $\mathcal{R}_1$ to transformed feature representations $\tilde{\mathcal{R}}_1$ compatible with new feature representations space $\mathcal{R}_2$. We define a feature transformation function as $\phi_{1\to2} : \mathcal{R}_1 \to \tilde{\mathcal{R}}_1$. We also input the feature representations of new data $\mathcal{X}_2$ learned from old model, i.e., $g_{w_1}(x)$, to get the transformed feature representations of new data, i.e., $\phi_{1\to2}(g_{w_1}(x))$. To keep the transformed space compatible with the new feature representation space, we train the transformation function $\phi_{1\to2}$ by making the $\phi_{1\to2}(g_{w_1}(x))$ and $g_{w_2}(x)$ similar, where $x \in \mathcal{X}_2$. The loss function is defined as $L_{FT}(x) = 1 - cos\big(\phi_{1\to2}(g_{w_1}(x)), g_{w_2}(x)\big)$, which is used to train the function $\phi_{1\to2}$ to transform feature representations between different feature spaces. Then, we can attain the transformed old feature representations $\tilde{\mathcal{R}}_1 = \phi_{1\to2}(\mathcal{R}_1)$, which is in the same space as $\mathcal{R}_2$.

Balancing Global Feature Representation Space. We have obtained a global feature representation space including the transformed representations of stored old data and new representations of new available data. We adopt the same integral probability metrics as baseline model to make sure that the representation distributions are balanced for treatment and control groups in the global feature representation space. In addition, we define a potential outcome function $h_{\theta_2} : (\tilde{\mathcal{R}}_1, \mathcal{R}_2) \times \mathcal{T} \to \mathcal{Y}$. Let $\hat{y}_i^M = h_{\theta_2}\big(\phi_{1\to2}(r_i), t\big)$, where $r_i \in \mathcal{M}_1$, and $\hat{y}_j^D = h_{\theta_2}\big(g_{w_2}(x_j), t\big)$, where $x_j \in \mathcal{D}_2$ denote the inferred observed outcomes. We aim to minimize the mean squared error in predicting factual outcomes for global feature representations including transformed old feature representations and

new feature representations: $L_G = \frac{1}{\tilde{n}_1}\sum_{i=1}^{\tilde{n}_1}(\hat{y}_i^M - y_i^M)^2 + \frac{1}{n_2}\sum_{j=1}^{n_2}(\hat{y}_j^D - y_j^D)^2$, where $\tilde{n}_1$ is the number of units stored in $\mathcal{M}_1$ by herding algorithm, $y_i^M \in \mathcal{M}_1$, and $y_j^D \in \mathcal{D}_2$.

In summary, the objective function of our continual causal effect learning model is $L = L_G + \alpha Wass(P, Q) + \lambda L_{w_2} + \beta L_{FD} + \delta L_{FT}$, where α, λ, β, and δ denote the hyper-parameters controlling the trade-off among $Wass(P, Q)$, L_{w_2}, L_{FD}, L_{FT}, and L_G in the final objective function.

3.2 Overview of CERL

In the above sections, we have provided the baseline and continual causal effect learning models. When the continual causal effect learning model for the second data is trained, we can extract the $\mathcal{R}_2 = \{g_{w_2}(x)|x \in \mathcal{D}_2\}$ and $\tilde{\mathcal{R}}_1 = \{\phi_{1\to2}(r)|r \in \mathcal{M}_1\}$. We define a new memory set as $\mathcal{M}_2 = \{\mathcal{R}_2, \mathcal{Y}_2, \mathcal{T}_2\} \cup \phi_{1\to2}(\mathcal{M}_1)$, where $\phi_{1\to2}(\mathcal{M}_1)$ includes $\tilde{\mathcal{R}}_1$ and the corresponding $\{\mathcal{Y}, \mathcal{T}\}$ stored in $\mathcal{M}_1$. Similarly, to satisfy the pre-specified memory constraint, $\mathcal{M}_2$ can be reduced by conducting the herding algorithm to store the same number of feature representations from treatment and control groups. We only store the new memory set $\mathcal{M}_2$ and new model g_{w_2}, which are used to train the following model and balance the global feature representation space. It is unnecessary to store the original data ($\mathcal{D}_1$ and $\mathcal{D}_2$) any longer.

We follow the same procedure for the subsequently available observational data. When we obtain the new observational data $\mathcal{D}_d$, we can train $h_{\theta_d}(g_{w_d})$ and $\phi_{d-1\to d} : \mathcal{R}_{d-1} \to \tilde{\mathcal{R}}_{d-1}$ based on the continual causal effect learning model. Besides, the new memory set is defined as: $\mathcal{M}_d = \{\mathcal{R}_d, \mathcal{Y}_d, \mathcal{T}_d\} \cup \phi_{d-1\to d}(\mathcal{M}_{d-1})$. So far, our model $h_{\theta_d}(g_{w_d})$ can estimate causal effect for all seen observational data regardless of the data source and it doesn't require access to previous data. The detailed procedures of our CERL method are summarized in Algorithm 1.

Algorithm 1 Continual Causal Effect Representation Learning

Data: Given d incrementally available observational data from $\mathcal{D}_1$ to $\mathcal{D}_D$
if $\{\mathcal{X}, \mathcal{Y}, \mathcal{T}\} \in \mathcal{D}_1$ **then**
  *** Train baseline causal effect model $h_{\theta_1}(g_{w_1})$ ***
  $w_1, \theta_1 =$ OPTIMIZE$(L_Y + \alpha Wass(P, Q) + \lambda L_{w_1})$
  $\mathcal{R}_1 = \{g_{w_1}(x)|x \in \mathcal{D}_1\}$
  $\mathcal{M}_1 =$ HERDING$\{\mathcal{R}_1, \mathcal{Y}_1, \mathcal{T}_1\}$
else
  for $\{\mathcal{X}, \mathcal{Y}, \mathcal{T}\} \in \mathcal{D}_2, ..., \mathcal{D}_D$ **do**
    *** Train continual causal effect model $h_{\theta_d}(g_{w_d})$ ***
    $w_d, \theta_d, \phi_{d-1\to d} =$ OPTIMIZE$(L_G + \alpha Wass(P, Q) + \lambda L_{w_2} + \beta L_{FD} + \delta L_{FT})$
    $\tilde{\mathcal{R}}_{d-1} = \phi_{d-1\to d}(\mathcal{R}_{d-1})$
    $\mathcal{R}_d = \{g_{w_d}(x)|x \in \mathcal{D}_d\}$
    $\mathcal{M}_d =$ HERDING$\big(\{\mathcal{R}_d, \mathcal{Y}_d, \mathcal{T}_d\} \cup \{\tilde{\mathcal{R}}_{d-1}, \mathcal{Y}_{d-1} \in \mathcal{M}_{d-1}, \mathcal{T}_{d-1} \in \mathcal{M}_{d-1}\}\big)$
  end
end

4 Experiments

The traditional benchmarks for causal effect estimation are not suitable for evaluating the performance of continual causal effect inference methods due to three reasons: (1) all the data are available at a time and they cannot continue to grow; (2) all the data are from one single stationary source; (3) the size of data is usually small so that we don't face memory constraint problem. However, in practice, a mass of observational data is incrementally growing from different non-stationary data distributions. For example, the electronic health records are "big health data", which are composed of different healthcare databases always collected from different hospitals or other organizations even across multiple countries. Besides, they are incessantly growing to update the databases in real time. Therefore, we generate a series of synthetic datasets that mimic the electronic health records collected from different sources. In this section, we will adapt the representative causal effect estimation method to continual causal effect estimation under three strategies and compare them with our model from three criteria.

4.1 SYNTHETIC DATASET

Our synthetic data include confounders, instrumental, adjustment, and irrelevant variables. The interrelations among these variables, treatments, and outcomes are illustrated in Figure 2. The number of observed variables in the vector $X = (C^\intercal, Z^\intercal, I^\intercal, A^\intercal)^\intercal$ is set to 100, including 35 confounders in C, 35 adjustment variables in A, 10 instrumental variables in Z, and 20 irrelevant variables in I. The model used to generate the continuous outcome variable Y in this simulation is the partially linear regression model, extending the ideas described in Robinson (1988); Jacob et al. (2019); Chu et al. (2020), i.e., $Y = \tau((C^\intercal, A^\intercal)^\intercal)T + g((C^\intercal, A^\intercal)^\intercal) + \epsilon$, where ϵ are unobserved covariates, which follow a standard normal distribution $N(0, 1)$ and $E[\epsilon|C, A, T] = 0$. $T \overset{ind.}{\sim} \text{Bernoulli}(e_0((C^\intercal, Z^\intercal)^\intercal))$ and $e_0((C^\intercal, Z^\intercal)^\intercal)$ is the propensity score, which represents the treatment selection bias based on their own confounders C and instrumental variables Z. Because we aim to simulate multiple data sources $\{\mathcal{D}_d; d = 1, ..., D\}$, the vector of all observed covariates $X = (C^\intercal, Z^\intercal, I^\intercal, A^\intercal)^\intercal$ is sampled from different multivariate normal distribution with mean vector $\mu_C^d, \mu_Z^d, \mu_I^d$, and μ_A^d and different random positive definite covariance matrices Σ^d.

We totally simulate five different data sources with five different multivariate normal distributions to represent the incrementally available observational data. In each data source, we randomly draw 10000 samples including treatment units and control units. Therefore, for five datasets, they have different selection bias, magnitude of covariates, covariance matrices for variables, and number of treatment and control units. To ensure a robust estimation of model performance, for each data source, we repeat the simulation procedure 10 times and obtain 10 synthetic datasets. The details of data simulation are provided in Section A of Appendix.

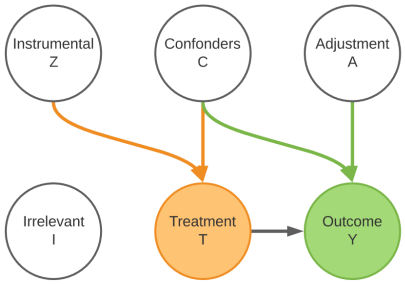

Figure 2: The types of variables.

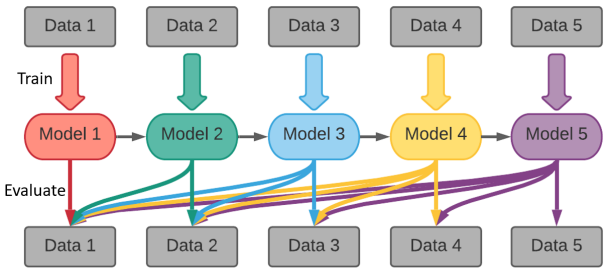

Figure 3: The work flow of task.

4.2 RESULTS AND ANALYSIS

Task Description. The five observational data are incrementally available in sequence, and the model will continue to estimate the causal effect without having access to previous data. To evaluate the capability of continual learning without forgetting, after training one new data, we ask model to estimate the causal effect for the test dataset composed of samples from previous and current data sources, as illustrated in Figure 3. We adopt two commonly used evaluation metrics. The first one is the error of ATE estimation, which is defined as $\epsilon_{\text{ATE}} = |\text{ATE} - \widehat{\text{ATE}}|$, where ATE is the true value and $\widehat{\text{ATE}}$ is an estimated ATE. The second one is the error of expected precision in estimation of heterogeneous effect (PEHE) Hill (2011), which is defined as $\epsilon_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}(\text{ITE}_i - \widehat{\text{ITE}}_i)^2$, where ITE_i is the true ITE for unit i and $\widehat{\text{ITE}}_i$ is an estimated ITE for unit i.

Extended Classical Causal Effect Estimation Model. We employ three strategies to adapt traditional causal effect estimation models to incrementally available observational data: (A) directly apply the model previously trained based on original data to new observational data; (B) utilize newly available data to re-train the previously learned model; (C) store all previous data and combine with new data to re-train the model from scratch. For these three strategies, we conduct the experiments based on the counterfactual regression model (CFR) (Shalit et al., 2017), which is a representative causal effect estimation method. Because CFR is not designed for continual causal effect estimation, we only use one previous data and another newly available data. As shown in Table 1, strategy CFR-A performs well on previous data, but declines on new dataset; strategy CFR-B shows the

Table 1: Performance on two sequential data. We present the mean value of $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} on test sets from two datasets. The standard deviations are tiny, so they are not reported.

	Previous data			New data		
Strategy	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	Result	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	Result
CFR-A	1.47	0.35	✓	2.51	0.73	×
CFR-B	1.82	0.47	×	1.63	0.45	✓
CFR-C	1.49	0.36	✓	1.62	0.44	✓
CERL	1.49	0.37	✓	1.63	0.44	✓

catastrophic forgetting problem where the performance on previous dataset is poor; strategy CFR-C performs well on both previous and new data, but it re-trains the whole model using previous and new data. However, if there is a memory constraint or a barrier to accessing previous data, the strategy CFR-C cannot be conducted. Our CERL has a similar performance to strategy CFR-C, while CERL does not require access to previous data. It can demonstrate our model's superiority from extensibility, adaptability, and accessibility.

CERL Performance Evaluation. We further evaluate the performance of CERL from three perspectives, i.e., the impact of memory constraint, effeteness of cosine normalization, and its robustness to hyper-parameters. As shown in Figure 4 (a) and (b), as the model continually learns a new dataset, every time when finishing training one new dataset, we report the $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} on test sets composed of previous data and new data. Our model with memory constraints has a similar performance to the ideal situation, where all data are available to train the model from scratch. However, our model can effectively save memory space, e.g., when facing the fifth dataset, our model only stores 1000, 5000, or 10000 feature representations, but the ideal situation needs to store $5 \times 10000 = 50000$ observations with all covariates. For the cosine normalization, we perform an ablation study of CERL, where we remove cosine normalization in the representation learning procedure. We find the $\sqrt{\epsilon_{\text{PEHE}}}$ increases from 1.80 and 1.92 and ϵ_{ATE} from 0.55 to 0.61. Next, we explore the model's sensitivity to the most important parameter α and δ, which controls the representation balance and representation transformation. From Fig. 4 (c) and (d), we observe that the performance is stable over a large parameter range. In addition, the parameter β for feature representation distillation is set to 1 (Rebuffi et al., 2017; Iscen et al., 2020).

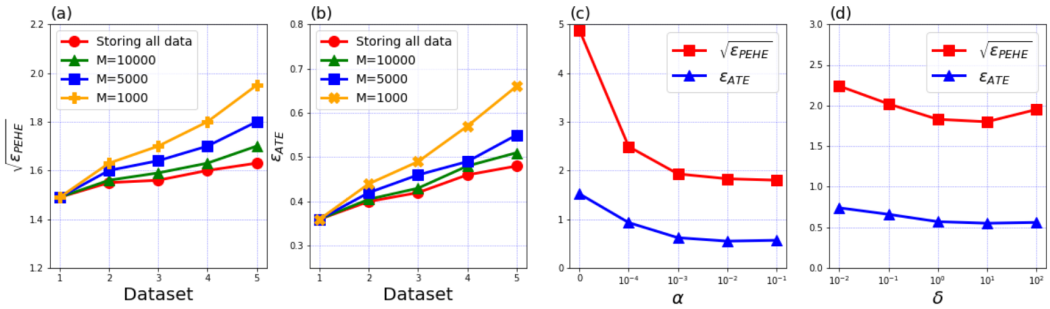

Figure 4: Performance of CERL under different settings.

5 Conclusion

It is the first time to propose the continual lifelong causal effect inference problem and the corresponding evaluation criteria. As the real world evidence is becoming more prominent, how to integrate and utilize these powerful data for causal effect estimation becomes a new research challenge. To address this challenge, we propose the Continual Causal Effect Representation Learning method for estimating causal effect with observational data, which are incrementally available from non-stationary data distributions. Extensive experiments demonstrate the superiority of our method over baselines for continual causal effect estimation.

REFERENCES

Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 3424–3432, 2017.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Zhixuan Chu, Stephen L Rathbun, and Sheng Li. Matching in selective and balanced representation space for treatment effects estimation. *arXiv preprint arXiv:2009.06828*, 2020.

Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5138–5146, 2019.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

Johanna Hardin, Stephan Ramon Garcia, and David Golan. A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, pp. 1733–1762, 2013.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 831–839, 2019.

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

Ahmet Iscen, Jeffrey Zhang, Svetlana Lazebnik, and Cordelia Schmid. Memory-efficient incremental learning through feature adaptation. *arXiv preprint arXiv:2004.00713*, 2020.

Daniel Jacob, Wolfgang Karl Härdle, and Stefan Lessmann. Group average treatment effects for observational studies. *arXiv preprint arXiv:1911.02688*, 2019.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029, 2016.

Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2008.

Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, pp. 929–939, 2017.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Chunjie Luo, Jianfeng Zhan, Xiaohe Xue, Lei Wang, Rui Ren, and Qiang Yang. Cosine normalization: Using cosine similarity instead of dot product in neural networks. In *International Conference on Artificial Neural Networks*, pp. 382–391. Springer, 2018.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pp. 931–954, 1988.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Saeed Samet, Ali Miri, and Eric Granger. Incremental learning of privacy-preserving bayesian networks. *Applied Soft Computing*, 13(8):3657–3667, 2013.

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

Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pp. 465–472, 1990.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL `https://doi.org/10.1080/01621459.2017.1319839`.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1121–1128, 2009.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *arXiv preprint arXiv:2002.02770*, 2020.

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.

Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 1131–1140, 2020.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

A SIMULATION PROCEDURE

Our synthetic data include confounders, instrumental, adjustment, and irrelevant variables. The interrelations among these variables, treatments, and outcomes are illustrated in Figure 2. The number of observed variables in the vector $X = (C^\intercal, Z^\intercal, I^\intercal, A^\intercal)^\intercal$ is set to 100, including 35 confounders in C, 35 adjustment variables in A, 10 instrumental variables in Z, and 20 irrelevant variables in I. The model used to generate the continuous outcome variable Y in this simulation is the partially linear regression model, extending the ideas described in Robinson (1988); Jacob et al. (2019); Chu et al. (2020):

$$Y = \tau((C^\intercal, A^\intercal)^\intercal)T + g((C^\intercal, A^\intercal)^\intercal) + \epsilon, \tag{1}$$

where ϵ are unobserved covariates, which follow a standard normal distribution $N(0, 1)$ and $E[\epsilon|C, A, T] = 0$. $T \overset{ind.}{\sim}$ Bernoulli$(e_0((C^\intercal, Z^\intercal)^\intercal))$ and $e_0((C^\intercal, Z^\intercal)^\intercal)$ is the propensity score, which represents the treatment selection bias based on their own confounders C and instrumental variables Z. Because we aim to simulate multiple data sources $\{\mathcal{D}_d; d = 1, ..., D\}$, the vector of all observed covariates $X = (C^\intercal, Z^\intercal, I^\intercal, A^\intercal)^\intercal$ is sampled from different multivariate normal distribution with mean vector $\mu_C^d, \mu_Z^d, \mu_I^d,$ and μ_A^d and different random positive definite covariance matrices Σ^d.

For each data source, except for the different magnitude of mean vector and structure of covariance matrix, the simulation procedure is the same. Let D be the diagonal matrix with the square roots of the diagonal entries of Σ on its diagonal, i.e., $D = \sqrt{diag(\sigma)}$, then the correlation matrix is given as:

$$R = D^{-1}\Sigma D^{-1}. \tag{2}$$

We use algorithm 3 in Hardin et al. (2013) to simulate positive definite correlation matrices consisting of different types of variables. Our correlation matrices are based on the hub correlation structure which has a known correlation between a hub variable and each of the remaining variables (Zhang & Horvath, 2005; Langfelder et al., 2008). Each variable in one type of variables is correlated to the hub-variable with decreasing strength from specified maximum correlation to minimum correlation, and different types of variables are generated independently or with weaker correlation among variable types. Defining the first variable as the hub, for the ith variable $(i = 2, 3, ..., n)$, the correlation between it and the hub-variable in one type of variables is given as:

$$R_{i,1} = \rho_{\text{max}} - \left(\frac{i-2}{d-2}\right)^{\gamma} (\rho_{\text{max}} - \rho_{\text{min}}), \tag{3}$$

where ρ_{max} and ρ_{min} are specified maximum and minimum correlations, and the rate γ controls rate at which correlations decay.

After specifying the relationship between the hub variable and the remaining variables in the same type of variables, we use Toeplitz structure to fill out the remainder of the hub correlation matrix and get the hub-Toeplitz correlation matrix R_{type} for other type of variables. Here, R is the $n \times n$ matrix having the blocks $R_Z, R_C, R_A,$ and R_I along the diagonal and zeros at off-diagonal elements. This yields a correlation matrix with nonzero correlations within the same type and zero correlation among other types. The amount of correlations among types which can be added to the positive-definite correlation matrix R is determined by its smallest eigenvalue.

The function $\tau((C^\intercal, A^\intercal)^\intercal)$ describes the true treatment effect as a function of the values of adjustment variables A and confounders C; namely $\tau((C^\intercal, A^\intercal)^\intercal) = (\sin((C^\intercal, A^\intercal)^\intercal \times b_\tau))^2$ where b_τ represents weights for every covariate in the function, which is generated by uniform$(0, 1)$. The variable treatment effect implies that its strength differs among the units and is therefore conditioned on C and A. The function $g((C^\intercal, A^\intercal)^\intercal)$ can have an influence on outcome regardless of treatment assignment. It is calculated via a trigonometric function to make the covariates nonlinear, which is defined as $g((C^\intercal, A^\intercal)^\intercal) = (\cos((C^\intercal, A^\intercal)^\intercal \times b_g))^2$. Here, b_g represents a weight for each covariate in this function, which is generated by uniform$(0, 1)$. The bias is attributed to unobserved covariates which follow a random normal distribution $N(0, 1)$. The treatment assignment T follows the Bernoulli distribution, i.e., $T \overset{ind.}{\sim}$ Bernoulli$(e_0((C^\intercal, Z^\intercal)^\intercal))$ with probability

$e_0((C^\intercal, Z^\intercal)^\intercal) = \Phi(\frac{a-\mu(a)}{\sigma(a)})$, where $e_0((C^\intercal, Z^\intercal)^\intercal)$ represents the propensity score, which is the cumulative distribution function for a standard normal random variable based on confounders C and instrumental variables Z, i.e., $a = \sin((C^\intercal, Z^\intercal)^\intercal \times b_a)$, where b_a is generated by uniform$(0, 1)$.

We totally simulate five different data sources with five different multivariate normal distributions to represent the incrementally available observational data. In each data source, we randomly draw 10000 samples including treatment units and control units. Therefore, for five datasets, they have different selection bias, magnitude of covariates, covariance matrices for variables, and number of treatment and control units. To ensure a robust estimation of model performance, for each data source, we repeat the simulation procedure 10 times and obtain 10 synthetic datasets.