Informative Robust Causal Representation for Generalizable Deep Learning

Anonymous authors
Paper under double-blind review

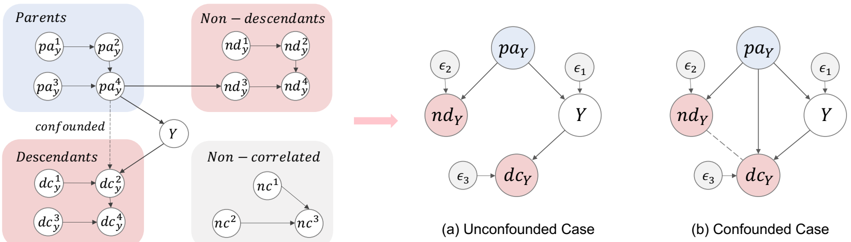
Abstract

In many real-world scenarios, such as image classification and recommender systems, it is evidence that representation learning can improve model's performance over multiple downstream tasks. Existing learning approaches rely on establishing the correlation (or its proxy) between features and the downstream task (labels), which typically results in a representation containing cause, effect and spurious correlated variables of the label. Its generalizability may deteriorate because of the unstability of the non-causal parts. In this paper, we propose to learn causal representation from observational data by regularizing the learning procedure with mutual information measures according to our hypothetical causal graph. The optimization involves a counterfactual loss, based on which we deduce a theoretical guarantee that the causality-inspired learning is with reduced sample complexity and better generalization ability. Extensive experiments show that the models trained on causal representations learned by our approach is robust under adversarial attacks and distribution shift.

1 Introduction

Learning representations from purely observations concerns the problem of finding a low-dimensional, compact representation which is beneficial to prediction models for multiple downstream tasks. It is widely applied in many real-world applications like recommendation system, searching system etc.(Sun et al., 2018; Okura et al., 2017; Zhang et al., 2017; Shi et al., 2018). Generally, the feature representation learning is to map the collected features into a representation space, which is then leveraged for training a prediction model. Most of the representation learning methods are designed based on the information of correlation between the feature and downstream labels only. Some recent works reckon that the representation produced by this schema is not able to achieve robustness and generalization (Schölkopf et al., 2021; Peters et al., 2017; Zhou et al., 2021), since spurious correlated features miss essential information which actually influence the system (Pearl, 2009). For example, obtaining thermometer enables the prediction of the air temperature, but manually putting the thermometer into hot water does not change the air temperature. It is obvious that thermometer scale is not always a stable feature to predict temperature, because non-causal relations exist. Therefore, the causal information rather than correlation is more robust and general for prediction models.

Causal representation learning is an effective approach for extracting invariant, cross-domain stable causal information, which is believed to be able to improve sample efficiency by understanding the underlying generative mechanism from observational data (Schölkopf et al., 2021; Ay & Polani, 2008). Recently, some approaches learn the causal representation making use of certain causal property provided by the specified priors. One perspective is to learn feature representations with its structure specified by causal models, assuming that the raw data contains both cause variables and effect variables, like causal disentangled representation learning form images (Träuble et al., 2020; Yang et al., 2021; Shen et al., 2020; Suter et al., 2019; Dittadi et al., 2020) and physical environments (Ahmed et al., 2020; Sontakke et al., 2021). The representations have good interpretability, but probably suffer from redundancy because the learning process may generate some parts unrelated to the downstream tasks. Another perspective is to utilize the causal relation between features and labels (Suter et al., 2019; Kilbertus et al., 2018; Schölkopf et al., 2012; Wang & Jordan, 2021). They are interpreted as causal or anti-causal learning, tackling cases such as the causal direction between observational feature set $\mathbf{X}$ and downstream labels $\mathbf{Y}$ is $\mathbf{X} \rightarrow \mathbf{Y}$ and finding the cause information for



(a) Unconfounded Case

(b) Confounded Case

Figure 1: The figure demonstrates a typical case of a causal system, and we extend to two more general settings: (a) the causal graph without confounders, (b) the causal graph with confounder.

downstream prediction. This only covers a set of simple cases like handwritten numeral recognition, where the handwritten numeral $\mathbf{X}$ is produced by the number $\mathbf{Y}$ in brains. However, its effectivity is not guaranteed in circumstances where causal graph is complex and unknown previously.

In this paper, we deal with the general problem of learning minimal sufficient causal representations for downstream prediction. Considering more complex scenario like recommendation system and fault detection system. Fig. 1 (left) describes the relations among observations and predicted labels, which can be generalized to the setting of Fig. 1 (a)(b). Generally, the observational data $\mathbf{X} = [\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}, \mathbf{nc_Y}]$, which consists of a mixture of the minimal sufficient parent modes $\mathbf{pa_Y}$, non-descendants $\mathbf{nd_Y}$, descendants $\mathbf{dc_Y}$ or uncorrelated features $\mathbf{nc_Y}$ of $\mathbf{Y}$. For unconfounded case, Fig. 1 (a), we formalize a basic causal graph to describe relations among $\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}$. Note that no edge between $\mathbf{nd_Y}$ and Y exists, otherwise one can make adjustments by merging $\mathbf{nd_Y}$ into $\mathbf{pa_Y}$ or $\mathbf{dc_Y}$. In confounded case, Fig. 1 (b), we extend the basic one by considering causal edges between the nodes in $\mathbf{pa_Y}$ or $\mathbf{nd_Y}$ with $\mathbf{dc_Y}$. To learn a good representation $\mathbf{Z}$ from $\mathbf{X}$ consisting of the aforementioned components, we propose an approach from an information-theoretic perspective. Treating the information propagation along the causal graph as a natural generative process, a fact is that the information contained in root $\mathbf{pa_Y}$ suffers from degradation along causal paths, based on which we deduce inequalities on the mutual information between node pairs and design a method to learn the $\mathbf{pa_Y}$ from observational data. Based on the inequality, we deduce a tractable mutual information based causal representation learning objective. To find a robust $\mathbf{pa_Y}$, we take a perspective of adversarial attack, and propose an counterfactual vulnerable (CV) term to quantify the necessary and sufficient representation of causal information. In addition, we theoretically analyze the generalization ability under finite sample by information theory, and provide a condition that $\mathbf{Z} = \mathbf{pa_Y}$ is the minimizor of the finite-sample upper bound of a downstream prediction error over all possible solutions in the space of $\mathbf{Z}$. The theory provides an intuition for the claim that sample efficiency is improved when causal information is obtained.

Our Contribution

(i) We propose an information-theoretical approach to learn causal representation from data, using an explicit causal graphical model to describe the generative process of the real-world system.

(ii) We propose a novel quantification of the causal effect of the representation on the downstream labels by measuring the interventional vunlerability, based on which a robust learning approach is proposed correspondingly to optimize our model for causality-inspired learning. This integrates the concept of intervention in causality with the domain of representation learning.

(iii) We theoretically analyze the sample efficiency of the learning approach by giving a generalization error bound with respect to finite sample size. The theorem depicts a quantitative link between the amount of causal information contained in the learned representation, and the sample complexity of the model on downstream tasks.

(iv) Comprehensive experiments to verify the merits of method are conducted, including testing cases for the model's generalization ability when adversarial attack in representation space and distribution shift on dataset exist.

2 Related Works

Causal Representation Learning is a set of approaches to find reusable causal information and causal mechanism from observational data. Motivated by learning generalizable and stable information from data, causal representation approaches from several different perspectives have been proposed in literature. Under the framework of corss-domain learning, the pioneering work (Zhou et al., 2021; Wang & Jordan, 2021; Shen et al., 2021) consider the heterogeneity across multiple domains under the out-of-distribution settings (Gong et al., 2016; Li et al., 2018; Magliacane et al., 2017; Zhang et al., 2015). They learn causal representations from observational data by invariant causal mechanism across multi-domains. Leveraging and combining the idea of causal structure learning, some works use structural causal models to describe causal relationship inside the mixed observational data (Yang et al., 2021; Shen et al., 2020) and perform learning by minimizing a loss containing a structure learning part. On the other hand, based on the recently proposed independence between cause and mechanism principle (ICM), several work focus on the assymmetry between cause and effect. (Sontakke et al., 2021; Steudel et al., 2010) use the asymmetrical Komolgorov Complexity (Janzing & Schölkopf, 2010; Cover, 1999) relationship between cause and effect for causal learning, and similar ideas are utilized by (Parascandolo et al., 2018; Steudel et al., 2010). Compared with previous work, different from ICM, our paper considers the relationship between cause and effect from the perspective of mutual information Belghazi et al. (2018); Cheng et al. (2020), and we utilize the characteristics that cause information will reduce along the causal generative process, a phenonemon also widely know as data processing inequality (Kullback, 1997; Cover, 1999). We put our attentions on the generalization ability of the causal representation. Most existing causal inference in generalization focuses on distribution shift setting (Vapnik, 1999; Rojas-Carulla et al., 2018; Meinshausen, 2018; Peters et al., 2016). Arjovsky et al. (2019) aim at finding representation containing invariant information and analyze the generalization ability of method. Different with them, in our paper, we theoretically consider the generalization ability of causal representation under finite sample in probability approximate correctly view (PAC) (Shalev-Shwartz & Ben-David, 2014; Shamir et al., 2010), and demonstrate that causal representation has low generalization error, which also supports the claim that causal information is sample efficient.

3 PRELIMINARIES

Structural Causal Models.(Pearl, 2009) In causality, the causal graphical model is represented quantitatively by a functional model named Structural Equation Model (SCM). The variable Y is represented as a function of its parental variable $\mathbf{X}$, and additional noise $\boldsymbol{\epsilon}$,

$$Y = f(\mathbf{X}, \boldsymbol{\epsilon}), \mathbf{X} \perp \boldsymbol{\epsilon}$$

Under different assumptions, the model can be estimated from observational data up to certain degree by linear and nonlinear regression, or independence test based approaches. The information propogation is described by the functional models, which also enables the estimation of causal effect as well and causal direction under the graphical model.

Counterfactual Estimation.(Pearl, 2009) Pearl's causality gives a 3-step approach – 'abduction-action-prediction', for counterfactual estimation. In first abduction step, let $\mathbf{O} = \{\mathbf{X}, Y\}$ denote observational data, $\mathbf{X}$ is the cause of Y, we should infer the posterior of the exogenous variable $q(\boldsymbol{\epsilon}|\mathbf{O})$. Action approach is implemented by the language of do-operations, where $do(\mathbf{X} = \mathbf{x}')$ sets the variable to be a value of $\mathbf{x}'$ and the prediction approach is to observe the change of the output under such action. The counterfactual result is thus the interventional output specified by the SCMs. The information propogation along the causal path results in the observed counterfactual predictions which are later used as a base for quantifying the causality inside the system.

Mutual Information.(Cover, 1999) We explore the causal characteristics on mutual information level. Mutual information is an entropy based measure of the mutual dependence between variables.

Definition 1. The mutual information between two random variables $\mathbf{X}, \mathbf{Z}$ is define

$$I(\mathbf{X}; \mathbf{Z}) = \int_{\mathcal{Z}} \int_{\mathcal{X}} p_{XZ}(\boldsymbol{x}, \boldsymbol{z}) \log \left(\frac{p_{XZ}(\boldsymbol{x}, \boldsymbol{z})}{p_X(\boldsymbol{x}) p_Z(\boldsymbol{z})} \right) d\boldsymbol{x} d\boldsymbol{z} \tag{1}$$

Data Processing Inequality (DPI).(Cover, 1999) DPI is an information theoretic concept which describes the decreasing of the information along the Markov chain.

Definition 2. (Data Processing Inequality) If three random variables form the Markov chain $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$, there exist an inequality:

$$I(\mathbf{X}; \mathbf{Z}) \geq I(\mathbf{X}; \mathbf{Y}) \tag{2}$$

4 METHOD

In this section, we demonstrate a method to specify the minimal sufficient parents information $\mathbf{pa_Y}$ from mixed observation $\mathbf{X}$. We firstly analyze the information propogation among different causal variables under two typical causal graphs, based on which we propose an objective function with mutual information constraint for casual representation learning. To optimize the model under such function to fulfill our hypothetical causal structure, an counterfactual vulnerability based approach is also introduced.

4.1 CAUSAL REPRESENTATION LEARNING BASED ON MUTUAL INFORMATION

Denote $\mathbf{X} \in \mathcal{X}$ as d-dimensional observational data like context information or features in real-world system, and $Y \in \mathcal{Y}$ as the labels of downstream tasks. Each pair of sample $(\mathbf{x}$,$y)$ is drawn i.i.d from joint distribution $p(\mathbf{x}, y)$. As is shown in Fig. 1, we use $\mathbf{pa_Y} = [\mathbf{pa_Y^1}, \mathbf{pa_Y^2}, \cdots \mathbf{pa_Y^{p_1}}] \in \mathbb{R}^{p_1}$ to denote the variables including all observable parent nodes of Y in the causal graph. Similarly, $\mathbf{dc_Y} \in \mathbb{R}^{p_2}$ and $\mathbf{nd_Y} \in \mathbb{R}^{p_3}$ denote the descendant and non-descendant nodes of $\mathbf{Y}$, respectively. We give two possible graphical formulations of the relationship between $\mathbf{X}$ and $\mathbf{Y}$ in Fig. 1. Fig. 1 (a) is the unconfounded scenario with no confounding node between $\mathbf{pa_Y}$, Y and $\mathbf{dc_Y}$ exists, and Fig. 1 (b) shows the confounded case. The only difference lies on the additional edge from $\mathbf{pa_Y}$ to $\mathbf{dc_Y}$, since some nodes in $\mathbf{pa_Y}$ will direct point to the nodes in $\mathbf{dc_Y}$, which makes $\mathbf{pa_Y}$ confound $\mathbf{dc_Y}$ and Y. The data generative process is formulated by a Structure Causal Model (SCM) as

$$\begin{aligned} \mathbf{Y} = \mathbf{g_1}(\mathbf{pa_Y}, \boldsymbol{\epsilon}_1), \mathbf{nd_Y} = \mathbf{g_2}(\mathbf{pa_Y}, \boldsymbol{\epsilon}_2), \mathbf{dc_Y} = \mathbf{g_3}(Y, \boldsymbol{\epsilon_3}) \\ \mathbf{Pa_Y} \perp \boldsymbol{\epsilon_1}, \mathbf{Pa_Y} \perp \boldsymbol{\epsilon_2}, Y \perp \epsilon_3, \end{aligned} \tag{3}$$

where $\boldsymbol{\epsilon_1}$, $\boldsymbol{\epsilon_2}$ and ϵ_3 are assumed to be Gaussian noise, with a distribution $\mathcal{N}(\mathbf{0}, \beta\mathbf{I})$. Since the information flow among the variables described by the aforementioned causal graph is important, we consider the relationship among $\mathbf{pa_Y}$, $\mathbf{dc_Y}$ and $\mathbf{nd_Y}$ instead of considering that between $\mathbf{X}$ and $\mathbf{Y}$ only. The difficulty lies on distinguishing among $\mathbf{pa_Y}$, $\mathbf{dc_Y}$ and $\mathbf{nd_Y}$, when one only observes $\mathbf{X}$, a mixture of them. To start with, we firstly consider the unconfounded case (Fig. 1 (a)). Let $I(\mathbf{X} : Y)$ denote the mutual information between $\mathbf{X}$ and Y, and $H(\mathbf{X})$ denote the entropy of $\mathbf{X}$. From Data Processing Inequality (DPI) we can find the the inequality of mutual information inside SCM.

$$I(\mathbf{pa_Y}; \mathbf{nd_Y}, \mathbf{dc_Y}) \leq I(\mathbf{pa_Y}; \mathbf{nd_Y}, Y) \tag{4}$$

In confounded case shown in Fig. 1 (b), DPI does not apply since $\mathbf{dc_Y}$ is generated by both $\mathbf{pa_Y}$ and Y, $\mathbf{dc_Y}$. We thus decompose the mutual information in confounded case as

$$\begin{aligned} I(\mathbf{pa_Y}; \mathbf{nd_Y}, \mathbf{dc_Y}) &= I(\mathbf{pa_Y}; \mathbf{nd_Y}, Y, \mathbf{dc_Y}) - \underbrace{I(\mathbf{pa_Y};Y|\mathbf{dc_Y}, \mathbf{nd_Y})}_{\geq 0} \\ &\leq I(\mathbf{pa_Y}; \mathbf{nd_Y}, Y) + I(\mathbf{pa_Y}; \mathbf{dc_Y}|Y, \mathbf{nd_Y}) \\ &= I(\mathbf{pa_Y}; \mathbf{nd_Y},Y) + I(\mathbf{pa_Y};\boldsymbol{\epsilon_3}) = I(\mathbf{pa_Y}; \mathbf{nd_Y},Y) \end{aligned} \tag{5}$$

The equality is achieved when all the function deterministic in causal system and $H(\boldsymbol{\epsilon}) = 0$. Compared with Eq. 4, the right hand side of inequality in confounded case is with an additional entropy term $H(\mathbf{pa_Y})$. However, the information specified by the causal model is unknown when only the observation $\mathbf{X}$ is given. We develop an algorithm to learn representations based on such hypothetical structure using the presented inequalities. Let $\mathbf{Z} = \phi(\mathbf{X})$ denote representation extracted from original observation $\mathbf{X}$, where $\phi : \mathcal{X} \to \mathcal{Z}$ is the representation extraction function. For the causal system shown in Fig. 1, an important fact is that $\mathbf{pa_Y}$ is the minimal sufficient statistics of the observational data and $\mathbf{Y}$. If the mapping function maps the $\mathbf{X}$ to be $\mathbf{pa_Y}$, it is a function that finds the minimal sufficient statistics of causal system. The minimal sufficient is defined as follows:

Definition 3. (Minimal Sufficient Statistic (Lehmann & Scheffé, 2012)). Let $\mathbf{X}, Y$ be random variables. $\mathbf{Z}'$ is sufficient for Y if and only if for $\forall \mathbf{x} \in \mathcal{X}, \mathbf{z}' \in \mathcal{Z}, y \in \mathcal{Y}, p(\mathbf{x}|\mathbf{z}',y) = p(\mathbf{x}|\mathbf{z}')$. A sufficient statistic $\mathbf{Z}^*$ is minimal if and only if for any sufficient statistic $\mathbf{Z}$, there exists a deterministic function f such that $\mathbf{Z}^* = \mathbf{f}(\mathbf{Z})$ almost everywhere w.r.t $\mathbf{X}$.

In this paper, we argue that a satisfactory solution of the representation is that $\mathbf{Z}$ is equal to $\mathbf{pa_Y}$. To optimize the model under such objective based on the Eq. 4, we formulate this as a minmax optimization problem, and provide theoretical analyses for such approach. The following theorem, proven in Appendix A, illustrates the equivalence between satisfying Eq. 4 and 5 and obtaining minimal sufficient statistics.

Theorem 1. *Let $\mathbf{Z}', \mathbf{Z} \in \mathcal{Z}$, $\mathbf{Z} = \phi(\mathbf{X})$ is minimal sufficient statistics of $(Y, \mathbf{nd_Y})$ if and only if*

$$\begin{aligned} \mathbf{pa_Y} &= \arg\min_{\mathbf{Z}} I(\mathbf{Z}; \mathbf{nd_Y}, \mathbf{dc_Y}) \\ \textit{s.t.} I(\mathbf{Z}; Y, \mathbf{nd_Y}) &= \max_{\mathbf{Z}'} I(\mathbf{Z}'; Y, \mathbf{nd_Y}) \end{aligned} \tag{6}$$

By Theorem 1, when the underlying causal information is unavailable from observational data, we use above minmax approach to identify the causal information. The process of finding an optimal representation $\mathbf{z} = \phi(\mathbf{x})$ is alternatively formulated as maximizing the following Lagrangian function:

$$\delta(\phi) = \max_{\phi} \underbrace{I(\phi(\mathbf{X}); Y, \mathbf{nd_Y})}_{①} - \lambda \underbrace{I(\phi(\mathbf{X}); \mathbf{nd_Y}, \mathbf{dc_Y})}_{②} \tag{7}$$

Note that since the information of $\mathbf{nd_Y}$ and $\mathbf{dc_Y}$ is not revealed, the above objective function is not able to be optimized directly. To get an tractable form of this objective function, we firstly present the inequalities below (proven in Appendix A).

Lemma 1. *Suppose the features and labels are $\mathbf{X}, Y$, where $\mathbf{X}$ consists of the minimal sufficient parents, descendants and non-descendants as $\mathbf{X} = [\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}]$. The following inequality holds*

1. $I(\mathbf{pa_Y}; \mathbf{nd_Y}, \mathbf{dc_Y}) \leq I(\mathbf{pa_Y}; \mathbf{X})$
2. $I(\mathbf{pa_Y}; Y) \leq I(\mathbf{pa_Y}; \mathbf{nd_Y}, Y)$

The equality is achieved when all the functions in causal system is deterministic and $H(\boldsymbol{\epsilon}) = 0$. We use Lemma 1 (1) and (2) to substitute ① and ② in Eq.7 respectively, to get the transformed lower bound of the objective function. The new objective function is tractable since one no longer needs to specify $\mathbf{dc_Y}$ and $\mathbf{nd_Y}$ in advance.

$$\delta(\phi) \geq L(\phi) = \max_{\phi} I(\phi(\mathbf{X}); Y) - \lambda I(\phi(\mathbf{X}); \mathbf{X}) \tag{8}$$

The above objective function coincides with deep Information Bottleneck (IB). The difference is that IB is deduced from Rate Distortion Theorem in information theory, and it holds under the structure of Markov Chain instead of a causal graph (i.e. Fig. 1). In this paper, the IB setting is generalized into causal space, by bridging minimal sufficient statistics with root cause variables in the hypothetical causal graph. Although the tractable objective (Eq. 8) enables the optimization of the model with the mapping function ϕ, one challenge that it cannot correctly discern $\mathbf{pa_Y}$ and $\mathbf{dc_Y}$ like intractable objective Eq.7. Since $\mathbf{dc_Y}$ and $\mathbf{pa_Y}$ are both closely related to $\mathbf{Y}$, the learned $\phi(\mathbf{X})$ may be a mixture of them, which makes the representation not optimal to our expectation. In the next section, we follow the counterfactual estimation process to get sufficient and necessary cause $\mathbf{pa_Y}$.

4.2 Learning Representation by Counterfactual Estimation of PNS

In this section, we introduce a method to learn the minimal sufficient parental information of the labels. The core idea is to consider the counterfactual identifiable probability of necessary and sufficient (PNS) cause of predicted labels (Pearl, 2009; Wang & Jordan, 2021). Under this schema, obtaining the cause information is transformed into a tractable task satisfying counterfactual PNS on Y. We design a robust intervention vulnurability term to quantify the degree of satisfied PNS, and obtain a set of parameters that maximize the PNS probability, in order to find sufficient and necessary causal information.

Definition 4. (PNS Distribution Pearl (2009)) Suppose we observe a data point with representation $\mathbf{Z} = \mathbf{z}$ and label $\mathbf{Y} = \mathbf{y}$. The probability of necessary and sufficiency (pns) of $\mathbb{I}\{\mathbf{Z} = \mathbf{z}\}$ for $\mathbb{I}\{Y = y\}$:

$$\text{PNS} = \underbrace{\mathbb{E}_{p(\mathbf{Z} \neq \mathbf{z}, Y \neq y)} P(Y(\mathbf{Z} = \mathbf{z}) = y | \mathbf{Z} \neq \mathbf{z}, Y \neq y)}_{①\text{sufficient}} + \underbrace{\mathbb{E}_{p(\mathbf{Z} = \mathbf{z}, Y = y)} P(Y(\mathbf{Z} \neq \mathbf{z}) \neq y | \mathbf{Z} = \mathbf{z}, Y = y)}_{②\text{necessary}} \tag{9}$$

PNS is identified if and only if $\mathbf{Z}$ is the parent of Y (Pearl, 2009). To be more specific, sufficient (Eq. 9 ①) describes the ability of the causal representation to generate labels and necessary (Eq.

9 ②) measures the probability of negative counterfactual label if we intervene on representation $\mathbf{z}$. Although theoretically well sound, the estimation of PNS is with some difficulty since $P(Y(\mathbf{Z}=\mathbf{z})=y|\mathbf{Z}\neq\mathbf{z},Y\neq y)$ and $P(Y(\mathbf{Z}\neq\mathbf{z})\neq y|\mathbf{Z}=\mathbf{z},Y=y)$ are both counterfactual distributions, which means based on the observation that $\mathbf{Z}\mathbf{z},Yy$, the probability of $Y=y$ if $\mathbf{Z}$ is intervened to be $\mathbf{z}$. As illustrated in Preliminary, counterfactual estimation normally follows a 3-step procedure in Pearl's framework, under which we propose a constrained method for counterfactual estimation. Considering the causal generative process $\mathbf{Y}=f(\mathbf{pa_Y},\boldsymbol{\epsilon}_1)$, the $\boldsymbol{\epsilon}_1$ is regarded as a random noise perturbing the $\mathbf{pa_y}$ inside a ball with finite diameter. We treat the inference approach as the process of adversarial attack (Szegedy et al., 2013; Ben-Tal et al., 2009; Biggio & Roli, 2018) and define the 'Actions'-step in counterfactual estimation as

$$\begin{aligned}\mathbf{z}'&=\mathbf{z}+\boldsymbol{\epsilon}_1,\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta)\\ \bar{\mathbf{z}}'&=\bar{\mathbf{z}}+\boldsymbol{\epsilon}_1,\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)\end{aligned}\tag{10}$$

where $\mathcal{B}(\mathbf{z},\beta)$ is Wasserstein ball, in which the p-th Wasserstein distance (Panaretos & Zemel, 2019) W_p [1]between z and z' is smaller than β. Then, we define an counterfactual vulnerability (CV) to measure PNS from the perspective of mutual information. The defined counterfactual vulnerability term quantifies the vulnerability of prediction model (classification) under input perturbations, which is formed later. It adjusts the parameterized $\mathbf{Z}$ to fit PNS, which later can be estimated via maximum likelihood. To simplify the learning method, we consider an objective equivalent to Definition 4 as:

$$\begin{aligned}&\mathbb{E}_{p(\mathbf{Z}\neq\mathbf{z},Y\neq y)}P(Y(\mathbf{Z}=\mathbf{z})=y|\mathbf{Z}\neq\mathbf{z},Y\neq y)\\ &-\mathbb{E}_{p(\mathbf{Z}=\mathbf{z},Y=y)}P(Y(\mathbf{Z}\neq\mathbf{z})=y|\mathbf{Z}=\mathbf{z},Y=y)\end{aligned}\tag{11}$$

Definition 5. (Counterfactual Vulnerability) Let $\mathbf{Z}',\bar{\mathbf{Z}}'$ denote intervened variables on $\mathbf{Z}=\phi(\mathbf{X})$, $\forall\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta),\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)$, D and D' denote datasets sample from $p(\mathbf{z}',\mathbf{y})$ and $p(\bar{\mathbf{z}}',\mathbf{y})$, the vulnerability of robust counterfactual estimation is defined as

$$\begin{aligned}&\min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta),\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)}CV_{\mathcal{B}}\\ =&\min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta)}\frac{1}{|D|}\sum_{\mathbf{z}',\mathbf{y}}\log P(Y=y|\mathbf{Z}'=\mathbf{z}')-\min_{\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)}\frac{1}{|D'|}\sum_{\bar{\mathbf{z}}',\mathbf{y}}\log P(Y=y|\bar{\mathbf{Z}}'=\bar{\mathbf{z}}')\\ =&\min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta)}H(Y|\mathbf{Z}')-\min_{\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)}H(Y|\bar{\mathbf{Z}}')+H(Y)-H(Y)\\ =&\min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta)}I(Y;\mathbf{Z}')-\min_{\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)}I(Y;\bar{\mathbf{Z}}')\end{aligned}\tag{12}$$

Remark. We can interpret the approach under the Pearl's 'Abduction-Action-Prediction' three-level framework. 'Abduction'-step: In Definition 5, search for the worst $\boldsymbol{\epsilon}_1$ under minimum process of CV term, which capture the vulnerability in system. 'Action'-step: The Eq. 10 describes. 'Prediction'-step: we use the formulation of SCMs $y=f(\mathbf{z},\boldsymbol{\epsilon}_1)$ to predict counterfactual results.

Combining the CV term with original objective $\delta(\phi)$, we get the final objective function optimized by minmax approach. To simplify the objective function, we only consider to optimize $I(\mathbf{Z}':Y)$ rather than $I(\mathbf{Z}:Y)+I(\mathbf{Z}':Y)$ since if the worst case $I(\mathbf{Z}':Y)$ is satisfied, $I(\mathbf{Z}':Y)$ is satisfied. The robust optimization objective function is $L_{\text{rb}}\phi$, where

$$\begin{aligned}&\max_{\phi}\min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta),\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)}I(\phi(\mathbf{X});Y)-\lambda I(\phi(\mathbf{X});\mathbf{X})+CV_{\mathcal{B}}\\ \geq&L_{\text{rb}}(\phi)=\max_{\phi}\min_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta),\bar{\mathbf{z}}'\in\mathcal{B}(\bar{\mathbf{z}},\beta)}\underbrace{I(\mathbf{Z}';Y)-\lambda I(\mathbf{Z};\mathbf{X})}_{①positive}-\underbrace{I(\bar{\mathbf{Z}}';Y)}_{②negative}\end{aligned}\tag{13}$$

Our model learning is accompanlished by the minmax procedure. Intuitively speaking, the minimization procedure helps to identify the counterfactual interventional output under pertubations, or equalvalently, infer the exougeous variables form a perspective of adversarial learning. The maximization procedure is to identify a mapping function that learns the representation most likely to satisfy the PNS condition. The optimizatin of positive term aims at finding sufficient parents' information, and the negative term is to find necessary parents' information, so that the final optimization objective can extract minimal sufficient parents from observation data in high probability.

[1] $\mathrm{W}_p(\mu,\nu)=\left(\inf_{\gamma\in\Gamma(\mu,\nu)}\int_{\mathcal{Z}\times\mathcal{Z}}\Delta\left(\boldsymbol{z},\boldsymbol{z}'\right)^p d\gamma\left(z,z'\right)\right)^{1/p}$, $\Gamma(\mu,\nu$ is the collection of all probability measures on $\mathcal{Z}\times\mathcal{Z}$

Table 1: Overall Results on CPC and PCIC

Dataset	Method	p=∞				p=2			
	Metrics	AUC	ACC	advAUC	advACC	AUC	ACC	advAUC	advACC
CPC	base(robust)	0.5	0.1017	0.5	0.1017	0.5	0.1012	0.5	0.0928
	base(standard)	0.7144	0.7076	0.4506	0.4305	0.7103	0.7612	0.4489	0.4017
	IB(standard)	0.7113	0.7265	0.5776	0.6952	0.7124	0.7066	0.5647	0.6642
	r-CVAE(robust)	0.7182	0.7749	0.6694	0.7196	0.7139	0.7574	0.6611	0.7426
	r-CVAE(standard)	0.7183	0.7703	0.6467	0.6809	0.7124	0.7692	0.6402	0.691
	CaRR(robust)	**0.7271**	**0.7834**	**0.6851**	**0.7501**	**0.7224**	**0.7874**	**0.6776**	**0.7607**
	CaRR(standard)	0.7225	0.7543	0.6606	0.6995	0.7175	0.778	0.6661	0.7378
PCIC	base(robust)	0.5534	0.5875	0.5388	0.6257	0.5605	0.6498	0.5264	0.6287
	base(standard)	0.6177	0.6517	0.5231	0.589	0.6269	0.6615	0.519	0.5581
	IB(standard)	0.6242	0.6532	0.5741	0.6199	0.6216	0.6537	0.5768	0.6233
	r-CVAE(robust)	0.6261	0.6699	0.6314	0.6557	0.6324	0.6586	0.6272	0.6459
	r-CVAE(standard)	0.635	0.6537	0.6126	0.64	0.6282	0.6508	0.6161	0.64
	CaRR(robust)	0.6429	0.6787	**0.6335**	**0.6714**	**0.6465**	**0.6782**	**0.6330**	**0.6645**
	CaRR(standard)	**0.648**	**0.6802**	0.6173	0.6709	0.6361	0.6778	0.6208	0.6414

5 Sample Complexity

In this section, we theoretically analyze the proposed algorithm by the probability approximately correct (PAC) framework. We start from the perspective of information theory, and it in fact can be generalized to deep learning models. We answer the question that why causal representation containing cause information enhances the generalization ability. Different with traditional PAC learning theorem, we analyze the risk by mutual information, which follows the framework of information bottleneck (Shamir et al., 2010). We provide a finite sample bound of generalization ability. The bound measures the relationship between $I(\mathbf{Z};Y)$ and its estimation $\hat{I}(\mathbf{Z};Y)$. Here, we provide theoretical justification with following theory (proven in Appendix A):

Theorem 2. *Let $\mathbf{Z} = \phi(\mathbf{X})$, where $\phi : \mathcal{X} \to \mathcal{Z}$ be a fixed arbitrary function, determined by a known conditional probability distribution $p(\mathbf{z}|\mathbf{x})$. Let m be sample size drawn from $p(\mathbf{X}, Y)$. For any confidence parameter $0 < \delta < 1$, it holds with a probability of at least $1 - \delta$, that*

1. General case ($\mathbf{Z} = \phi(\mathbf{X})$)

$$|I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(|\mathcal{Y}|\sqrt{|\mathcal{Z}|}\log(m) + \frac{1}{2}\sqrt{|\mathcal{Z}|}\log(|\mathcal{Y}|)\right) + \frac{2}{e}|\mathcal{Y}|}{\sqrt{m}} \tag{14}$$

where $m \geq \frac{C}{4}\log(|\mathcal{Y}|/\delta)|\mathcal{Z}|\mathrm{e}^2$

2. Ideal case ($\mathbf{Z} = \phi(\mathbf{X}) = \mathbf{pa_Y}$)

$$|I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(|\mathcal{Y}|\sqrt{\beta}\log(m) + \frac{1}{2}\sqrt{|\mathcal{Z}|}\log(|\mathcal{Y}|)\right) + \frac{2}{e}|\mathcal{Y}|}{\sqrt{m}} \tag{15}$$

where $m \geq C\log(|\mathcal{Y}|/\delta)\beta\mathrm{e}^2$

Remark. The theorem provides a generalization bound under finite sample settings. It shows that when representation $\mathbf{Z}$ fully contains parents' information $\mathbf{pa_Y}$, we achieve a sample complexity bound as $m \geq C\log(|\mathcal{Y}|/\delta)\beta\mathrm{e}^2$, where β refers to Eq. 3. The minimum number of samples needed reduces from $|\mathcal{Z}|$ to β, which is a better bound since in most of cases we assume $|\mathcal{Z}| \gg \beta$. This shows that $\mathbf{z} = \mathbf{pa_Y}$ gives the reduced sample complexity and tightened generalization bound. The theorem also serves as a general solution to causality prediction problems, supporting the claim that a better prediction is achieved with causal variables, compared to that with correlated variables.

6 Experiments

In this section, we conduct extensive experiments to verify the effectiveness of our framework. In the following, we begin with the experiment setup, and then report and analyze the results.

6.1 Datasets

Our experiments are based on several real-world benchmarks. We focus on the application of recommendation system and evaluate our method on click through rate (CTR) prediction.**Yahoo!**

R3[2] is an online music recommendation dataset, which contains the user survey data and ratings for randomly selected songs. The dataset contains two part. The uniform (OOD) set and the nonuniform (I.I.D.) set. **Coat Shopping Dataset**[3] is a commonly used dataset which collected from web-shop rating on clothing. The self-selected ratings are the I.I.D. set and the uniformly selected ratings are the OOD set. **PCIC**[4] is a recently released dataset for debiased recommendation, where we are provided with the user preferences on uniformly exposed movies. **CPC** This is a dataset for CTR prediction, which includes events from real-world searching system.

6.2 Experimental Setup

Implementation of Our Method. We name our method as **CaRR**. All the objective functions are defined under information-theoretical formulation. We evaluate Eq. 13 in two parts. The first positive part (Eq. 13 ①) is evaluated by the following parameterized objective (Alemi et al., 2016):

$$I(\mathbf{Z}';Y) - \lambda I(\mathbf{Z};\mathbf{X}) \geq \mathbb{E}_D[\mathbb{E}_{\mathbf{z}'\in\mathcal{B}(\mathbf{z},\beta)}[\log p_g(y|\mathbf{z}')] - \lambda\mathcal{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}))] \tag{16}$$

we use PGD attack (Madry et al., 2017) with ∞-norm and 2-norm to get intervened Z'. We set $p_{\boldsymbol{\theta}}(\mathbf{z})$ as $\mathcal{N}(y,1)$ to avoid trivial representations. For the negative term (Eq. 13 ②), let $\bar{\mathbf{z}} = \mathbf{z} + b$, while b is hyperparameter denoting degree of bias. In our experiments, we set $b = 0.8$ when $y = 0$, $b = -0.8$ when $y = 1$. Then we use negative cross entropy to approximate mutual information. More implementation details are shown in Appendix B.

Compared Method. For all the compared methods, we use the same model architecture, with different training strategies. The model consists of representation learning module $\mathbf{z} = \phi(\mathbf{x})$ and the downstream prediction module $\hat{\mathbf{y}} = g(\mathbf{z})$, with each module implemented by neural networks. **Base** model has no additional constraints on representation, and the optimization is to minimize the cross-entropy between y and learned $\hat{y}$. We involve a recently proposed variational estimation with information bottleneck (**IB**) (Alemi et al., 2016), extend the condition VAE (CVAE (Sohn et al., 2015)) by robust training process as **r-CVAE**, whose objective function is similar with CaRR but without a negative term (Eq. 13 ②). We conduct ablation studies by comparing our proposed method **CaRR** with the r-CVAE to evaluate the effectiveness of negative term. We evaluate our method on two main aspects: (i) **Generalization** of the model under distribution shifts and (ii) **Robustness** under adversarial attack on representation space. For (i), we evaluate our method on OOD and I.I.D. setting on Yahoo! R3 and Coat. For (ii), the standard mode of adversarial attack ($\beta = 0$) means that we do not perturb original $\mathbf{z}$. In robust mode, we set $\beta = \{0.1, 0.2, 0.1, 0.3\}$ for PCIC, Yahoo! R3, Coat, and CPC, respectively.

Metrics. We use commonly used evaluation metrics AUC/ACC (Rendle et al., 2012; Gunawardana & Shani, 2009) in CTR prediction and additionally evaluate metrics AUC/ACC on advasarially perturbed evaluation dataset as adv-ACC/ adv-AUC (Madry et al., 2017).

6.3 Result Analysis

Table 1 shows overall experimental results on CPC and PCIC, based on which we find that in most cases, our method achieves better performance in terms of AUC and ACC, compared to base methods. For example, both standard and robust modes of CaRR achieve the best AUC 64.29%, 64.8% respectively, which is the best among all the compared methods on PCIC. This observation demonstrates the effectiveness of our idea. In robust training mode, our method gets best performance when adversarial metrics are considered. For example, in PCIC dataset, our method reaches 63.35%, which increases 9.47% against base methods on adv-AUC. Robust training of CaRR is also better than the standard training, winning with a margin around 1.62%. The results show that the robust learning process with exogenous variables involved enhances the adversarial performance on perturbed samples. On the other hand, in standard training mode, CaRR achieves better adversarial performance than baselines including base method and IB. Although the robust training deteriorates the performance of on normal dataset, it will help to identify the causal representation, which benefits downstream prediction under adversarial attack. For instance, we find that standard training of CaRR on PCIC has an AUC of 64.8%, which is better than the performance under robust training (64.29%). But contrary conclusions are drawn on adversarial performance. The result supports that causal representation we learned is more robust. The performance of base method in robust training mode is

[2] https://webscope.sandbox.yahoo.com/catalog.php?datatype=r

[3] https://www.cs.cornell.edu/ schnabts/mnar/

[4] https://competition.huaweicloud.com/information/1000041488/introduction

Table 2: Overall Results on Yahoo!R3-OOD and Yahoo!R3-IID

Dataset	Method	p=∞				p=2			
	Metrics	AUC	ACC	advAUC	advACC	AUC	ACC	advAUC	advACC
Yahoo!R3-OOD	base(robust)	0.5	0.4508	0.5	0.4508	0.5	0.4545	0.5	0.4537
	base(standard)	0.6198	0.6097	0.5212	0.5189	0.621	0.6099	0.5139	0.5188
	IB(standard)	0.6181	0.6063	0.5333	0.5149	0.6184	0.6069	0.5431	0.5255
	r-CVAE(robust)	0.6122	0.623	0.5883	0.5907	0.6119	0.6251	0.5866	0.5875
	r-CVAE(standard)	0.6186	0.6252	0.5836	0.584	0.6183	0.6273	0.5807	0.5794
	CaRR(robust)	0.6271	**0.6290**	**0.6008**	**0.6009**	0.6278	**0.6297**	**0.603**	**0.6021**
	CaRR(standard)	**0.6293**	0.6285	0.587	0.5862	**0.6284**	0.6297	0.5918	0.594
Yahoo!R3-IID	base(robust)	0.5	0.6001	0.5	0.5997	0.5	0.6	0.5	0.6
	base(standard)	0.7334	0.7483	0.6267	0.6251	0.7346	0.752	0.6260	0.6103
	IB(standard)	0.7291	0.7513	0.6361	0.6721	0.7348	0.7521	0.6418	0.6775
	r-CVAE(robust)	0.7431	0.7299	0.7097	0.6999	0.7382	0.7239	0.7081	0.7006
	r-CVAE(standard)	0.7428	0.7302	0.7063	0.6984	0.7416	0.7286	0.7035	0.6957
	CaRR(robust)	**0.7460**	0.7330	**0.7162**	**0.7076**	**0.7436**	**0.7291**	**0.7199**	**0.7116**
	CaRR(standard)	0.7447	**0.7344**	0.7053	0.6998	0.742	0.7276	0.7069	0.6981

worst in most of cases, indicating that robust training process will largely influence the learning of the model and ruin the prediction model. Table 2 shows the results on Yahoo! R3, which all contain I.I.D. validation and test sets, and also OOD dataset. We find that our method has a better generalization ability. For example, in Yahoo! R3 OOD, our method increases the performance by 1.9% and 8.2%, in terms of ACC and adv-ACC, compared with base method. The performance of r-CVAE is close to CaRR, since it is a modified version of our method, which only includes the positive term in Eq. 13 but with the negative one dropped. The difference between performance of CaRR and that of r-CVAE shows the effectiveness of the negative term in the objective function of CaRR. Fig. 2 demonstrates how robust training degree ($\beta = \{0.1, 0.3, 0.5, 0.7, 1.0\}$) influences the downstream prediction under adversarial settings. We conduct the experiments on the attacked real-world dataset by PGD attacker. From Fig. 2, we find that our method is better than base method, because the base model's ability on standard prediction is broken by adversarial training. When β is small, our method behaves closely to the r-CVAE in all the datasets. When β gets larger, the difference between performance of CaRR and that of r-CVAE continuously enlarges in Yahoo!R3. In PCIC, the gap becomes the largest among all when $\beta = 0.5$, and narrows down to 0 when $\beta = 0.7$. Because in our framework, we explicitly deploy a model to achieve more robust representations, while others fail.

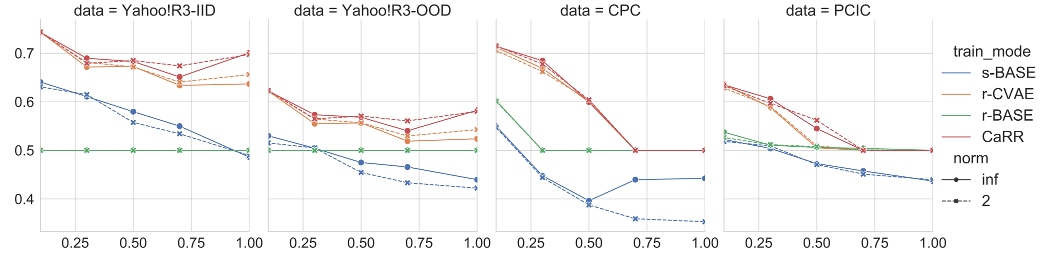


Figure 2: Results under different adversarial perturbation degree β on three datasets. Axis-x is the attack degree β. Axis-y is the adv-AUC under attacked test dataset.

7 Conclusions

In this paper, we deal with the problem of representation learning for robust deep models. We argue that when the observations contain mixed information of the labels, that is, the cause, effect and other unrelated variables, learning a representation which preserves mostly the cause information results in satisfactory generalizability of the models. By information-theoretically analyzing our hypothetical graphical model, we propose a causality-inspired representation learning method by regularized mutual information based approach, which achieves effective learning via counterfactual based model tuning, with guaranteed sample complexity reduction under certain assumptions. Extensive experiments on real data set show the effectiveness of our algorithm, supporting our claim of robust learning.

Reproducibility Statement All the supplementary materials are included in Appendix. Appendix A contains the proof of Theorem 1, Theorem 2 and Lemma 1. The implementation detail of method and the experimental results on Coat dataset are available in Appendix B and Appendix C.

REFERENCES

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Nihat Ay and Daniel Polani. Information flows in causal networks. *Adv. Complex Syst.*, 11(1): 17–41, 2008. doi: 10.1142/S0219525908001465. URL https://doi.org/10.1142/S0219525908001465.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 530–539. PMLR, 2018. URL http://proceedings.mlr.press/v80/belghazi18a.html.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1779–1788. PMLR, 2020. URL http://proceedings.mlr.press/v119/cheng20b.html.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. *arXiv preprint arXiv:2010.14407*, 2020.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848. PMLR, 2016.

Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(12), 2009.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Erich Leo Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation-part i. In *Selected Works of EL Lehmann*, pp. 233–268. Springer, 2012.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *arXiv preprint arXiv:1707.06422*, 2017.

Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pp. 6–10. IEEE, 2018.

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1933–1942, 2017.

Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pp. 4036–4044. PMLR, 2018.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*, 2020.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Chuan Shi, Binbin Hu, Wayne Xin Zhao, and S Yu Philip. Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 31(2): 357–370, 2018.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015.

Sumedh A Sontakke, Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. Causal curiosity: Rl agents discovering self-supervised experiments for causal representation learning. In *International Conference on Machine Learning*, pp. 9848–9858. PMLR, 2021.

Bastian Steudel, Dominik Janzing, and Bernhard Schölkopf. Causal markov condition for submodular information measures. *arXiv preprint arXiv:1002.4020*, 2010.

Zhu Sun, Jie Yang, Jie Zhang, Alessandro Bozzon, Long-Kai Huang, and Chi Xu. Recurrent knowledge graph embedding for effective recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 297–305, 2018.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pp. 6056–6065. PMLR, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? on the generalization of representations learned from correlated data. *arXiv e-prints*, pp. arXiv–2006, 2020.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. Joint representation learning for top-n recommendation with heterogeneous information sources. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1449–1458, 2017.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

A Theoretical Proof

A.1 Proof of Theorem 1

follows directly from the following two lemmas. We denote by $\mathcal{F}(\mathbf{X})$ the set of probabilistic functions of $\mathbf{X}$ into an arbitrary target space, and by $\mathcal{S}(Y)$ the set of sufficient statistics for Y.

Lemma 2. *Let $\mathbf{Z}$ be a probabilistic function of $\mathbf{X}$. Then $\mathbf{Z}$ is a sufficient statistic for (Y,$\mathbf{nd_Y}$) if and only if $I(Y, \mathbf{nd_Y}; \mathbf{pa_Y}) = I(Y, \mathbf{nd_Y}; \mathbf{Z}) = \max_{\mathbf{Z}' \in \mathcal{F}(\mathbf{X})} I(Y, \mathbf{nd_Y}; \mathbf{Z}')$*

Proof. The proof of Lemma is similar with Lemma 12 in Shamir et al. (2010), the difference is that we focus on the joint variable of Y, $\mathbf{nd_Y}$ rather than Y. For every $\mathbf{Z}'$ which is a probabilistic function of $\mathbf{X}$, we have Markov Chain $Y, \mathbf{nd_Y} - \mathbf{X} - \mathbf{Z}'$, so we have $I(Y, \mathbf{nd_Y}; \mathbf{X}) \geq I(Y, \mathbf{nd_Y}; \mathbf{Z}')$. We also have Markov Chain $Y, \mathbf{nd_Y} - \mathbf{Z} - \mathbf{X}$, so we have $I(Y, \mathbf{nd_Y}; \mathbf{X}) \leq I(Y, \mathbf{nd_Y}; \mathbf{Z})$ Then assume $I(Y, \mathbf{nd_Y}; \mathbf{Z}) = I(Y, \mathbf{nd_Y}; \mathbf{X})$. Since we also have Markov Chain $Y, \mathbf{nd_Y} - \mathbf{X} - \mathbf{Z}$, it follows that $Y, \mathbf{nd_Y}$ and $\mathbf{X}$ are conditionally independent given $\mathbf{Z}$, hence $\mathbf{Z}$ is a sufficient statistic.

□

Lemma 3. *Let $\mathbf{Z}$ is sufficient statistics of Y. Then $\mathbf{Z}$ is minimal sufficient statistic for $\mathbf{Y}$ if and only if $I(\mathbf{nd_Y}, \mathbf{dc_Y}; \mathbf{pa_Y}) = I(\mathbf{nd_Y}, \mathbf{dc_Y}; \mathbf{Z}) = \max_{\mathbf{Z}' \in \mathcal{S}(Y, \mathbf{nd_Y})} I(\mathbf{nd_Y}, \mathbf{dc_Y}; \mathbf{Z}')$*

Proof. First, let $\mathbf{Z}$ be a minimal sufficient statistic, and let $\mathbf{Z}'$ be some sufficient statistic. Since there is an function $\mathbf{Z} = f(\mathbf{Z}')$, we have Markov Chain $(\mathbf{nd_Y}, \mathbf{dc_Y}) - Y - \mathbf{Z}' - \mathbf{Z}$, we could have $I(\mathbf{nd_Y}, \mathbf{dc_Y}; \mathbf{Z}) \leq I(\mathbf{nd_Y}, \mathbf{dc_Y}; \mathbf{Z})$. Similarity, following the proof process of Lemma 13 in **?**, we get the above Lemma. □

A.2 Proof of Lemma 1

Proof. For the Lemma 1 (1) is hold since

$$I(\mathbf{pa_Y}; \mathbf{nd_Y}, \mathbf{dc_Y}) \leq I(\mathbf{pa_Y}; \mathbf{nd_Y}, \mathbf{dc_Y}, \mathbf{pa_Y}) \leq I(\mathbf{pa_Y}; \mathbf{X}). \tag{17}$$

For Lemma 1 (2) in main text.

Firstly, we introduce Kolmogorove Complexity $K(\mathbf{x})$, which denotes the shortest description length of string $\mathbf{x}$. Previous works give some useful technical results about Kolmogorove Complexity in causality.

Definition 6 (Algorithmic Mutual Information)**.** Let x, y be two strings. Then the algorithmic mutual information of x, y is:

$$I(\mathbf{x} : \mathbf{y}) :\stackrel{\pm}{=} K(\mathbf{y}) - K(\mathbf{y}|\mathbf{x}^*). \tag{18}$$

The mutual information is the number of bits that can be saved in the description of $\mathbf{y}$ when the shortest description of $\mathbf{x}$ is already known.

Lemma 4 (Entropy and Kolmogorov Complexity (Vapnik, 1999))**.** *Let $\mathbf{x} = \mathbf{x}_1, \mathbf{x}_2 \cdots, \mathbf{x}_n$ be a sting whose symbols $\mathbf{x}_j \in \mathcal{X}$ are drawn i.i.d. from a probability distribution $P(\mathbf{X})$ over the finite alphabet $\mathbf{X}$. Slightly overloading notation, set $P(\mathbf{x}) := P(\mathbf{x}_1) \cdots P(\mathbf{x}_n)$.. Let $H(\cdot)$ denote the Shannon entropy of a probability distribution. Then there is a constant c for n such that*

$$H(P(\mathbf{X})) \leq \frac{1}{n} E(K(\mathbf{x}|n)) \leq H(P(\mathbf{X})) + \frac{|\mathcal{X}| \log n}{n} + \frac{c}{n} \tag{19}$$

where $E(\cdot)$ is short hand for the expected value with respect to $P(\mathbf{X})$. Hence

$$\lim_{n \to \infty} \frac{1}{n} E(K(\mathbf{x})) = H(\mathbf{X}) \tag{20}$$

Lemma 5. *(Recursive Form (Janzing & Schölkopf, 2010)) Given the strings $x_1, \cdots, x_n$ and a directed acyclic graph G. Then the Kolmogrove Complexity has the recursive form:*

$$K(\mathbf{x}_1, \ldots, \mathbf{x}_n) :\stackrel{\pm}{=} \sum_{j=1}^{n} K\left(\mathbf{x}_j \mid \mathbf{pa}_j^*\right) \tag{21}$$

$$
\begin{aligned}
I(\mathbf{nd_y}, y : \mathbf{pa_y}) &:\overset{\pm}{=} K(\mathbf{nd_y}, y) - K(\mathbf{nd_y}, y|\mathbf{pa_y}^*) \\
&:\overset{\pm}{=} K(\mathbf{nd_y}, y) - K(\mathbf{nd_y}|\mathbf{pa_y}^*) - K(y|\mathbf{pa_y}^*) \\
&:\overset{\pm}{=} K(y) + K(\mathbf{d_y}) - I(\mathbf{nd_y} : y) \\
&\quad - K(\mathbf{nd_y}|\mathbf{pa_y}^*) - K(y|\mathbf{pa_y}^*) \\
&:\overset{\pm}{=} I(y : \mathbf{pa_y}) + \underbrace{I(\mathbf{nd_y} : \mathbf{pa_y}) - I(\mathbf{nd_y} : y)}_{\geq 0} \\
&:\overset{+}{\geq} I(y : \mathbf{pa_y})
\end{aligned} \tag{22}
$$

Lemma 4 already shows that

$$
\lim_{n\to\infty} \frac{1}{n} E(I(\mathbf{nd_y}, y : \mathbf{pa_y}) = I(\mathbf{nd_y}, y; \mathbf{pa_y}, \lim_{n\to\infty} \frac{1}{n} E(y : \mathbf{pa_y}) = I(y; \mathbf{pa_y}) \tag{23}
$$

We can accomplish the proof of Lemma 1 □

A.3 Proof of Theorem 2

The proof follows process in Shamir et al. (2010) Theorem 3. The sketch of proof contains two steps: (i) we decompose the original objective $|I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})|$ into two parts. (ii) for each part, we deduce the deterministic finite sample bound by concentration of measure arguments on L2 norms of random vector.

$$
|I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq |H(Y|\mathbf{Z}) - \hat{H}(Y|\mathbf{Z})| + |H(Y) - \hat{H}(Y)| \tag{24}
$$

Let $h(x)$ denote a continuous, monotonically increasing and concave function.

$$
h(x) = \begin{cases} 0 & x = 0 \\ x\log(1/x) & 0 < x \leq 1/\mathrm{e} \\ 1/\mathrm{e} & x > 1/\mathrm{e} \end{cases} \tag{25}
$$

for the term $|H(Y|Z) - \hat{H}(Y|Z)|$

$$
\begin{aligned}
|H(Y|\mathbf{Z}) - \hat{H}(Y|\mathbf{Z})| &= \left| \sum_{\mathbf{z}} (p(\mathbf{z})H(Y \mid \mathbf{z}) - \hat{p}(\mathbf{z})\hat{H}(Y \mid \mathbf{z})) \right| \\
&\leq \left| \sum_{\mathbf{z}} p(\mathbf{z})(H(Y \mid \mathbf{z}) - \hat{H}(Y \mid \mathbf{z})) \right| + \left| \sum_{\mathbf{z}} (p(\mathbf{z}) - \hat{p}(\mathbf{z}))\hat{H}(Y \mid \mathbf{z}) \right|
\end{aligned} \tag{26}
$$

For the first summand in this bound, we introduce variable $\boldsymbol{\epsilon}$ to help decompose $p(y|\mathbf{z})$, where $\boldsymbol{\epsilon}$ is independent with the parents $\mathbf{pa_y}$ (i.e. $\boldsymbol{\epsilon} \perp \mathbf{pa_y}$)

$$
\begin{aligned}
\left| \sum_{\mathbf{z}} p(\mathbf{z})(H(Y \mid \mathbf{z}) - \hat{H}(Y \mid \mathbf{z})) \right| &\leq \left| \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} (\hat{p}(y \mid \mathbf{z}) \log(\hat{p}(y \mid \mathbf{z})) - p(y \mid \mathbf{z}) \log(p(y \mid \mathbf{z}))) \right| \\
&\leq \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} h(|\hat{p}(y \mid \mathbf{z}) - p(y \mid \mathbf{z})|) \\
&= \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} h\left(\left| \sum_{\boldsymbol{\epsilon}} p(\boldsymbol{\epsilon} \mid \mathbf{z})(\hat{p}(y \mid \mathbf{z}, \boldsymbol{\epsilon}) - p(y \mid \mathbf{z}, \boldsymbol{\epsilon})) \right| \right) \\
&= \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} h(\|\hat{\mathbf{p}}(y \mid \mathbf{z}, \boldsymbol{\epsilon}) - \mathbf{p}(y \mid \mathbf{z}, \boldsymbol{\epsilon})\| \sqrt{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))})
\end{aligned} \tag{27}
$$

where $\frac{1}{m}V(x)$ denote the variance of vector x. For the second summand in Eq. 26.

$$
\left| \sum_{\mathbf{z}} (p(\mathbf{z}) - \hat{p}(z))\hat{H}(Y \mid z) \right| \leq \|\mathbf{p}(\mathbf{z}) - \hat{\mathbf{p}}(\mathbf{z})\| \cdot \sqrt{V(\hat{\mathbf{H}}(Y \mid \mathbf{z}))} \tag{28}
$$

For the summand $|H(Y) - \hat{H}(Y)|$:

$$\begin{aligned} |H(Y) - \hat{H}(Y)| &= \left| \sum_{y} p(y) \log(p(y)) - \hat{p}(y) \log(\hat{p}(y)) \right| \\ &\leq \sum_{y} h(|p(y) - \hat{p}(y)|) \\ &= \sum_{y} h \left(\left| \sum_{\mathbf{z}} \sum_{\boldsymbol{\epsilon}} p(\boldsymbol{\epsilon} \mid \mathbf{z})(p(\mathbf{z})p(y|\boldsymbol{\epsilon}) - \hat{p}(\mathbf{z})p(y|\boldsymbol{\epsilon})) \right| \right) \\ &\leq \sum_{y} h(\|\mathbf{p}(\mathbf{z})p(y|\boldsymbol{\epsilon}) - \hat{\mathbf{p}}(\mathbf{z})p(y|\boldsymbol{\epsilon})\| \sqrt{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))}) \end{aligned} \tag{29}$$

Combining above bounds we get:

$$\begin{aligned} |I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq & \sum_{y} h(\|\mathbf{p}(\mathbf{z}, y|\boldsymbol{\epsilon}) - \hat{\mathbf{p}}(\mathbf{z}, y|\boldsymbol{\epsilon})\| \sqrt{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))}) \\ & + \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} h(\|\hat{\mathbf{p}}(y \mid \mathbf{z}, \epsilon) - \mathbf{p}(y \mid \mathbf{z}, \boldsymbol{\epsilon})\| \sqrt{V(\mathbf{p}(\boldsymbol{\epsilon} \mid z))}) \\ & + \|\mathbf{p}(\mathbf{z}) - \hat{\mathbf{p}}(\mathbf{z})\| \cdot \sqrt{V(\hat{\mathbf{H}}(Y \mid \mathbf{z}))} \end{aligned} \tag{30}$$

Let $\boldsymbol{\rho}$ be a distribution vector of arbitrary cardinality, and let $\hat{\boldsymbol{\rho}}$ be an empirical estimation of $\boldsymbol{\rho}$ based on a sample of size m. Then the error $\|\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}\|$ will be bounded with a probability of at least $1 - \delta$

$$\|\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}\| \leq \frac{2 + \sqrt{2 \log(1/\delta)}}{\sqrt{m}} \tag{31}$$

Following the proof of Theorem 3 in Shamir et al. (2010), to make sure the bounds hold over $|\mathcal{Y}| + 2|$ quantities, we replace δ in Eq. 31 by $\delta/(|\mathcal{Y}| + 2|$, than substitute $\|\mathbf{p}(\mathbf{z}, y|\boldsymbol{\epsilon}) - \hat{\mathbf{p}}(\mathbf{z}, y|\boldsymbol{\epsilon})\|$ $\|\hat{\mathbf{p}}(y \mid \mathbf{z}, \boldsymbol{\epsilon}) - \mathbf{p}(y \mid \mathbf{z}, \boldsymbol{\epsilon})\|$, $\|\mathbf{p}(\mathbf{z}) - \hat{\mathbf{p}}(\mathbf{z})\|$, by Eq. 31.

$$\begin{aligned} |I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq & (2 + \sqrt{2 \log((|\mathcal{Y}| + 2)/\delta)}) \sqrt{\frac{V(\hat{\mathbf{H}}(Y \mid \mathbf{z}))}{m}} \\ & + 2|\mathcal{Y}| h \left(2 + \sqrt{2 \log((|\mathcal{Y}| + 2)/\delta)} \sqrt{\frac{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))}{m}} \right) \end{aligned} \tag{32}$$

There exist a constant C, where $2 + \sqrt{2 \log((|\mathcal{Y}| + 2)/\delta)} \leq \sqrt{C \log((|\mathcal{Y}|)/\delta)}$. From the fact that variance of any random variable bounded in [0, 1] is at most 1/4, we analyze the bound under two different cases:

In general case ($\mathbf{z} = \phi(\mathbf{x})$),

$$V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z})) \leq \frac{|\mathcal{Z}|}{4} \tag{33}$$

let m denote the number of sample, we get a lower bound of m, which is also known as sample complexity.

$$m \geq \frac{C}{4} \log(|\mathcal{Y}|/\delta) |\mathcal{Z}| \mathrm{e}^2 \tag{34}$$

In ideal case($z = pa_y$) $z \perp \epsilon$:

$$V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z})) \leq \beta \tag{35}$$

$$m \geq \frac{C}{4} \log(|\mathcal{Y}|/\delta) |\beta| \mathrm{e}^2 \tag{36}$$

$$\sqrt{\frac{C \log(|\mathcal{Y}|/\delta) V(\mathbf{p}(\epsilon \mid \mathbf{z}))}{m}} \leq \sqrt{\frac{C \log(|\mathcal{Y}|/\delta) |\mathcal{Z}|}{4m}} \leq 1/\mathrm{e} \tag{37}$$

Then, from the fact that (Shamir et al. (2010)):

$$\begin{aligned} h\left(\sqrt{\frac{\nu}{m}}\right) &= \left(\sqrt{\frac{\nu}{m}}\log\left(\sqrt{\frac{m}{v}}\right)\right) \\ &\leq \frac{\sqrt{v}\log(\sqrt{m})+1/\mathrm{e}}{\sqrt{m}}, \end{aligned} \tag{38}$$

We can get the upper bound of second summand in Eq. 32 as follows

$$\begin{aligned} &\sum_{y} h\left(\sqrt{C\log((|\mathcal{Y}|)/\delta)}\sqrt{\frac{V(\mathbf{p}(\epsilon\mid z))}{m}}\right) \\ \leq &\frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\log(m)\left(|\mathcal{Y}|\sqrt{V(\mathbf{p}(\boldsymbol{\epsilon}\mid\mathbf{z}))}\right)+\frac{2}{\mathrm{e}}|\mathcal{Y}|}{2\sqrt{m}} \end{aligned} \tag{39}$$

In general case:

$$Eq.39 \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\log(m)\left(|\mathcal{Y}|\sqrt{|\mathcal{Z}|}\right)+\frac{2}{\mathrm{e}}|\mathcal{Y}|}{2\sqrt{m}} \tag{40}$$

In ideal case:

$$Eq.39 \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\log(m)\left(|\mathcal{Y}|\sqrt{\beta}\right)+\frac{2}{\mathrm{e}}|\mathcal{Y}|}{2\sqrt{m}} \tag{41}$$

For the first summand in Eq. 32, we follow the fact (Shamir et al. (2010) Theorem 3) that:

$$V(\mathbf{H}(Y\mid\mathbf{z})) \leq \frac{|Z|\log^2(|\mathcal{Y}|)}{4} \tag{42}$$

Finally we accomplish the proof of Theorem 2.

B Experimental Details

B.1 Datasets

Yahoo! R3 The nonuniform (OOD) set contains samples of users deliberately selected and rate the songs by preference, which can be considered as a stochastic logging policy. For the uniform (I.I.D.) set, users were asked to rate 10 songs randomly selected by the system. The dataset contains 14,877 users and 1,000 items. The density degree is 0.812%, which means the dataset only records 0.812% of rating pairs.

CPC The dataset contains 85000 samples for training and 15000 samples for validation and test. The recommended item list will be exposed to query of the system by nonuniform recommendation policy. The data includes 29 dimensions of matching features, which include query and item features.

Coat The dataset was collected by an online web-shop interface.In training dataset, users were asked to rate 24 coats selected by themselves from 300 item sets. In test dataset, it collects the userrates on 16 random items from 300 item sets. Just as Yahoo! R3,the training dataset is a non-uniform dataset and the test dataset is uniform dataset. The dataset provides side information of both users and item sets. The feature dimension of user/item pair is 14/33.

PCIC The dataset was collected from a survey by questionnaire about the rate and reason why the audience like or dislike the movie. Movie features is collected form movie-review pages. The training data is biased dataset that 1000 users were asked to rate the movies they care from 1720 movies. The validation and test set is the user preference on uniformly exposed movies. The density degree is 0.241%.

For evaluation, Yahoo! R3 and Coat dataset both have two validation (include test) datasets. The I.I.D. set is 1/3 data from nonuniform logging policy, and OOD set consists of the data generated under a uniform policy. For PCIC dataset, we train our method on non-uniform datasets and perform evaluations on uniform dataset.

B.2 Implementation Details

The hyper-parameters are determined by grid search. In specific, the learning rate and batch size are tuned in the ranges of $[10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}]$ and $[64, 128, 256, 512, 1024]$, respectively. The weighting parameter λ is determined in $[0.001]$. Perturbation degree are $\beta = \{0.1, 0.2, 0.1, 0.3\}$ for Coat, Yahoo!R3, PCIC and CPC separately. The representation dimension is empirically set as 64. All the experiments are conducted based on a server with a 16-core CPU, 128g memories and an RTX 5000 GPU. The deep model architecture is shown as follows:

(1)Representation learning method $\phi(\mathbf{x})$: If dataset is Yahoo!R3 or PCIC, in which only user id and item id is the input, we should firstly use an embedding layer. The representation function architecture is.

- Concat(Embedding(user id, 32), Embedding(item id, 32))
- Linear(64, 64), ELU()
- Linear(64, representation dim), ELU()

Then for the dataset Coat and CPC, the feature dimension is 29 and 47 separately. It do not use embeding layer at first. The representation function architecture is.

- Linear(64, 64), ELU()
- Linear(64, representation dim), ELU()

(2)Downstream Prediction Model $g(\mathbf{z})$:

- Linear(representation dim, 64), ELU()
- Linear(64, 2)

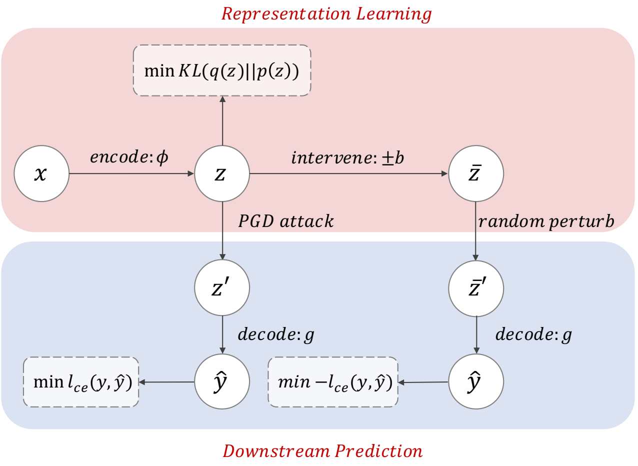


Figure 3: The figure demonstrates the model architecture of CaRR

The figure shows the model architecture. The model consists of two parts, the representation learning part and downstream prediction part. As illustrated in Section 6.2, our final objective is:

$$\mathbb{E}_D[\mathbb{E}_{\mathbf{z}' \in \mathcal{B}(\mathbf{z},\beta)}[\log p_g(\mathbf{y}|\mathbf{z}')] - \lambda \mathcal{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z})) - \mathbb{E}_{\hat{\mathbf{z}}' \in \mathcal{B}(\hat{\mathbf{z}},\beta)}[\log p_g(\mathbf{y}|\hat{\mathbf{z}}')]] \quad (43)$$

For the representation learning part, we firstly use encode function $\phi(\cdot)$ to get representation $\mathbf{z}$ and get the intervened $\hat{\mathbf{z}}$. Then we perturb the learned $\mathbf{z}$ by PGD attack procedure and perturb the $\hat{\mathbf{z}}$ by random perturbation to find the worst case correspnding to the worst downstream loss. Finally we put $\mathbf{z}'$ and $\hat{\mathbf{z}}'$ into the downstream prediction model $g(\cdot)$ to calculate y. The likelihood in Eq. 43 is estimated by cross entropy loss. Note that the perturbation approach would block the gradient propagation between representation learning process and downstream prediction by some implementation ways. Thus we use the conditional Gaussian prior $p_\theta(\mathbf{z}) = \mathcal{N}(y\mathbf{1}, \mathbf{I})$ rather than standard Gaussian distribution $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to calculate KL term. If gradient propagation is blocked, by using conditional prior, the learning process of representation $\mathbf{z}$ and exogenous $\boldsymbol{\epsilon}$ embedded in $\mathbf{z}'$ will not be influenced. The form of conditional Gaussian prior is more general $p_\theta(\mathbf{z}) = \mathcal{N}(\zeta(y), \mathbf{I})$, where $\zeta(\cdot)$ could be any non trivial function like linear function even neural network.

C Additional Results

Due to the page limit in main text, we demonstrate the additional test results on Coat dataset in this section. The table contains both IID and OOD setting, the base method on adversarial AUC and ACC achieves 70.34% and 71.04% on Coat dataset, but unsatisfactory on other dataset, possibly because the coat dataset is not sensitive to adversarial attack.

Dataset	Method	p=∞				p=2			
	Metrics	AUC	ACC	advAUC	advACC	AUC	ACC	advAUC	advACC
Coat-OOD	base(robust)	0.5586	0.5569	0.5479	0.5451	0.5593	0.556	0.5441	0.5412
	base(standard)	0.5659	0.5724	0.3874	0.4024	0.5642	0.5687	0.3128	0.3317
	IB(standard)	0.5659	0.5681	0.4701	0.4796	0.5659	0.5713	0.5442	0.5495
	r-CVAE(robust)	0.5616	0.5569	0.5595	0.5549	0.5611	0.5569	0.5585	0.5543
	r-CVAE(standard)	0.5636	0.5607	0.554	0.55	0.5632	0.5588	0.5511	0.5444
	CaRR(robust)	**0.5716**	**0.5730**	**0.566**	**0.5657**	**0.5712**	**0.5731**	**0.5653**	**0.5671**
	CaRR(standard)	0.5689	0.5710	0.5600	0.5594	0.5699	0.5695	0.5604	0.5598
Coat-IID	base(robust)	0.7156	0.7232	0.7034	0.7107	0.7195	0.7261	0.7001	0.7057
	base(standard)	0.7191	0.7217	0.4911	0.487	0.7235	0.7255	0.3642	0.3515
	IB(standard)	0.7162	0.72	0.6023	0.6017	0.7182	0.7222	0.694	0.696
	r-CVAE(robust)	0.7172	0.7243	0.7149	0.722	0.7138	0.72	0.7113	0.7176
	r-CVAE(standard)	0.7199	0.7278	0.7095	0.7169	0.717	0.7264	0.7086	0.7006
	CaRR(robust)	**0.7255**	0.7331	**0.7205**	**0.7265**	**0.7257**	**0.7325**	**0.7223**	**0.7306**
	CaRR(standard)	0.7253	**0.7350**	0.7095	0.7155	0.725	0.7321	0.7162	0.7232

Table 3: Additional results on PCIC with standard error

PCIC			AUC	std	ACC	std	adv_AUC	std	adv_ACC	std
standard	p=2	CaRR	0.6416	0.0078	0.6803	0.0014	0.619	0.004	0.6625	0.0041
		r-CVAE	0.6328	0.0023	0.6725	0.0042	0.5893	0.0419	0.6429	0.0201
	p=∞	CaRR	0.6447	0.0041	0.6817	0.0043	0.6148	0.011	0.664	0.0104
		r-CVAE	0.6358	0.014	0.6779	0.0066	0.6138	0.0062	0.6601	0.0048
robust	p=2	CaRR	0.6363	0.0045	0.6709	0.0042	0.6332	0.0024	0.6576	0.0006
		r-CVAE	0.63	0.0075	0.674	0.0069	0.6187	0.0051	0.6493	0.0013
	p=∞	CaRR	0.639	0.007	0.6761	0.0024	0.6225	0.0057	0.6638	0.001
		r-CVAE	0.6363	0.0066	0.6733	0.0058	0.6088	0.0098	0.6596	0.0124

Table 4: Additional results on Yahoo!R3 OOD with standard error

Yahoo!R3 OOD			AUC	std	ACC	std	adv_AUC	std	adv_ACC	std
standard	p=2	CaRR	0.6276	0.0001	0.6255	0.0022	0.5917	0.0071	0.5917	0.0072
		r-CVAE	0.6233	0.0005	0.6243	0.002	0.5865	0.0022	0.5872	0.0025
	p=∞	CaRR	0.629	0.0011	0.6257	0.0002	0.5966	0.0049	0.5965	0.0042
		r-CVAE	0.6253	0.0023	0.6249	0.0014	0.5855	0.0016	0.5863	0.0019
robust	p=2	CaRR	0.6242	0.0009	0.6307	0.0012	0.6008	0.0009	0.601	0.0016
		r-CVAE	0.6191	0.0013	0.6241	0.0051	0.5882	0.0014	0.5907	0.0009
	p=∞	CaRR	0.6238	0.0011	0.6284	0.0017	0.5993	0.0019	0.5999	0.0026
		r-CVAE	0.6186	0.001	0.6235	0.0028	0.5886	0.0014	0.5912	0.0012

Table 5: Additional results on Yahoo!R3 I.I.D. with standard error

Yahoo!R3 I.I.D.			AUC	std	ACC	std	adv_AUC	std	adv_ACC	std
standard	p=2	CaRR	0.7493	0.0004	0.7495	0.0015	0.7188	0.0015	0.7072	0.0013
		r-CVAE	0.7487	0.0001	0.7529	0.0027	0.7202	0.0029	0.7099	0.0027
	p=∞	CaRR	0.7497	0.0004	0.7503	0.0019	0.7191	0.0023	0.7099	0.0026
		r-CVAE	0.7488	0.0001	0.7515	0.0008	0.7191	0.0021	0.7072	0.0015
robust	p=2	CaRR	0.7374	0.0024	0.7158	0.0061	0.7247	0.0026	0.7159	0.0036
		r-CVAE	0.7376	0.0018	0.7151	0.0045	0.7194	0.0020	0.7082	0.0021
	p=∞	CaRR	0.7378	0.0015	0.7168	0.0015	0.7210	0.0031	0.7107	0.0040
		r-CVAE	0.7341	0.0007	0.7093	0.0035	0.7180	0.0017	0.7080	0.0016

Table 6: Additional results on Coat OOD with standard error

Coat OOD			AUC	std	ACC	std	adv_AUC	std	adv_ACC	std
standard	p=2	CaRR	0.5725	0.0005	0.5732	0.0005	0.5608	0.0003	0.5601	0.0004
		r-CVAE	0.5671	0.0005	0.5649	0.0006	0.5586	0.0002	0.554	0.0001
	p=∞	CaRR	0.5705	0.0013	0.5718	0.0017	0.5643	0.0001	0.5659	0.0006
		r-CVAE	0.5656	0.0005	0.5643	0.0007	0.5527	0.0074	0.5478	0.0081
robust	p=2	CaRR	0.5705	0.0015	0.5675	0.0015	0.5674	0.0002	0.565	0.0012
		r-CVAE	0.5634	0.0014	0.5591	0.0018	0.5572	0.0009	0.5522	0.0003
	p=∞	CaRR	0.5707	0.0017	0.5681	0.0024	0.5653	0.0019	0.5659	0.0011
		r-CVAE	0.5629	0.0017	0.5586	0.0028	0.559	0.0004	0.5544	0.0007

Table 7: Additional results on Coat I.I.D. with standard error

Coat I.I.D.			AUC	std	ACC	std	adv_AUC	std	adv_ACC	std
standard	p=2	CaRR	0.7248	0.0011	0.7305	0.0016	0.7069	0.0023	0.7125	0.0036
		r-CVAE	0.7129	0.0009	0.7206	0.0022	0.7023	0.0041	0.7059	0.0061
	p=∞	CaRR	0.7283	0.0013	0.7355	0.0015	0.7125	0.0007	0.7196	0.001
		r-CVAE	0.7106	0.0029	0.7184	0.0033	0.7029	0.0008	0.7106	0.0094
robust	p=2	CaRR	0.7265	0.0032	0.7331	0.0027	0.7196	0.0046	0.7261	0.0042
		r-CVAE	0.7087	0.0005	0.7169	0.0016	0.7058	0.002	0.7141	0.0036
	p=∞	CaRR	0.7276	0.0028	0.7339	0.002	0.7208	0.0023	0.727	0.0019
		r-CVAE	0.7147	0.0023	0.7222	0.0026	0.7105	0.0039	0.7181	0.0043