# HOMEWORK ON CAUSALITY

**Qixun Wang**
School of Intelligence Science and Technology
Peking University
qixun.wang@pku.edu.cn

## 1 MAIN CONTENTS

### 1.1 CAUSAL CONFUSION IN IMITATION LEARNING (DE HAAN ET AL., 2019)

This paper focuses on the causal misidentification problem in imitation learning. When there is distribution shifts between the testing and training environments, the model may not correctly capture the true causal features due to the spurious correlations between some factors in the observed state and the actions. The aim of De Haan et al. (2019) is to find the true causes. They propose a two-stage framework to address the problem. First they train a network mapping a causal graph $G$ and observations $X$ to actions $A$. $G$ is drawn uniformly at random over all $2^n$ possible causal graphs where $n$ is the dimension of the state $X = [X_1, ..., X_n]$. In order to gain robustness to distribution shifts, they propose two likelihood maximization algorithms to obtain the true causal graph $G$ for the cases that expert actions are available or not, respectively.

### 1.2 CAUSAL EFFECT INFERENCE WITH DEEP LATENT-VARIABLE MODELS (LOUIZOS ET AL., 2017)

This paper aims at learning causal effects when confounders between the interventions are the outcomes are hidden. They assume that proxies $\mathbf{t}$ of the confounders are available, as is shown in Figure 1. More concretely, the goal of the paper is to recover the individual treatment effect (ITE) as well as the average treatment effect (ATE):

$$\mathrm{ITE}(x) := \mathbb{E}[\mathbf{y} \mid \mathbf{X} = x, do(\mathbf{t} = 1)] - \mathbb{E}[\mathbf{y} \mid \mathbf{X} = x, do(\mathbf{t} = 0)], \quad ATE := \mathbb{E}[\mathrm{ITE}(x)] \tag{1}$$

Louizos et al. (2017) first prove that the recovery of ATE can be achieved by recovering the joint

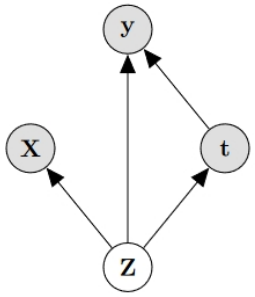


Figure 1: Causal graph considered in Louizos et al. (2017)

probability $p(\mathbf{Z}, \mathbf{x}, \mathbf{t}, \mathbf{y})$. To be specific, they prove that

$$p(\mathbf{y} \mid \mathbf{X}, do(\mathbf{t} = 1)) = \int_{\mathbf{Z}} p(\mathbf{y} \mid \mathbf{t} = 1, \mathbf{Z}) p(\mathbf{Z} \mid \mathbf{X}) d\mathbf{Z}. \tag{2}$$

To calculate $p(\mathbf{y} \mid \mathbf{t} = 1, \mathbf{Z})$ and $p(\mathbf{Z} \mid \mathbf{X})$, they propose to train a VAE to estimate the probability functions by minimizing the evidence lower bound:

$$\mathcal{L} = \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{z}_i \mid \mathbf{x}_i, t_i, y_i)} \left[ \log p\left(\mathbf{x}_i, t_i \mid \mathbf{z}_i\right) + \log p\left(y_i \mid t_i, \mathbf{z}_i\right) + \log p\left(\mathbf{z}_i\right) - \log q\left(\mathbf{z}_i \mid \mathbf{x}_i, t_i, y_i\right) \right]. \tag{3}$$

The first two terms are likelihoods (reconstruction), and the last two terms represent the KL divergence between the distribution of $\mathbf{Z}$ predicted by the encoder ("model network" in the paper) and the prior distribution of $\mathbf{Z}$. The final objective is obtained by adding two auxiliary distributions $p(t|x)$ and $p(y|x,t)$, which are for the prediction of the intervention $\mathbf{t}$ and the outcome $\mathbf{y}$ of the new sample $x$:

$$\mathcal{F}_{\text{CEVAE}} = \mathcal{L} + \sum_{i=1}^{N} \left(\log q\left(t_i = t_i^* \mid \mathbf{x}_i^*\right) + \log q\left(y_i = y_i^* \mid \mathbf{x}_i^*, t_i^*\right)\right). \tag{4}$$

After obtaining all the probabilistic functions, finally we can compute

$$q(\mathbf{Z} \mid \mathbf{X}) = \sum_{\mathbf{t}} \int q(\mathbf{Z} \mid \mathbf{t}, \mathbf{y}, \mathbf{X}) q(\mathbf{y} \mid \mathbf{t}, \mathbf{X}) q(\mathbf{t} \mid \mathbf{X}) d\mathbf{y} \tag{5}$$

and combine it with $p(y|t,z)$ to calculate $p(\mathbf{y}|\mathbf{t},\mathbf{Z})$.

### 1.3 Fairness in Decision-Making The Causal Explanation Formula (Zhang & Bareinboim, 2018)

This paper tackles the problem of detecting and distinguishing three types of discrimination, namely, direct, indirect, and spurious. They consider a causal graph with a mediator $W$, an observed confounder $Z$ (Figure 2). They present a hiring decision making example on this causal graph. In this example, if there is no discrimination, $Z$ (education background) should be the only factor affecting $Y$ (hire or not) and the religious belief $X$ and whether the applicant lives close to a religious area shouldn't affect the probability of hiring. They first reveal previous discrimination measures cannot detect the spurious discrimination, i.e., the path $X \leftarrow Z \rightarrow Y$ through which $X$ affects $Y$, though they can capture the so-called direct discrimination path $X \rightarrow Y$ and the indirect path $X \rightarrow W \rightarrow Y$. However, the discrimination does indeed exists since the total variation $TV_{x_0,x_1}(y) = P\left(y \mid x_1\right) - P\left(y \mid x_0\right) > 0$ when $Y = 1$ (hired). To achieve more fine-grained explanation and detection, Zhang & Bareinboim (2018) propose three counterfactual measures to identify the underlying discriminatory mechanisms, namely, Counterfactual Direct Effect (Ctf-DE)

$$DE_{x_0,x_1}(y \mid x) = P\left(y_{x_1,W_{x_0}} \mid x\right) - P\left(y_{x_0} \mid x\right), \tag{6}$$

Counterfactual Indirect Effect (Ctf-IE)

$$IE_{x_0,x_1}(y \mid x) = P\left(y_{x_0,W_{x_1}} \mid x\right) - P\left(y_{x_0} \mid x\right), \tag{7}$$

and Counterfactual Spurious Effect (Ctf-SE)

$$SE_{x_0,x_1}(y) = P\left(y_{x_0} \mid x_1\right) - P\left(y \mid x_0\right). \tag{8}$$

The authors further prove that Ctf-DE, Ctf-IE and Ctf-SE can identify corresponding discrimination

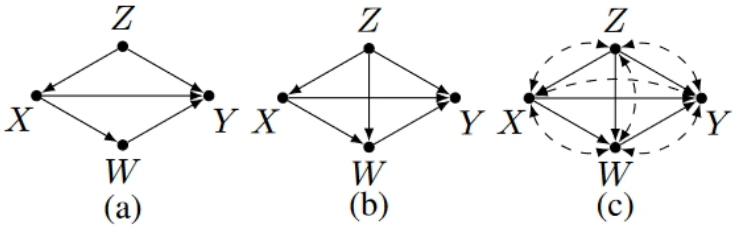


Figure 2: Causal graph considered in Zhang & Bareinboim (2018)

paths (Property 1-3 in Zhang & Bareinboim (2018)). Moreover, they prove a theorem saying that the total variation can be decomposed into the three proposed terms:

$$\begin{aligned} TV_{x_0,x_1}(y) =& SE_{x_0,x_1}(y) + IE_{x_0,x_1}\left(y \mid x_1\right) - DE_{x_1,x_0}\left(y \mid x_1\right), \\ TV_{x_0,x_1}(y) =& DE_{x_0,x_1}\left(y \mid x_0\right) - SE_{x_1,x_0}(y) - IE_{x_1,x_0}\left(y \mid x_0\right). \end{aligned} \tag{9}$$

This provides a quantitative explanation for the disparities observed in TV. It is worth noting that this theorem is non-parametric, which means that it works for any functional form of the underlying (generating) structural functions and for any distribution of the unobserved exogenous variables. Zhang & Bareinboim (2018) also show how to practically compute Ctf-DE, Ctf-IE and Ctf-SE under standard model (Figure 2 (b)), linear-standard model (the underlying structural functions are linear) and extended model (Figure 2 (c)) (Theorem 2-4 in Zhang & Bareinboim (2018)).

## 2 CORRELATIONS BETWEEN THE WORKS

Generally speaking, a common critical issue across the three papers is **modeling the probability when an intervention $x$ is imposed**, either interventional (Louizos et al., 2017; De Haan et al., 2019) or counterfactual (Zhang & Bareinboim, 2018). They adopt different approaches to achieve this goal. We summarize the characteristics of the papers in Table 1.

| | Explicit modeling of the probability | Parameterized | Observed confounders |
|---|---|---|---|
| De Haan et al. (2019) | ✗ | ✓ | ✓ |
| Louizos et al. (2017) | ✓ | ✓ | ✗ |
| Zhang & Bareinboim (2018) | ✓ | ✗ | ✓ & ✗ |

Table 1: Summary of the three papers.

In De Haan et al. (2019), they apply the notion of intervention in Functional Causal Models (FCMs) to the imitation learning setting. They argue that the distribution shifts in the state $X^t$ can be modeled by intervening on $X^t$, hence correctly modeling $p(A^t|\text{do}(X^t))$ will provide robustness to distribution shifts. They also claim that finding the true causal graph is the key to successful generalization. Here, we will further explain the relationship between the algorithm of De Haan et al. (2019), interventional distribution, and true causes. First, the training objective Equation (1) in De Haan et al. (2019) is actually modeling $p(A^t|\text{do}(X^t))$, i.e., the shifted distribution since Equation (1) takes all possible causal graphs $G$ (including the shifted and unshifted ones). Then, in order to recover the true causes, De Haan et al. (2019) propose to iteratively optimize the distribution of the causal graphs to minimize the loss w.r.t. the expert actions or maximizing the reward (like in reinforcement learning). The goal of step is recover the invariant (or true) probability $p(A^t|X^t)$. Although they do not explicitly solve $p(A^t|\text{do}(X^t))$, they use a deep network approximate it in a roundabout way. Note that the confounders ($X^{t-1}$ and $A^{t-1}$ are confounders between $X^t$ and $A^t$) De Haan et al. (2019) is observed.

In Louizos et al. (2017), their goal is to calculate $p(\mathbf{y} \mid \mathbf{X}, do(\mathbf{t} = 1))$. Different from De Haan et al. (2019), they use networks to explicitly model the probability functions. To deal with the unobserved confounder $\mathbf{Z}$, they propose a VAE to simultaneously discover the hidden confounders and infer how they affect treatment and outcome.

In Zhang & Bareinboim (2018), the target is to find an empirical measure that can identify the potential discrimination mechanisms (paths from the discrimination variable to the outcome variable). The measures mentioned in Zhang & Bareinboim (2018) often involve the the calculation of the interventional distributions, like Effect of Treatment On the Treated (ETT), Controlled Direct Effect (CDE), Natural Direct Effect (NDE), Natural Indirect Effect (NIE) and the proposed Ctf-DE, Ctf-IE, Ctf-SE. The practical computation method proposed in Zhang & Bareinboim (2018) is non-parameterized.

## 3 LIMITATIONS AND IMPROVEMENTS

### 3.1 LIMITATIONS AND IMPROVEMENTS OF DE HAAN ET AL. (2019)

**Limitations.** The expert query intervention (Algorithm 1) in De Haan et al. (2019) chooses to minimize the loss on a subset of the states $X$ with maximal variances in different causal graphs $G$. In the subsequent optimization, it simply optimizes the loss w.r.t. the expert prediction, without adding any the distributional robustness regularization. Thus the selection of the subset should be the main non-trivial source of generalization ability. However, the rationale of the selection of this state subset is unclear. According to my understanding, the author designed the selection scheme of state subset $\mathcal{S}'$ in this way because when choosing a state $X$ such that the change of $G$ can bring a large policy difference (large $D(X) = \mathbb{E}_G\left[D_{KL}\left(\pi_G(X), \pi_{mix}(X)\right)\right]$), optimizing the structure of $G$ will make it easier to cause the policy network prediction $\pi_G(X)$ to change toward the expert actions. However, this does not guarantee that $G$ ends up not using spurious causes.

**Improvements.** Improving the robustness to distribution shifts is a long-standing topic in machine learning communities. Among the topics, Out-of-Distribution (OOD) generalization and Domain Generalization (DG) are two recent hot topics. Their goal is to generalize the models well to un-

seen test distributions that differ from training ones. For addressing OOD generalization, invariant learning is one of the crucial strategies. The idea of invariant learning is to capture causal features that remain consistent across different environments[1] and have predictive power, thereby maintaining performance in the presence of distributional shifts. Numerous invariant learning methods have been proposed to tackle OOD problems in CV and NLP tasks (Arjovsky et al., 2020; Krueger et al., 2021; Bui et al., 2021; Rame et al., 2022; Shi et al., 2021; Mahajan et al., 2021; Wang et al., 2022; Yi et al., 2022). Here we adopt one classic invariant learning method V-REx (Krueger et al. (2021)) to improve the targeted intervention in De Haan et al. (2019). The idea of V-REx is to minimize the variance of loss across environments so that the model will have stable performance when facing new test distributions. The V-REx objective is:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim p_e(x,y)} \left[\ell(f_\theta(x), y)\right] + \lambda \text{Var}_{(x,y)\sim p_e(x,y)} \left[\ell(f_\theta(x), y)\right], \tag{10}$$

where Var is the variance, $\theta$ is the model parameter and $\lambda$ is a hyperparameter controlling the trade-off between the predictive power and the invariance of the model. We will introduce the V-REx objective as a regularization in Algorithm 1 in De Haan et al. (2019) by adding it to the loss function of the linear regression part:

$$\mathcal{L} \leftarrow \mathbb{E}_{s,a\sim\mathcal{T}} \left[\ell\left(\pi_G(s), a\right)\right] + \text{Var}_{s,a\sim\mathcal{T}} \left[\ell\left(\pi_G(s), a\right)\right]. \tag{11}$$

We consider each state $s$ as an environment in this imitation learning setting, since distributions of the state vary with time. Adding V-REx regularization may help to find the true causal graph $G$ since the true causes stay invariant no matter how the observed states change.

### 3.2 Limitations and Improvements of Louizos et al. (2017)

**Limitations.** This work has several limitations.

1. Cannot deals with counterfactuals.
2. Can only extract causality from low-dimension data, cannot scale to high-dimensional data, such as image classification problems.
3. It is hard to obtain massive proxies. In many real applications, however, it is prohibitively expensive or impossible to measure all the confounding factors for unbiased training. For example, it is often not affordable to annotate massively the confounding labels for the entire (attribute, text) corpus (Hu & Li, 2021).

**Improvements.** Try modeling the causal effects with as few proxies as possible.

### 3.3 Limitations and Improvements of Zhang & Bareinboim (2018)

**Limitations.** When dealing with the practical computation of the unobserved latent confounders, the proposed framework has to rely on the premise that $P(y_{x,w}|x', w')$ and $P(y_x|x')$ are identifiable, which is hard to achieve in practice. This limits the real-world application of this method.

**Improvements.** Design metrics that can detect the discrimination paths even if there are unobserved confounders. We can impose additional assumptions on the structure of the causal graphs to make the computation tractable.

### 3.4 Other Possible Future Research on Causality

I think it interesting to adopt methods in causal inference to address real-world OOD generalization problem. Among countless methods of various types that have been proposed to improve OOD generalization, finding the true causal effect is the only real guarantee of stable and good prediction performance under any data distributions (although it may be difficult to achieve in practice). There has been various of works focusing on finding the causal factors that remains invariant under distribution shifts or regularizing the network representations to capture invariant and predictive

[1] An environment in OOD generalization or DG settings represents a collection of data in which the data has some shared characteristic. For example, in PACS dataset (Li et al., 2017), the set of images with photo/art/cartoon/sketch style can each be seen as an environment.

features (Wang et al., 2022; Liu et al., 2021; Kaur et al., 2022). One very recent work Kaur et al. (2022) (ICLR 2023 top 25%) provides a comprehensive theoretical view on the relationship between causality and generalization performance. Kaur et al. (2022) emphasize the importance of modeling the data generation process. In OOD scenarios, there are many kinds of distribution shifts. For instance, the spurious features (features that cannot provide stable predictive information about the targets) may be correlated with labels or not (this corresponds to the term "correlation/concept shift" or "diversity/covariate shift" in OOD community). Different distributional shift types will induce different causal graphs of the dataset. The key contribution of Kaur et al. (2022) is that they prove that finding the correct conditional independence between the variables using d-separation of the causal graph is the necessary condition for successful OOD generalization. This result highlights the need for correctly modeling the causal graph for each OOD task. I think there are two directions worth exploring in the future: (1) finding ways to construct causal graph for OOD datasets more accurately with less human annotations (2) following Kaur et al. (2022), derive a sufficient condition for successful generalization from the view of causality.

## REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. In *ICML*, 2020.

Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. In *NeurIPS*, volume 34, pp. 21189–21201, 2021.

Pim De Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhiting Hu and Li Erran Li. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34:24941–24955, 2021.

Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. *arXiv preprint arXiv:2206.07837*, 2022.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, pp. 5815–5826. PMLR, 2021.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34:6155–6170, 2021.

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

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML*, pp. 7313–7324. PMLR, 2021.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *ICML*, pp. 18347–18377. PMLR, 2022.

Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *CVPR*, pp. 375–385, 2022.

Mingyang Yi, Ruoyu Wang, Jiacheng Sun, Zhenguo Li, and Zhi-Ming Ma. Breaking correlation shift via conditional invariant regularizer. In *ICLR*, 2022.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.