Exploiting Causal Chains for Domain Generalization

Olawale E. Salaudeen
Department of Computer Science
University of Illinois at Urbana-Champaign
oes2@illinois.edu

Oluwasanmi Koyejo
Department of Computer Science
University of Illinois at Urbana-Champaign
sanmi@illinois.edu

Abstract

Invariant Causal Prediction provides a framework for domain (or out-of-distribution) generalization – predicated on the assumption of invariant causal mechanisms that are constant across the data distributions of interest. Accordingly, given a sufficient number of distinct training distributions, the Invariant Risk Minimization (IRM) objective was proposed to learn this stable structure. However, recent work has identified the limitations of IRM when extended to data-generating mechanisms that are different from those considered in its formulation. This work considers a chain generative process where domain-specific exogenous factors influence all features – but the target is free of direct domain-specific influences. We propose a target conditioned representation independence (TCRI) constraint, which enforces the mediative effect of the observed target with respect to the causal chain of latent features we aim to identify. We empirically show a setting where this approach outperforms both Empirical Risk Minimization (ERM) and IRM.

1 Introduction

Domain generalization aims to develop models that generalize to any arbitrary distribution, provided that the distribution is structured in some expected way. A strategy that has recently received much attention is Invariant Causal Predictions (ICP; (Peters et al., 2016)), which assumes that while some aspects of the data distributions may vary across domains, the causal structure (or data-generating mechanisms), are the same. One approach that follows this strategy is Invariant Risk Minimization (IRM; (Arjovsky et al., 2019)), which proposes an objective that aims to learn a feature representation that yields a shared optimal linear predictor across domains[1]. Like other works (Rosenfeld et al., 2020), we observe settings where IRM fails to recover said predictor and propose an alternative learning strategy in these settings. In particular, we impose a different Markov property than IRM, motivated by the assumed chain data generating mechanism and show empirically that it leads to a more domain-general predictor than ERM and IRM in the linear Gaussian setting.

Contributions. This work considers the chain generative process where the estimand mediates a set of causal (predecessor) and anticausal (successor) features that are exogenously influenced by domain-specific factors – the target, however, is free of direct domain-specific influences. We show empirically that IRM fails under this data-generating process. Instead, we propose a target conditioned representation independence (TCRI) constraint, which enforces the mediative effect of the observed target on the causal chain of latent features we aim to identify. We show that this approach outperforms both Empirical Risk Minimization (ERM) and IRM in the average and worst-case on new test distributions.

2 Related Work

Machine learning algorithms are evaluated by their ability to generalize, i.e., generate reasonable predictions for unseen examples. Often, learning frameworks are designed to exploit some shared

[1] Domain and environment are used interchangeably

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

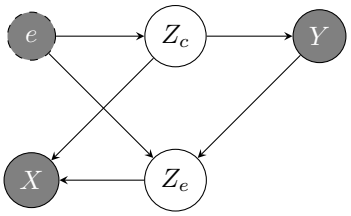

Figure 1: Graphical model depicting the structure of our data generating process - shaded nodes indicate observed variables. X represents the observed features, Y represents observed targets, and e represents domain influences. There is an explicit separation of domain-general (causal) Z_c and domain-specific (anticausal) Z_e features, causal, and anticausal, respectively.

structure between the examples available during learning and expected examples *in the wild*. Thus, a common structural assumption is that the training and testing examples are drawn independently and from the same distribution, i.e., independent and identically distributed (iid). Given the iid assumption, Empirical Risk Minimization (ERM; (Vapnik, 1991)) and its variants give strong generalization guarantees and are effective in practice. Nevertheless, many practical problems are non-stationary with respect to the train and test domain, and ERM can fail catastrophically under this setting. To address this limitation, many works have developed theories and practices for learning under distribution shift. Still, dealing with this task remains a challenge.

Following the ICP strategy (Peters et al., 2016), Arjovsky et al. (2019) propose an objective for learning representations ϕ of features x which, when conditioned on, yields a distribution on targets y that is consistent with the observed domain, that is $y|\phi(x^{e_i}) \sim y^{e_i}$ for all domains e_i. The successors of this work impose stronger assumptions of invariance, such as on higher-order conditional moments (Krueger et al., 2021). However, Rosenfeld et al. (2020) provides analysis of IRM for classification and shows a generative model where the IRM objective can fail to recover the optimal invariant predictor. They also provide the necessary conditions for IRM to work, which are difficult to satisfy in practice.

Other works have defined domain generalization by a notion of extrapolation and find that ERM remains optimal in the linear regime (Rosenfeld et al., 2021). However, there are many critical problems, such as healthcare, where robustness to worst-case distribution shifts is vital. In such settings, their notion of extrapolation only characterizes a subset of potential extrapolations.

3 Model

We consider the causal graph in Figure 1 and the equivalent structural equation model (or structural causal model ($\mathcal{SCM}$ 1; (Pearl, 2010)). We assume that the observed data is drawn from a set of E_{tr} training domains $\mathcal{E}_{tr} = \{e_i : i = 1, 2, \ldots, E_{tr}\}$, all generated from $\mathcal{SCM}$ 1, thereby fixing the mechanisms by which the observed distribution is generated:

$$\mathcal{SCM}(e) \coloneqq \begin{cases} z_c^{(e)} \sim P_{Z_c}^{(e)} \\ y^{(e)} = f_y\left(z_c^{(e)}\right) + \epsilon & \text{where } \epsilon \perp\!\!\!\perp z_c^{(e)}, \\ z_e^{(e)} = f_{z_e}\left(y^{(e)}\right) + \eta^{(e)} & \text{where } \eta^{(e)} \perp\!\!\!\perp y^{(e)}, \end{cases} \tag{1}$$

where P_{Z_c} is a probability distribution, and f_y, f_{z_e} are generative mechanisms of y and f_{z_e}, respectively. Consequently, these mechanisms are invariant across domains, i.e., $\mu_{e_i}(y \mid z_c) = \mu(y \mid z_c)$ and $\mu_{e_i}(z_e \mid y) = \mu(z_e \mid y)\, \forall e_i \in \mathcal{E}$, where $\mathcal{E}$ is the set of all possible domains.

Under the Markov assumption, we can immediately read off some properties of any distribution induced by the data generating process shown in Figure 1: (i) $e \perp\!\!\!\perp Y \mid Z_c$, (ii) $Z_c \perp\!\!\!\perp Z_e \mid Y, e$, and (iii) $Y \not\perp\!\!\!\perp e \mid X$. We consider the asymptotic setting, thus we can avoid any finite sample effects. It follows that a feature transformation that yields causal variables is sufficient to obtain a domain general predictor, though it may not be the unique or globally optimal solution on a given set of training distributions. Furthermore, as shown in Figure 1, we observe an unknown function of the latent variables z_c and z_e, $x = h(z_c, z_e)$. Generally, $\mu_{e_i}(y \mid z_e) \neq \mu_{e_j}(y \mid z_e)$ and $\mu_{e_i}(y \mid z_c, z_e) \neq \mu_{e_j}(y \mid z_c, z_e)$ for $i \neq j$, so $\mu_{e_i}(y \mid x) \neq \mu_{e_j}(y \mid x)$.

In the following sections, we will identify when invariance on the train environments implies invariance on the test environments under these model assumptions.

4 Feature Representations

One proposed approach to exploit this assumed generative model is Invariant Risk Minimization (IRM) (Arjovsky et al., 2019), which attempts to learn a mapping $\Phi : \mathcal{X} \mapsto \mathcal{H}$ (observed feature space to a latent feature space) from which an invariant predictor $\theta : \mathcal{H} \mapsto \mathcal{Y}$ (latent feature space to the target feature space) can be learned by directly optimizing the following objective:

$$\min_{\Phi,\theta} \sum_{e_i \in \mathcal{E}_{tr}} \mathcal{R}^{e_i}(\theta \circ \Phi) + \lambda \|\nabla_\theta \mathcal{R}^{e_i}(\theta \circ \Phi)\|_2,$$

where $\mathcal{R}^e$ is the environment (e) dependent risk.

We empirically show that this objective does not recover an invariant predictor when the data is generated according to $\mathcal{SCM}$ (1).

The IRM objective exploits one Markov property of the data generating process, namely $Y \perp\!\!\!\perp e \,|\, Z_c$. However, this objective alone is not enough to obtain a representation with an optimal invariant predictor – typically because there exist invariant predictors under the training distributions that use anticausal features and do not generalize on test distributions. We propose to enforce another Markov property of the graph instead, $Z_c \perp\!\!\!\perp Z_e | Y, e$, which we call target conditioned representation independence (TCRI). A representation satisfying TCRI cannot use both causal and anticausal features because the two sets are [2]d-separated by the target. Thus, enforcing this criterion restricts the feasible solutions to representations that strictly use Z_c or Z_e. Based on this, we can design models aimed at learning Z_c.

We aim to learn a domain-general representation Φ, a domain-specific representation Ψ, where $\Phi(X) \perp\!\!\!\perp \Psi(X) \,|\, Y \,\forall e_i \in \mathcal{E}_{tr}$. Clearly, $\Psi(X) \sim$ *noise* satisfies this property, so we also need additional constraints on Ψ. The natural constraint is that Ψ should yield a representation on which a non-trivial domain-specific predictor can be learned – meaning that the feature transformation by Φ is correlated with y.

5 Method

Our proposed objective contains three terms, each related to the properties desired of the learned representations, as follows,

$$\mathcal{L} = \mathcal{L}_\Phi + \beta \mathcal{L}_\Psi + \rho TCRI, \tag{2}$$

where β and ρ are hyperparemeters.

We let $\mathcal{L}_\Phi$ be the average empirical risk generated by Φ, θ_c across training distributions, where $\Phi : \mathcal{X} \mapsto \mathcal{H}_\Phi$, $\theta_c : \mathcal{H}_\Phi \mapsto \mathcal{Y}$. This term aims to learn a representation that yields a good linear predictor on average across training distributions:

$$\mathcal{L}_\Phi = \frac{1}{E_{tr}} \sum_{e_i \in \mathcal{E}_{tr}} \mathcal{R}^{e_i}\left(\theta_c \circ \Phi\right),$$

where $\mathcal{R}^{e_i}$ denotes the empirical risk achieved on domain e_i.

We also learn a domain-specific representation Ψ, which is constrained to be (i) conditionally independent of our (fixed) domain-general representation given the target and environment, and (ii) amenable to a good domain-specific predictor. We save discussion of (i) for later in this section. With respect to (ii), given a domain-specific representation, $\Psi : \mathcal{X} \mapsto \mathcal{H}_\Psi$, we define a set of domain-specific predictors $\{\theta_{e_i} : \mathcal{H}_\Psi \mapsto \mathcal{Y} : i = 1, \ldots, E_{tr}\}$. We enforce that Ψ maps to a feature space from which a good domain-specific linear predictor can be learned:

$$\mathcal{L}_\Psi = \sum_{e_i \in \mathcal{E}_{tr}} \mathcal{R}^{e_i}\left(\theta_{e_i} \circ \Psi\right).$$

To enforce conditional independence between the invariant and variant representation, given the outcome, we consider the Hilbert-Schmidt Independence Criterion (HSIC). We first remove the effect of the outcome $Y^{(e_i)}$ from both representations $\Phi(X^{(e_i)})$ and $\Psi(X^{(e_i)})$ – by regressing it out for continuous $\mathcal{Y}$.

We use the V-statistic-based HSIC estimate, as an independence test on these residuals, for the TCRI constraint (more details on HSIC can be found in Gretton et al. (2007)). For two generic random

[2]Random variables X, Y are said to be d-separated by Z if $X \perp\!\!\!\perp Y \,|\, Z$.

Table 1: Relative ratio of mean squared error, computed as $\frac{\text{Algorithm}}{\text{ERM}}$. Unlike accuracy, the raw metrics are uninformative, so we elect to illustrate relative performance via these ratios, using Empirical Risk Minimization (ERM) as a baseline. Smaller $\downarrow$ is better. We compute the error for each domain independently. Average here refers to the average error across observed domains, with respect to both the train and test set. Worst Case similarly refers to the largest error on a single domain.

Algorithm	**Average**		**Worst Case**	
	Train	**Test**	**Train**	**Test**
ERM	baseline			
IRM	1.08	1.24	1.92	2.16
TCRI (ours)	1.38	1.20	0.16	0.11
causal (oracle)	1.29	1.13	0.11	0.06

variables X, Y for which we want to determine independence ($X \perp\!\!\!\perp Y$?), define

$$\widehat{HSIC}(X, Y) = \frac{1}{n^2}\text{trace}(\mathbf{K}_{XX'}\mathbf{H}_n\mathbf{K}_{YY'}\mathbf{H}_n),$$

where $\mathbf{K}_{XX'} \in \mathbb{R}^{n\times n}$, $\mathbf{K}_{YY'} \in \mathbb{R}^{n\times n}$ are Gram matrices, $\mathbf{K}^{i,j}_{XX'} = \phi(X_i, X_j)$, $\mathbf{K}^{i,j}_{YY'} = \psi(Y_i, Y_j)$, $\mathbf{H}_n = \frac{1}{n}\mathbf{I}_n\mathbf{I}_n^\top$ is a centering matrix, $\mathbf{I}_n$ is the $n \times n$ dimensional identity matrix, $\mathbf{1}_n$ is the n-dimensional vector whose elements are all 1, and $^\top$ denotes the transpose. Alternatively, one may use the conditional cross-covariance when appropriate (linearity and Gaussianity). For three generic random variables X, Y Z for which we want to determine conditional independence ($X \perp\!\!\!\perp Y \mid Z$?), define

$$\Sigma_{X,Y|Z} = \Sigma_{XY} - \Sigma_{XZ}\Sigma^{-1}_{ZZ}\Sigma_{YZ}.$$

All together, the objective function (2) is given by:

$$\min_{\Phi,\Psi,\theta_c,\theta_1,\theta_2,\ldots,\theta_{E_{tr}}} \sum_{e_i\in\mathcal{E}_{tr}} \left[\mathcal{R}^{e_i}\left(\theta_c \circ \Phi\right) + \beta\mathcal{R}^{e_i}\left(\theta_{e_i} \circ \Psi\right) + \rho TCRI^{e_i}\left(\Phi, \Psi\right)\right], \tag{3}$$

where $TCRI$ may be given by $\widehat{HSIC}$ or the norm of the conditional cross-covariance.

After minimizing this objective, only the invariant representation and its predictor, $\theta_c \circ \Phi$, are used for prediction.

Remark 1. *(Representations induced by TCRI) It is not clear that Φ and Ψ are distinguishable in general. It may be possible to find Ψ, strictly using Z_e, that yields a better invariant predictor than the corresponding Φ strictly using Z_c. However, we did not observe such a case in our experiments.*

6 Experiments

We present results on a simulated Linear Gaussian SEM. Additional evaluation is left for an extended version. We keep the same structure as the $\mathcal{SCM}$ (1) where

$$z_c^{(e)} \sim \mathcal{N}\left(0, (\sigma_c^2)^e I_{d_c}\right), \qquad f_y(z_c) = z_c^{(e)}\alpha, \qquad f_{z_e}(y) = y^{(e)}\gamma$$

$$\epsilon \sim \mathcal{N}\left(0, (\sigma_\epsilon^2)^{(e)}\right), \qquad \eta \sim \mathcal{N}\left(0, (\sigma_\eta^2)^{(e)} I_{d_e}\right).$$

We generate E_{tr} and E_{te} environments, train and test respectively. We let $h(z_c, z_e)$ be a concatenation of the two features sets and let all functions be linear, $\Phi, \Psi : \mathbb{R}^{d_c+d_e} \mapsto \mathbb{R}$ – where d_c, d_e are the dimensions of the causal and anticausal latent variables respectively. We also let $\theta_c, \theta_e : \mathbb{R} \mapsto \mathbb{R}$ but fix $\theta_c : \theta(x) \mapsto x$ to be a dummy predictor. We also use mean square error as our loss function and the $l-1$ norm of the conditional cross-covariance as the TCRI constraint.

We delineate environments via parameters $(\sigma_\epsilon^2)^{(e_i)} = e_i$, $(\sigma_\eta^2)^{(e_i)} = e_i \cdot I_{d_e}$, where I is identity, and set all other model parameters to one, i.e., $\alpha = [1]^{d_c}$ and $\gamma = [1]^{d_e}$. We let $d_c = d_e = 1$, then randomly select $E_{tr} = 2$ environment parameters, $e_1^{tr}, e_2^{tr} \in [1, 3]$, and select $E_{te} = 100$ with $e_1^{te}, \ldots, e_{100}^{te} \in [3, 20]$ as the testing environment parameters – 20 is arbitrarily chosen to capture the

behavior of error w.r.t environment parameters. We choose $E_{tr} = 2$ to show the limiting case of minimal training environments and $d_c = d_e = 1$ to more clearly examine the feature transformation.

We find that the general trends of relative errors shown in Table 1hold for any arbitrary selection of train and test distributions parameters. The IRM approach outperforms both TCRI and the true causal model on average across all observed training distributions; however, it necessarily exploits the non-domain general feature to do this – having a higher coefficient for the anticausal dimension compared to the others. In new environments, however, IRM yields greater error than ERM, matching existing results, e.g., Rosenfeld et al. (2020). Our proposed TCRI constraint and the true causal model perform best on new test distributions in both average and worst cases. In our experiments, TCRI did not always recover the exact causal parameters; however, it consistently yielded a closer model than IRM.

7 Conclusion and Future Work

Domain (out-of-distribution) generalization remains an essential and unsolved task, and Invariant Causal Prediction continues to be a promising strategy to achieve this. We exploit the mediative effect of the target on the causal and anticausal features from the assumed causal chain data-generating mechanism and propose an objective function that enforces this property as a means to learn a feature representation that maps to the domain-general causal features. We empirically show this method outperforms Empirical Risk Minimization (ERM) and Invariant Risk Minimization (IRM) in the scalar linear Gaussian $\mathcal{SCM}$ setting when considering average and worst-case error on new test distributions. Our future work includes more empirical evaluation of a variety of real-world datasets.

References

Martín Arjovsky, L. Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

A. Gretton, K. Fukumizu, C. Teo, Le Song, B. Schölkopf, and Alex Smola. A kernel statistical test of independence. In *NIPS*, 2007.

David Krueger, Ethan Caballero, J. Jacobsen, A. Zhang, Jonathan Binas, Rémi Le Priol, and Aaron C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, 2021.

J. Pearl. Causal inference. In *NIPS Causality: Objectives and Assessment*, 2010.

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)*, pages 947–1012, 2016.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. An online learning approach to interpolation and extrapolation in domain generalization. *arXiv preprint arXiv:2102.13128*, 2021.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *NIPS*, volume 91, pages 831–840, 1991.