Exploiting Causal Chains for Domain Generalization

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, the Invariant Risk Minimization (IRM) objective has been proposed to learn this stable structure, given sufficient training distributions. Unfortunately, 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 the generative process with causal (predecessor) and anticausal (successor) features where environment-specific exogenous factors influence all features – but the target is free of direct environment-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.

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; [5]), which assumes that some aspects of the data distributions may vary across environments, but the causal structure (or data-generating mechanisms) are the same. One approach that follows this strategy is Invariant Risk Minimization (IRM) [1], whose proposed objective aims to learn a feature representation that yields a shared optimal linear predictor in each environment. Like other works [7], we observe settings where IRM fails and we propose an alternative learning strategy in these settings. In particular, we impose a different Markov property than IRM from the assumed data generating mechanism and show empirically that it leads to a better invariant predictor than IRM in the linear Gaussian setting.

Contributions. This work considers the generative process with causal and anticausal features where environment-specific exogenous factors influence all features – but the target is free of direct environment-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 both 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 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; [8]) and its variants give strong generalization guarantees and are effective in practice. Nevertheless, many practical problems are non-stationary with respect to the

Preprint. Under review.

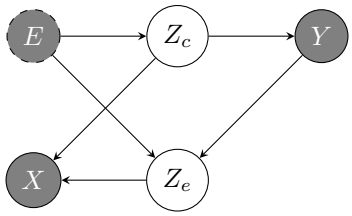

Figure 1: Graphical model depicting the structure of our data generating process - shaded nodes indicate observed variables. Note that there is an explicit separation of domain general (causal) Z_c and domain specific (anticausal) Z_e features, causal and anticausal respectively. X represents the observed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

train and test domain, and ERM can fail catastrophically under this setting. To address this limitation, many works have developed the theory and practice of learning with distribution shift. Still, dealing with distribution shifts remains a challenge.

Arjovsky et al. [1] propose an objective for learning representations ϕ which, when conditioned on, yields a distribution on targets y consistent with the observed domain, that is $y|\phi(x) \sim y^e$ for domain e. The successors of this work impose stronger assumptions of invariance of higher order conditional moments [3]. However, Rosenfeld et al. [7] provide analysis of IRM for classification and show 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. Other works have defined domain generalization by a notion of extrapolation and find that ERM remains optimal in the linear regime [6]. However, there are many critical problems, such as in healthcare, where robustness to worst case distribution shifts is critical. 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; [4]). We assume that the observed data is drawn from a set of E training environments $\mathcal{E} = \{e : 1, 2, \ldots, E\}$, all generated from $\mathcal{SCM}$ 1, thereby fixing the mechanisms by which the observed distribution is generated:

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

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., the environment-specific distributions $\mu_e(y \,|\, z_c) = \mu(y \,|\, z_c)\, \forall e \in \mathcal{E}$ and $\mu_e(z_e \,|\, y) = \mu(z_e \,|\, y)\, \forall e \in \mathcal{E}$.

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) $E \perp\!\!\!\perp Y \,|\, Z_c$, (ii) $Z_c \perp\!\!\!\perp Z_e \,|\, Y, E$, and (iii) $Y \not\perp\!\!\!\perp E \,|\, X$. We consider the asymptotic setting, thus we can avoid any finite sample effects. It follows that a feature transformation that only uses causal variables is sufficient for domain generalization, though it may not be the unique or globally optimal solution on a 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, $\mu_e(y \,|\, z_e) \neq \mu_{e'}(y \,|\, z_e)$ and $\mu_e(y \,|\, z_c, z_e) \neq \mu_{e'}(y \,|\, z_c, z_e)$ for $e \neq e'$.

The question remains, under these model assumptions, when does invariance on the train environments imply invariance on the test environments

4 Feature Representations

One proposed approach to exploit this assumed generative model is Invariant Risk Minimization (IRM) [1], which attempts to learn a mapping $\Phi : \mathcal{X} \mapsto \mathcal{H}$ from the observed feature space to an latent feature space from which an invariant predictor $\theta : \mathcal{H} \mapsto \mathcal{Y}$, from the latent feature space to the

target feature space, can be learned by directly optimizing the objective:

$$\min_{\Phi,\theta} \sum_{e\in\mathcal{E}} \mathcal{R}^e(f_\theta \cdot \Phi) + \lambda \|\nabla_\theta \mathcal{R}^e(f_\theta \cdot \Phi)\|_2,$$

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

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

The IRM objective aims to exploit one Markov property of the data generating process, namely $Y \perp\!\!\!\perp E \,|\, Z_c$. However, this objective alone is evidently 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 desired invariant representation instead, $Z_c \perp\!\!\!\perp Z_e | Y, E$, which we call target conditioned representation independence (TCRI). A representation satisfies TCRI cannot use both causal and anticausal features because the two sets are [1]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 a proxy of Z_c.

Specifically, in addition to learning a domain general representation Φ, we also aim to learn the domain specific representation Ψ, where $\Phi(X) \perp\!\!\!\perp \Psi(X) \,|\, Y, E$. 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 - non trivial meaning correlated with y.

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 were not able to find such a case in our experiments.*

5 Method

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

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

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 loss aims to learn a representation that yields a good linear predictor on average across training distributions.

We also learn a domain specific representation Ψ, which is constrained to be (i) conditionally independent of our 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, $\Psi : \mathcal{X} \mapsto \mathcal{H}_\Phi$, we define a set of domain specific predictors $\{(\theta_e : \mathcal{H}_\Phi \mapsto \mathcal{Y}) : e = 1, \ldots, E\}$. We enforce that Ψ maps to a feature space from which a good domain specific linear predictor can be learned, i.e.:

$$\mathcal{L}_\Phi = \sum_{e\in\mathcal{E}} \mathcal{R}^e\left(\theta_c \circ \Phi\right), \qquad \mathcal{L}_\Psi = \sum_{e\in\mathcal{E}} \mathcal{R}^e\left(\theta_e \circ \Psi\right),$$

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

To enforce conditional independence between the invariant and variant representation, given the outcome, we use the Hilbert-Schmidt Independence Criterion (HSIC; [2]). We first remove the effect of the outcome $Y^{(e)}$ from both representations $\Phi(X^{(e)})$ and $\Psi(X^{(e)})$ - by regressing out for continuous $\mathcal{Y}$ - and call the residuals $\tilde{\Phi}(X^{(e)})$ and $\tilde{\Psi}(X^{(e)})$, respectively.

We use the V-statistic-based HSIC estimate; more details can be found in [2]:

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

[1]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.

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

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.

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

$$\min_{\Phi,\Psi,\theta_c,\theta_1,\theta_2,\ldots,\theta_E} \sum_{e\in\mathcal{E}} \left[\mathcal{R}^e\left(\theta_c \circ \Phi\right) + \beta\mathcal{R}^e\left(\theta_e \circ \Psi\right) + \rho\widehat{HSIC}\left(\tilde{\Phi}^e, \tilde{\Psi}^e\right)\right]. \tag{3}$$

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

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 $z_c^{(e)} \sim \mathcal{N}\left(\mu_c^e, (\sigma_c^2)^e\right)$, $f_y(z_c) = \left\langle z_c^{(e)}, \alpha\right\rangle$, $f_{z_e}(y) = \left\langle y^{(e)}, \gamma\right\rangle$, $\epsilon \sim \mathcal{N}\left(\mu_\epsilon^{(e)}, (\sigma_\epsilon^2)^{(e)}\right)$, and $\eta \sim \mathcal{N}\left(\mu_\eta^{(e)}, (\sigma_\eta^2)^{(e)}\right)$. We generate E_{tr} and E_{te} environments, train and test respectively, with $\mu_\epsilon^e = \mu_\eta^e = 0$ and $(\sigma_\epsilon^e)^{(e)} = (\sigma_\eta^e)^{(e)}$. We let all models be linear functions, $\Phi, \Psi : \mathbb{R}^{d_c+d_e} \mapsto \mathbb{R}$ - d_c, d_e are the dimensions of the causal and anticausal latent variables respectively. We let $\theta_c, \theta_e : \mathbb{R} \mapsto \mathbb{R}$ but fix $\theta_c : \theta(x) \mapsto x$ to be a dummy predictor.

We delineate environments via parameters $(\sigma_\epsilon^e)^{(e)} = (\sigma_\eta^e)^{(e)} = e$, and we set all other model parameters to one. We then randomly select $E_{tr} = 2$ environment parameters, e, between 1 and 3, and select E_{te} intervals between 3 and 20 as the testing environment parameters, 20 arbitrarily chosen to capture asymptotic behavior of error w.r.t environment parameters.

We find that, in the worst case, the general trends of relative errors shown in Table 1 hold for any arbitrary selection of train and test distributions parameters. The IRM approach outperforms both TCRI and the true causal model on average across the training distributions, however, it necessarily exploits non-domain general features to do this. On new environments, however, IRM actually yields greater error than ERM, matching existing results e.g. Rosenfeld et al. [7]. Our proposed TCRI, and the true causal model perform best in both average and worst case.

7 Conclusion

Domain (out-of-distribution) generalization remains an important and unsolved task, and Invariant Causal Prediction continues to be a promising strategy to achieve this. We observe the mediative effect of the target on the causal and anticausal features, as a result of the causal chain of the data generating mechanisms, and design a model and objective function that enforces this property. We empirically show the improvements of this contribution outperform Empirical Risk Minimization (ERM) and Invariant Risk Minimization (Minimization) in the scalar linear Gaussian $\mathcal{SCM}$ setting. Our future work includes more empirical evaluation on a variety of real-world datasets.

References

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

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

[3] 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.

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

[5] J. Peters, Peter Buhlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *arXiv: Methodology*, 2015.

[6] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. An online learning approach to interpolation and extrapolation in domain generalization. *ArXiv*, abs/2102.13128, 2021.

[7] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *ArXiv*, abs/2010.05761, 2021.

[8] V. Vapnik. Principles of risk minimization for learning theory. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1992. URL `https://proceedings.neurips.cc/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf`.