A UNIFIED CAUSAL VIEW OF DOMAIN INVARIANT REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning methods can be unreliable when deployed in domains that differ from the domains on which they were trained. One intuitive approach for addressing this is to learn representations of data that are domain-invariant in the sense that they preserve data structure that is stable across domains, but throw out spuriously-varying parts. There are many approaches aimed at this kind of representation-learning, including methods based on data augmentation, distributional invariances, and risk invariance. Unfortunately, it is often unclear when a given method actually learns domain-invariant structure, and whether learning domain-invariant structure actually yields robust models. The key issue is that, in general, it's unclear how to formalize "domain-invariant structure". The purpose of this paper is to study these questions in the context of a particular natural domain shift notion that admits a natural formal notion of domain invariance. This notion is a formalization of the idea that causal relationships are invariant, but non-causal relationships (e.g., due to confounding) may vary. We find that whether a given method learns domain-invariant structure, and whether this leads to robust prediction, both depend critically on the true underlying causal structure of the data.

1 INTRODUCTION

Even when machine learning models have excellent performance on held-out data, they can perform poorly when deployed in the wild. Part of the problem is *domain shift*, a structural mismatch between the domain(s) used to train the model, and the domain(s) in which the model is deployed. A variety of techniques and heuristics have been proposed to mitigate domain shift problems. One popular class of approach—which we'll focus on in this paper—is to try to learn a representation of the data that is in some sense "invariant" across domains. Informally, the aim of such methods is to find a representation that captures the part of the data structure that is the "same" in all domains while discarding the part that varies across domains. It then seems intuitive that a predictor trained on top of such a representation would have stable performance even in new domains. Despite the intuitive appeal, there are fundamental open questions: when does a given method actually succeed at learning the part of the data that is invariant across domains? And, when does learning a domain invariant representation actually lead to more robust out-of-domain predictions?

There are many methods aimed at domain-invariant representation learning. When applied to broad ranges of real-world domain-shift benchmarks, there is no single dominant approach, and an attack that works well in one context may is often worse than vanilla empirical risk minimization in another (i.e., just ignoring the domain-shift problem). We'll study three broad classes of method:

Data augmentation Each example is perturbed in some way and we learn a representation that is the same for all perturbed versions. E.g., if $t(X)$ is a small rotation of an image X, then $\phi(X) = \phi(t(X))$. For example, (Krizhevsky et al., 2012; Hendrycks et al., 2019; Cubuk et al., 2019; Xie et al., 2020; Wei & Zou, 2019; Paschali et al., 2019; Hariharan & Girshick, 2017; Sennrich et al., 2015; Kobayashi, 2018; Nie et al., 2020).

Distributional invariance We learn a representation so that some distribution involving $\phi(X)$ is constant in all domains. There are three such distributional invariances that can be required to hold for all domains e, e':

marginal invariance: $\mathrm{P}^e(\phi(X)) = \mathrm{P}^{e'}(\phi(X))$ (Muandet et al., 2013; Ganin et al., 2016;

Albuquerque et al., 2020; Li et al., 2018a; Sun et al., 2017; Sun & Saenko, 2016; Matsuura & Harada, 2020);

conditional invariance: When Y is a label of interest, $\mathrm{P}^e(\phi(X) \mid Y) = \mathrm{P}^{e'}(\phi(X) \mid Y)$ (Li et al., 2018b; Long et al., 2018; Tachet des Combes et al., 2020; Goel et al., 2020)

sufficiency: $\mathrm{P}^e(Y \mid \phi(X)) = \mathrm{P}^{e'}(Y \mid \phi(X))$ (Peters et al., 2016; Rojas-Carulla et al., 2018; Wald et al., 2021).

Risk minimizer invariance For supervised learning, we learn a representation $\phi(X)$ so that there is a fixed (domain-independent) predictor w^* on top of $\phi(X)$ that minimizes risk in all domains (Arjovsky et al., 2019; Lu et al., 2021; Ahuja et al., 2020; Krueger et al., 2021; Bae et al., 2021).

In each case, the aim is to learn a representation that throws away information that varies 'spuriously' across domains while preserving information that is reliably useful for downstream tasks. However, the notion of what is thrown away is substantively different across all the approaches and it is unclear which, if any, is appropriate for any particular problem.

The principle challenge to answering our motivating questions is that it's unclear in general how to formalize the idea of "part of the data that is invariant across domains". To make progress, it is necessary to specify the manner in which different domains are related. In particular, we require an assumption that is both reasonable for real-world domain shifts and that precisely specifies what structure is invariant across domains. In many problems, it is natural to assume that causal relationships—determined by the unobserved real-world dynamics underpinning the data—should be the same in all domains. We'll use an assumption of this form; the first task is to translate it into a concrete notion of domain shift.[1]

Specializing to supervised learning with label Y and covariates X, we proceed as follows. The covariates X are caused by some (unknown) factors of variation. These factors of variation are also dependent with Y. For some factors of variation, jointly denoted as Z, the relationship between Y and Z is spurious in the sense that Y and Z are dependent due to an unobserved common cause. The distribution of this unobserved common cause may change across environments, which in turns means the relationship between Y and Z can shift. However, the structural causal relationships between variables will be the same in all environments—e.g., $\mathrm{P}(X \mid \mathrm{pa}(X))$ is invariant, where $\mathrm{pa}(X)$ denotes the (possibly unobserved) causal parents of X. We call a family of domains with this structure *Causally Invariant with Spurious Associations* (CISA).

Concretely, consider the problem of classifying images X as either Camel or Cow Y. In training, the presence of sand in the image background Z is strongly associated with camels. But, we may deploy our new classifier in an island domain where cows are frequently on beaches—changing the Z-Y association. Nevertheless, the causal relationships between the factors of variation—Y, Z, and others such as camera type or time of day—and the image X remain invariant.

In this example, a natural formalization of "domain invariant part" of the image is the part that does not changes if sand Z is added or removed from the background; an invariant representation learning method should learn a $\phi(X)$ that throws away such information. The aim of CISA is to be a reasonably broad notion of domain-shift that also allows us to formalize this intuitive notion of domain-invariance. Namely, under CISA, we define the domain-invariant part of X to be the part that is not causally affected by the spurious factors of variation Z. Accordingly, a representation ϕ is domain invariant if $\phi(X)$ is not causally affected by Z. We say a representation with this property is *counterfactually invariant to spurious factors* (CF-invariant for short).

We now return to the motivating questions: when does a given method actually succeed at learning the part of the data that is invariant across domains? And, when does learning a domain invariant representation actually lead to more robust out-of-domain predictions? In the context of CISA-compatible domain shifts, we can answer the first question by determining the conditions under which each approach learns a CF-invariant representation, and the second question by studying the relationship between CF-invariance and domain shift performance.

Informally, the technical contributions of the paper are:

[1]This kind of causal-invariance assumption is already used in the domain-shift literature, though the formal domain shift notion we'll use here differs from previous approaches, see Section 2.

1. Formalization of CISA and Counterfactually Invariant Representation Learning.
2. A characterization of the causal (real-world) conditions where each domain invariant representation learning method yields CF-invariant representations. This turns out to yield both a hierarchy of methods, and guidance for which methods are appropriate for a given problem.
3. Results relating invariant learning to robust learning. We'll find that, in general, there is not a simple relationship. For instance, learning a representation via risk minimizer invariance (an intuitive formalization of robustness) is impossible in general. However, we'll also find that there are (common) cases where there is a close connection. These results highlight the fundamental role of the underlying causal structure of the data.

2 Related Work

Causal invariance in domain generalization Several works (Peters et al. (2016); Rojas-Carulla et al. (2018); Arjovsky et al. (2019); Lu et al. (2021)) specify domain shifts with causal models. In these frameworks, the invariant predictor learns $P(Y \mid \mathrm{pa}(Y))$. However, this rules out, e.g., cases where Y causes X. CISA allows learning with more general structures, which we'll see is helpful in understanding the representation learning methods.

Domain generalization methods There are many methods for domain generalization; we give a categorization in the introduction. There have been a number of empirically-oriented surveys testing domain generalization methods in natural settings Koh et al. (2021); Wiles et al. (2021); Gulrajani & Lopez-Paz (2020). These find that no method consistently beats ERM, but many methods work well in at least some situations. Our aim here is to give theoretical insight into when each might work.

Robust methods The CISA framework is reasonable for many real-world problems, but certainly not all. There are other notions of domain shifts and differently motivated methods that do not fit under this framework. For example, many works (e.g., Sagawa et al., 2019; Liu et al., 2021; Ben-David et al., 2022) assume the testing domains are not too different from the training domains (e.g., test samples are drawn from the mixture of training distributions).These methods are complementary to the invariant representation learning approaches we study here; e.g., they would all fail in the two-bit-environments experiment in Appendix A.

3 Causal Invariance with Spurious Associations

Setup In domain generalization problems, we have training and test data from several related domains. The goal is to learn a predictor using data from training domains, and apply it to unseen test domains. The data comes in the form of (X, Y, E) where $X \in \mathcal{X}$ is the input, $Y \in \mathcal{Y}$ the label and $E \in \mathcal{E}$ is the domain index. We abstract each domain as a probability distribution P_e, where $X_i, Y_i \mid E_i = e \overset{\text{iid}}{\sim} P_e$. At training time, we have access to data from a finite set of domains $\mathcal{E}_{\text{train}} \subset \mathcal{E}$.

Our goal is to learn a representation ϕ so that $\phi(X)$ is in some sense invariant across the domains. Many methods (e.g., Arjovsky et al., 2019; Ganin et al., 2016; Goel et al., 2020) frame their learning procedures this way, though there are many distinct formal notions of "invariance". The first task is to establish a canonical notion of invariance.

Causal Invariance with Spurious Associations All methods for handling domain shifts must specify the structure that is preserved across domains, and the ways in which they are allowed to vary.[2] Here, we'll rely on a particular variant of the assumption that causal relationships are held fixed, while non-causal relationships may vary. To formalize this, we assume that X is caused by some (unobserved) factors of variation, and give a notion of what it means for these factors to be spuriously associated with Y:

Definition 1. We say a (latent) cause of X is a *spurious factor of variation* if it is not a cause of Y and there is some (latent) confounder that affects both it and Y. Call the set of all such causes the *spurious factors of variations*.

[2]Without such an assumption, the test domain could be chosen adversarially to have support only on points where the training-domain predictor makes mistakes.

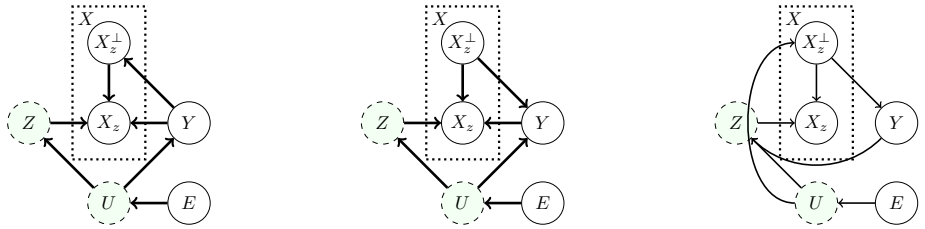


Figure 1: Examples of CISA-compatible causal structures. In each case, all causal relationships (solid arrows) are invariant across domains. Only the distribution of the unobserved confounder U can vary, changing the induced association between Y and the (unobserved) spurious factors Z.

Figure 1 shows examples of causal structures for prediction problems involving spurious factors of variation (Z). X is divided into two parts: $X_z^\perp$ denotes the part of X not affected by the spurious factors Z, and X_z the remaining part affected by Z. We do not assume the division of X is known a priori (indeed, Z is not even known!). Also note that $X_z^\perp$ may affect X_z, so these parts need not be independent. Formally, we use the notion of counterfactual invariance from Veitch et al. (2021).

Definition 2. Let ϕ be a function from $\mathcal{X}$ to $\mathcal{H}$ and Z the spurious factors. We say ϕ is *counterfactually invariant to spurious factors* (abbreviated CF-invariant), if $\phi(X(z)) = \phi(X(z'))$ $a.e.$, $\forall z, z' \in \mathcal{Z}$. Here, $X(z)$ is potential outcomes notation, denoting the X we would have seen had Z been set to z.

Then, we can formalize the part of X that's not affected by Z:

Definition 3. $X_z^\perp$ is a X-measurable variable such that, for all measurable functions f, f is CF-invariant iff $f(X)$ is $X_z^\perp$-measurable.[3]

Now, we characterize how the relationship between Z and Y changes across domains. We introduce U as the collection of all common causes of Z and Y and allow the distribution of U to vary across domains. Importantly, we assume that U does not confound the relationship between $X_z^\perp$ and Y, as otherwise no part of X would have a stable relationship with Y. Now we can put together all the ingredients and formally specify how the domains are related:

Definition 4. We say a set of domains $\mathcal{E}$ are *Causally Invariant with Spurious Associations* (CISA), if there are (unknown) spurious factors of variation Z, and unknown confounder U (that does not confound the relationship between $X_z^\perp$ and Y) so that $P_e(X, Y) = \int P_0(X, Y, Z|U)P_e(U)dZdU, \forall e \in \mathcal{E}$, where P_0 is a fixed distribution determined by the underlying dynamics of the system and $P_e(U)$ is a domain-specific distribution of the unobserved confounder.

Assumptions in domain shifts should be about the real-world processes that govern the domains from which the data is collected. This allows us to assess whether an assumption is reasonable for a given real-world situation. CISA is such an assumption; the spurious factors Z and confounder U are unknown and unobserved, but are assumed to correspond to real-world properties. This assumption about the real-world structure implies an assumption about the distributions $P_e(X, Y)$ in each domain. However, in general, the variables Z and U need not be identifiable from (i.e., uniquely determined by) the observational data. This point is only conceptually important, and does not affect the technical development in the following.

CF-invariant representation learning We now return to the question of what the right notion of domain-invariant representation is. In the case of CISA domains, there is a canonical notion for the part of X that has a domain-invariant relationship with Y. Namely, $X_z^\perp$, the part of X that is not affected by the spurious factors of variation. Accordingly, the goal of domain-invariant representation learning in this context is to find a representation ϕ such that $\phi(X)$ is counterfactually-invariant to Z (or, equivalently, $\phi(X)$ is $X_z^\perp$ measurable). Of course, we could satisfy this condition by simply throwing away all of the information in X (e.g., $\phi(X) = 0$ everywhere). So, we further look for the counterfactually-invariant representation that preserves the most predictive power for the label Y. Let $\Phi_{\text{cf-inv}}(\mathcal{E})$ denote the set of CF-invariant representations for CISA domains $\mathcal{E}$. Then, we take our domain-invariant representation learning objective to be:

$$\min_{\phi:\mathcal{X}\to\mathcal{H}, w:\mathcal{H}\to\mathcal{Y}} E_{P_{\mathcal{E}_{\text{train}}}}[L(Y, (w \circ \phi)(X))]$$
$$\text{subject to } \phi \in \Phi_{\text{cf-inv}}(\mathcal{E})$$

[3]Such a variable exists under weak conditions; e.g., Z discrete Veitch et al. (2021).

Here, the predictor w and loss function L capture the sense in which $\phi(X)$ should be predictive of Y.

The challenge here is that the spurious factors of variation are unknown and unobserved, so identifying the set of counterfactually-invariant representations is difficult. Now we can turn to understanding various approaches to domain adaptation methods as achievable relaxations of this ideal objective.

4 Causal View on Domain Invariant Representation Learning

4.1 Data Augmentation

Data augmentation is a standard technique in machine learning pipelines, and has been shown to (sometimes) help when faced with domain shifts Wiles et al. (2021). Our goal now is to understand when and why data augmentation might enable domain-invariant representation learning. The basic technique first applies pre-determined "label-preserving" transformations t to original features X to generate artificial data $t(X)$. There are two ways this transformed data can be used. The first option is to simply add the transformed data as extra data to a standard learning procedure. Alternatively, we might pass in pairs $(X_i, t(X_i))$ to our learning procedure, and directly enforce some condition that $\phi(X) \approx \phi(t(X))$ Garg et al. (2019). In both case, the natural questions are: What transformations are "label-preserving"? And, when do these techniques help with domain generalization?

We first formalize a notion of "label-preserving" for CISA domains. The key idea is that we can think of transformation $t(X)$ of X as being equivalent to changing some cause of X and then propagating this change through. For example, suppose a particular transformation t rotates the input images by 30 degrees, and W is the factor of variation corresponding to the angle away from vertical. Then, we can understand the action of t as $t(X(w)) = X(w + 30)$, where we again use the potential outcomes notation for counterfactuals. With this idea in hand, we see that a transformation is *label-preserving* in CISA domains if it is equivalent to a change that affects only spurious factors of variation. That is, label-preserving transformations cannot affect $X_z^{\perp}$. Otherwise, the transformation may change the invariant relationship with Y; replacing the background of cows with "sand" with "grass" doesn't change animal type; but distorting the part corresponding to the "cow" object may.

Definition 5. We say a data transformation $t : \mathcal{X} \to \mathcal{X}$ is *label-preserving* for CISA domains $\mathcal{E}$ if, for each $X(z)$ there is z' so that $t(X(z)) = X(z'), a.e.$.

Label preserving transformations leave the CISA invariant relationships (between $X_z^{\perp}$ and Y) alone, but can change the relationship between Y and the spurious factors of variation Z. Intuitively, if we have a 'large enough' set of such transformations, they can destroy the relationship between Y and Z that exists in the training data. So, we might expect that training with such data augmentation will automatically yield an optimal counterfactually-invariant predictor.

This is nearly correct, with the caveat that things can go wrong (for the naive training approach) if there is a part of X causally related to both Z and Y. That is, if there is a part of X that relies on the interaction between Z and Y. We follow Veitch et al. (2021) in formalizing how to rule out this case:

Definition 6. The spurious factors of variation Z are *purely spurious* if $Y \perp X | X_z^{\perp}, Z$

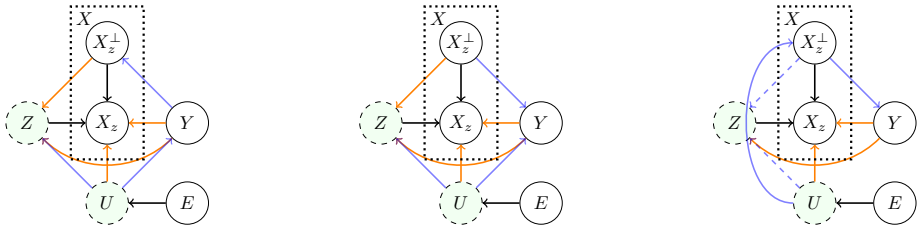
That is, without the unstable correlation (removed by conditioning on Z), $X_z^{\perp}$ is sufficient for Y.We can now state the main result connecting data augmentation and domain-invariance:

Theorem 7. *For a CISA domain, if the set of transformations $\mathcal{T}$ satisfies label-preserving and enumerates all potential outcomes of Z, then*

1. *If the model is trained to minimize risk on augmented data, and Z is purely spurious, or*
2. *If the model is trained to minimize risk on original data, with hard consistency regularization (i.e. enforcing $\phi(X) = \phi(t(X)), \forall t \in \mathcal{T}$)*

Then we recover the CF-invariant predictor that minimizes risk on original data.

Thus for CISA domains, ideal data augmentation (i.e. enumerating label-preserving transformations) will exactly learn CF-invariant representations. Moreover, this holds irrespective of what the true underlying causal structure is. Accordingly, such data augmentation would be the gold standard for domain-invariant representation learning.



(a) anti-causal **(b)** confounded-outcome **(c)** confounded-descendant

Figure 2: Every CISA compatible set of domains obeys exactly one of these causal structures. The black arrows are included in all graphs. The blue arrows are specific to different causal structures. The orange arrows are optional. At least one of the two dashed blue arrows in Figure 2c must exist.

However, in practice, it can be hard to satisfy the idealized conditions. Applying predefined transformations without thinking about the specific applications can lead to violations of the label preserving condition. For example in a bird classification task, changing color may really change the bird species (this is called manifold intrusion in Guo et al. (2019)). Further, heuristic transformations often cannot enumerate all potential outcomes of $X(z)$. Indeed, deep models can sometimes memorize predetermined transformations and fail to generalize with previously unseen transformations (Vasiljevic et al. (2016), Geirhos et al. (2018)).

Considering the limitations of using handcrafted transformations, a natural idea is to replace them with transformations learned from data. However, in practice, Z is unknown and we only observe the data domains E. Then, learning transformations must rely either on detailed structural knowledge of the problem (Robey et al., 2021, e.g.,), or on some distributional relationship between E, Y and X and $t(X)$ (e.g., Goel et al., 2020). Since $t(X)$ is only used for the representation learning, this is equivalent to learning based on some distributional criteria involving E, Y, and $\phi(X)$—the subject of the next section.

4.2 Distributionally Invariant Learning

Many domain-invariant representation learning methods work by enforcing some form of distributional invariance Section 1. There are three possible notions of distributional invariance, each of which appears in the literature. Early work tries to learn a representation ϕ so that $\phi(X) \perp\!\!\!\perp E$ while $\phi(X)$ is still predictive of Y Muandet et al. (2013); Ganin et al. (2016). The intuition is that predictions must be based on features that cannot discriminate between training and test domains. Later work recognized that that $P(Y|X)$ may be unstable across domains, and instead aims at $\phi(X) \perp\!\!\!\perp E|Y$ Li et al. (2018b); Long et al. (2018). Finally, some causally-motivated approaches have considered learning representations so that $Y \perp\!\!\!\perp E|\phi(X)$ Peters et al. (2016); Wald et al. (2021).

When the representation $\phi(X)$ is predictive of Y and Y is dependent with E, these three independence statements—$\phi(X) \perp E, \phi(X) \perp E|Y, Y \perp E|\phi(X)$—are the only possible distributional invariance relationships. It is also well-known that, if Y and E are not independent, then these distributional invariances are mutually exclusive. The question now is: when, if ever, are each of these distributional invariances the right approach for domain-invariant representation learning?

CISA compatible causal structures To answer this, the first step is to categorize the CISA domains into three types, corresponding to the underlying causal structure of the problem.

Theorem 8. *If a set of domains $\mathcal{E}$ satisfy CISA, then the underlying causal structure must correspond to exactly one of three types of causal graph: anti-causal, confounded-outcome, or confounded-descendant. These graphs are shown in Figure 2.*

The idea here is that for a set of CISA domains, the causal structure is the same in each domain. There are only three kinds of causal structure, distinguished by whether Y causes $X_z^\perp$ or $X_z^\perp$ causes Y (anti-causal vs confounded-outcome/descendant), and by whether the confounding affects Y directly or just affects some causal descendant of Y (confounded-outcome vs confounded descendant).

Anti-causal Image classification can be naturally viewed as an anti-causal problem. Various factors of variations such as lighting, background, angles, etc, and the object class Y, generate the image

X. Some of the factors of variation Z are confounded with Y—e.g., background and Y may be associated due to evolutionary pressures. The Cow/Camel on Grass/Sand example fits here.[4]

Confounded outcome The goal is to predict the helpfulness of a review. Each review receives a number of "helpful" votes Y, produced by site users. We use the review's text content X as covariates. The data is collected for different types of products E. The model's performance drops significantly when deployed in new product type. We think that the general sentiment Z of the review, and the helpfulness has unstable relationship across E: e.g. for books, customers write very positive reviews which are often voted favorably; for electronics this relationship is reversed.

Confounded descendant Consider predicting unemployment rate Y from a variety of economic factors X. It's not clear which factors cause Y directly, and which are descendants of Y. The relationships among X, Y change under certain big events, say financial crisis or pandemics. We denote those events as domains E, and take $U = E$. U might affect Y's descendants jointly with Y (through intermediate variable Z), but Y is not changed. Or U might affect $\mathrm{pa}(Y)$ directly, which changes Y. Notice that in this case, the CF-invariant representation is also the representation that uses only $\mathrm{pa}(Y)$. Thus, for this causal structure, the counterfactually invariant notion matches the traditional causally-invariant representation learning desiderata Peters et al. (2016); Arjovsky et al. (2019).

Distributionally invariant learning We now return to the question of when distributional invariance learns counterfactually-invariant representations. It turns out that the answer relies critically on the true causal structure of the problem:

Theorem 9. *If ϕ is a counterfactually-invariant representation,*

1. *if the underlying causal graph is anti-causal, $\phi(X) \perp\!\!\!\perp E|Y$;*
2. *if the underlying causal graph is confounded-outcome, $\phi(X) \perp\!\!\!\perp E$;*
3. *if the underlying causal graph is confounded-descendant, $Y \perp\!\!\!\perp E|\phi(X)$.*

Remark 10. This theorem looks similar to (Veitch et al., 2021, Thm. 3.2). This is deceptive; here we observe the environment E, whereas they assumed observations of the spurious factors Z.

In words: each of the distributional invariances arises as a particular implication of CF-invariance. This suggests a learning strategy that relaxes CF-invariance. Namely, for a given problem, determine the distributional invariance implied by CF-invariance and then enforce that distributional invariance. That is, we learn a representation according to:

$$\min_{\phi \in \Phi_{\mathrm{DI}}(\mathcal{E}_{\text{train}}), w} E_{P_{\mathcal{E}_{\text{train}}}}[L(Y, (w \circ \phi)(X))],$$

where $\Phi_{\mathrm{DI}}(\mathcal{E})$ is the set of representations matching the causal structure of the domains.

Notice that the right distributional invariance is necessary but not sufficient for CF-invariance. There are two reasons: first, the causal condition of CF-invariance is not implied by its distributional signature in Theorem 9. Second, the distributional invariance can only be measured on a limited number of training domains, which may provide a limited constraints. In this sense, distributional invariance is a relaxation of CF-invariance; i.e., $\Phi_{\text{cf-inv}}(\mathcal{E}) \subsetneq \Phi_{\mathrm{DI}}(\mathcal{E}_{\text{train}})$.

Also notice that in general enforcing the wrong distributional invariance will directly contradict CF-invariance, and *increase* dependency on the spurious factors of variation. Distributional invariance is only a relaxation when we actually get the underlying causal structure correct!

4.3 Invariant Risk Minimization

The Invariant Risk Minimization (IRM) paradigm Arjovsky et al. (2019) aims to find representations that admit a single predictor that has the optimal risk across all domains. That is, the set of IRM representation is

$$\Phi_{\mathrm{IRM}}(\mathcal{E}) := \{\phi : \exists\, w \text{ st } w \in \operatorname*{argmin}_{\bar{w}} E_{P_e}[L(Y, (\bar{w} \circ \phi)(X))]\ \forall e \in \mathcal{E}\} \tag{4.1}$$

[4]There is some controversy here since the label Y is often due to human annotators Lu et al. (2021).

Here, the question is: when, if ever, does the IRM procedure yield a CF-invariant predictor?

Again the answer will turn out to depend on the underlying causal structure. IRM is justified in the case where X includes both causes and descendants of Y, and the invariant predictor should use only information in the parents of Y. As explained above, this coincides with CF-invariance—and with distributional invariance—in the confounded-descendant case. Indeed, we can view IRM as a relaxation of $Y \perp\!\!\!\perp E|\phi(X)$. Instead of asking the distribution $P^e(Y|\phi(X) = h)$ to be invariant across domains for every $h \in \mathcal{H}$, we only require the risk minimizer under $P^e(Y|\phi(X) = h)$ to be invariant. So, a partial answer is that IRM is a relaxation of distributional invariance (and CF-invariance) when the underlying causal structure is confounded descendant.

However, the situation is harsher in the case of other causal structures. For anti-causal and confounded-outcome problems, the typical case is $\Phi_{\text{IRM}}(\mathcal{E}) = \emptyset$. For the anti-causal case, this is a consequence of *prior shift* Zhang et al. (2013). We can change the optimal predictor just by changing the prior distribution on Y without affecting the causal structure. Thus, there is no $X_z^{\perp}$-measurable predictor with invariant risk. The confounded-outcome case is even harder. In general, the generating function for Y could involve interaction between U and X. Then $P(Y|X_z^{\perp})$ can be quite different because of the shifting U. Thus the risk minimizers may be very different across domains.

Fortunately, for anti-causal problems there is a natural generalization of IRM that partially enforces the distributional invariance. Similarly, we can relax the invariance of $P^e(\phi(X)|Y = y)$ to the invariance of a re-weighted risk. We define generalized-IRM (g-IRM) as follows:

Definition 11. For anti-causal domains, we define the set of representations satisfying g-IRM as:

$$\Phi_{\text{g-IRM}}(\mathcal{E}) := \{\phi : \exists\, w \text{ st } w \in \operatorname*{argmin}_{\bar{w}} E_{P_e}[\frac{P_0(Y)}{P_e(Y)} L(Y, (\bar{w} \circ \phi)(X))], \forall e \in \mathcal{E}\}$$

where $P_0(.)$ is any baseline distribution for Y.

Then we can show the (generalized) IRM is a relaxation of the corresponding distributional invariance:

Theorem 12. *Let $\mathcal{E}$ satisfy CISA, then*

1. *if $\mathcal{E}$ is confounded-descendant, then $\Phi_{DI}(\mathcal{E}) \subset \Phi_{IRM}(\mathcal{E})$*
2. *if $\mathcal{E}$ is anti-causal, then $\Phi_{DI}(\mathcal{E}) \subset \Phi_{g\text{-}IRM}(\mathcal{E})$*

Although this simple generalization works for the anti-causal case, there is no such easy fix for the confounded-outcome case. For confounded outcome, the idea of having the same risk minimizer across domains does not make sense without further assumptions. The general case has $Y \leftarrow f(X, U, \eta)$ where η is noise independent of X, U. For example when $U = E$ we can write $Y \leftarrow f_E(X, \eta)$ so there could be arbitrarily different relationships between X and Y in every domain. It may be possible to circumvent this by making structural assumptions on the form of f; e.g., there is an invariant risk minimizer in the case where the effect of U and X is additive Veitch et al. (2021).

4.4 Relationship of Methods

Data augmentation training is the gold standard for CF-invariant representation learning if it's possible to enumerate all label-preserving transformations (Theorem 7). This is impossible in general as it requires direct manipulation of the spurious factors Z. Still, using augmentation with label-preserving transformations (but not exhaustively) enforces a relaxation of CF-invariance.

Distributional invariance relaxes CF-invariance if and only if chosen to match the underlying causal structures (Theorem 9). This can be a good option when full augmentation is not possible.

(generalized) IRM further relaxes distributional invariance for anti-causal and confounded-descendant problem, when it's chosen to match the causal structure of the problem (Theorem 12). It weakens the full independence criteria to use just the implication for a single natural test statistic: the loss of the model. This allows for more efficient algorithms Arjovsky et al. (2019).

5 Insights for Robust Prediction

Often, learning domain-invariant representations is an intermediate step towards learning robust predictors. Here, robust means that the predictor should have good performance when deployed in a

previously unseen domain. We now discuss the implications of our domain-invariant representation learning results for robust prediction.

whether robust prediction is even possible depends on the underlying causal structure For confounded-descendant problems, robust prediction is straightforward. Enforcing the correct distributional invariance on representation function ϕ leads to invariant risk in test domains for every value of $\phi(X)$. If we enforce the weaker IRM requirement, the optimal predictor (on top of ϕ) on training domains is optimal on test data. For anti-causal problems, there is no invariant predictor in general because $P_e(Y)$ can change across domains. However, there are simple adaptation methods to jointly estimate $P_e(Y)$ and adjust prediction during deployment Saerens et al. (2002), which can achieve near-optimal performance efficiently Lemberger & Panico (2020). Moreover, if the prior distribution doesn't shift very much between training and deployment, then the training-domain optimal predictor trained on the invariant representation will be nearly optimal in the test case. This kind of limited label shift seems common in practice. Accordingly, for both the confounded-descendant and (many) anti-causal cases, we can achieve robust prediction simply by naively training a predictor on top of the domain-invariant representation. For the confounded-outcome case, there is no notion of robust predictor without making some further structural assumptions.

no method works for all domain generalization problems Heuristic data augmentation enforces CF-invariance directly regardless of causal structures, but requires truly label-preserving transformations and only solves the invariant representation problem if the transformation set affects all spurious factors. Distributionally-invariant methods can work, but each approach is only valid if it matches the underlying causal structure of the problem. Enforcing the wrong distributional invariance can actually harm performance Veitch et al. (2021). Similarly, (generalized) invariant risk minimizer can work for some types of problems, but only when it matches the causal structure. This is consistent with the findings from various benchmark studies that there is no single method that does well in all tasks Wiles et al. (2021); Koh et al. (2021); Gulrajani & Lopez-Paz (2020).

data augmentation helps in most cases Label-preserving data augmentation won't hurt domain generalization and can often help. This is true no matter the underlying causal structure of the problem. This matches empirical benchmarks where data augmentations usually help domain generalization performance, sometimes dramatically Wiles et al. (2021); Koh et al. (2021); Gulrajani & Lopez-Paz (2020). For example, Wiles et al. (2021) finds that simple augmentations used in Krizhevsky et al. (2012) generally improves performance when "augmentations approximate the true underlying generative model".

pick a method matching the true causal structure Many papers apply distributional invariance approaches with no regard to the underlying causal structure of the problem. In particular, many tasks in benchmarks have the anti-causal structures, but the methods evaluated do not include those enforcing $\phi(X) \perp E|Y$ Koh et al. (2021). Tachet des Combes et al. (2020) find that methods enforcing $\phi(X) \perp E|Y$ consistently improve over methods that enforce $\phi(X) \perp E$—retrospectively, this is because they benchmark on anti-causal problems. Wiles et al. (2021) finds that learned data augmentation (Goel et al. (2020)) consistently improves performance in deployment—this method can be viewed as enforcing $\phi(X) \perp E|Y$; again, the benchmarks mostly have anti-causal structure.

6 Conclusion

We have introduced the CISA notion of domain shift and used it to study invariant representation learning methods. The key property of CISA is that it has a cannonical notion of "domain-invariant" representation: CF-invariance. We have seen that of data augmentation, distributional invariance learning, and risk invariant learning can be (sometimes) understood as relaxations of CF-invariance. The validity of the methods—as well as the connection between domain-invariance and domain-robust learning—depend fundamentally on the underlying causal structure of the data. As a further illustration of this point, in appendix A we find that two-bit-envs experiment — a simplification of colored MNIST, the canonical IRM demonstration example—actually has a causal structure mismatched to IRM, and show that simply changing the label distribution across environments causes IRM to fail. Further, we show this failure is fixed by using g-IRM, which is compatible with the true causal structure.