A Unified Causal View of Domain Invariant Representation Learning

Zihao Wang [1] **Victor Veitch** [1 2]

Abstract

Machine learning methods can be unreliable when deployed in domains that differ from the domains on which they were trained. To address this, we may wish to learn representations of data that are domain-invariant in the sense that we preserve data structure that is stable across domains, but throw out spuriously-varying parts. There are many representation-learning approaches of this type, including methods based on data augmentation, distributional invariances, and risk invariance. Unfortunately, when faced with any particular real-world domain shift, it is unclear which, if any, of these methods might be expected to work. The purpose of this paper is to show how the different methods relate to each other, and clarify the real-world circumstances under which each is expected to succeed. The key tool is a new notion of domain shift relying on the idea that causal relationships are invariant, but non-causal relationships (e.g., due to confounding) may vary. Considering this type of domain shift, a natural goal is to learn representations that are "Counterfactually Invariant". We find the popular domain-invariant representation learning methods enforce invariance that corresponds to the Counterfactual Invariance under different types of causal structures. Therefore, we should pick the method that matches the underlying causal structure.

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 training domain(s), and the deployed domain(s). Many methods have been developed to address it. One popular class of approach—which we'll focus on in this paper—is to try to learn a representation of the data which is in some sense invariant across domains. The intuition is that domain-invariant representations should capture the part of the data structure that is the "same" in all domains, so that, e.g., a predictor built on top of a domain-invariant representation will have good performance even when deployed in a previously unseen domain.

A principle challenge here is that it's unclear a priori how domain-invariance ought to be formalized and enforced. There exists many popular approaches for learning domain-invariant representations. When applied to broad ranges of real-world domain-shift benchmarks, there is no single dominant approach, and an it's even hard to beat the baseline empirical risk minimization. The aim of this paper is to clarify what assumptions on domain-shift structure justify (or contradict) common approaches to domain-invariant representation learning. We'll be particularly interested in the following three approaches:

Data augmentation We want our representations to b invariant to small perturbations: E.g., if $t(X)$ is a small rotation of an image X, then $\phi(X) = \phi(t(X))$. For example, [KSH12; Hen+19; Cub+19; Xie+20; WZ19; Pas+19; HG17; SHB15; Kob18; Nie+20].

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: $\mathrm{P}^e(\phi(X)) = \mathrm{P}^{e'}(\phi(X))$ [MBS13; Gan+16; Alb+20; Li+18a; SFS17; SS16; MH20];
conditional invariance: When Y is a label of interest, $\mathrm{P}^e(\phi(X) \mid Y) = \mathrm{P}^{e'}(\phi(X) \mid Y)$ [Li+18b; Lon+18; Com+20; Goe+20]
sufficiency: $\mathrm{P}^e(Y \mid \phi(X)) = \mathrm{P}^{e'}(Y \mid \phi(X))$ [PBM16; RC+18; Wal+21].

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 [Arj+19; Lu+21; Ahu+20; Kru+21; BCL21].

To understand the relationship of these notions of invariance, it is necessary to specify how different domains are

[1]Department of Statistics, the University of Chicago [2]Google Research. Correspondence to: Victor Veitch <victorveitch@gmail.com>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.

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. Specializing to supervised learning with label Y and covariates X, the approach we'll take proceeds 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: Y and Z are dependent due to an unobserved common cause U. The distribution of U may change across environments, thereby shifting the relationship between Y and Z. 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. Although Z (the background) is known in this example, it's important to note that we do not assume spurious factors Z are known a priori in general. CISA merely asserts the existence of such factors, but requires no knowledge of what they might be.

CISA is an assumption that is reasonable for many real-world situations and gives a canonical notion of domain-invariant representation. Namely, $\phi(X)$ should be the part of X that does not depend on the (unknown) spurious factors of variation Z. In the Camel/Cow example, $\phi(X)$ should exclude information in the image that changes if sand is added or removed from the background. We say a representation with this property is *counterfactually invariant to spurious factors* (CF-invariant for short). Under the CISA assumption, CF-invariance is the 'right' notion of domain-invariance—it exactly captures the part of X that has a stable relationship with Y across all domains. Accordingly, in the CISA setting, an ideal domain-invariant representation learning method should produce a CF-invariant representation.

We now return to the motivating questions: how are different approaches to domain-invariant representation learning related, and when might each work? The main idea of the paper is to answer these questions by determining the conditions under which each approach learns a CF-invariant representation. Informally, the technical contributions of the paper are:

1. Formalization of CISA and Counterfactually Invariant Representation Learning.
2. We first show that data augmentation techniques lead to counterfactually invariant representations for a suitable class of augmentations. The required condition is that the augmentations should be "label preserving" in the sense of (being equivalent to) affecting only spurious factors, and exhaustive in the sense of affecting all spurious factors.
3. Next, we show that distributional-invariance techniques can enforce (a relaxed form of) counterfactual invariance. This holds if and only if the particular distributional invariance is chosen to match the true underlying causal structure of the problem.
4. Finally, we show that risk minimizer invariance is impossible under CISA except for certain special causal structures. We then give causal-structure dependent generalizations of risk-invariant representation learning and show that these enforce (a relaxed version of) the causal-structure induced distributional invariance.

2. Related Work

Causal invariance in domain generalization Several works ([PBM16; RC+18; Arj+19; Lu+21; CB20]) specify domain shifts with causal models. In most frameworks, the invariant predictor learns $P(Y \mid \mathrm{pa}(Y))$. This is quite limiting. For example, it's impossible to have meaningful stable predictors in problems where Y causes X [Sch+12]. CISA allows learning with more general structures, which we'll see is helpful in understanding the representation learning methods. [CB20] allows more general causal structures, but it relies on limited parametric assumptions.

Domain generalization methods There are many methods for domain generalization; we give a categorization in the introduction. Our aim in this paper is to give theoretical insight into when each might work. Complementary to this, there have been a number of empirically-oriented surveys testing domain generalization methods in natural settings [Koh+21; Wil+21; GLP20; Hen+21]. These evaluations find that no method consistently beats ERM, but many methods work well in at least some situations. The question we're concerned with here is which situations are well-matched to which methods.

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., Sag+19; Liu+21; BDOR22] assume the testing domains are not too different from the training domains (e.g., test samples are drawn from the mixture

of training distributions); under this type of domain shift, "robustness" (e.g. good worst-case loss over training domains) is a more desired property than "invariance" as discarding spurious features would be too conservative for prediction. However, these methods would all fail in the simple Colored MNIST experiment in Appendix D.

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 $\mathcal{E}_{\text{train}} \subset \mathcal{E}$.

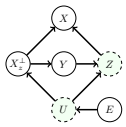


Figure 1: Examples of a causal structures compatible with the CISA assumption. Causal relationships (solid arrows) are invariant across domains. Only the distribution of the unobserved confounder U can vary.

Causal Invariance with Spurious Associations First we specify the structure that is preserved across domains, and the ways in which they are allowed to vary.[1] 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*.

Figure 1 shows examples of causal structures for prediction problems involving spurious factors of variation (Z). We will need notation for the part of X that is not affected by the spurious factors Z. Formally, we use the notion of counterfactual invariance from [Vei+21].

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 and denotes the value of X we would have seen had Z been set to 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.[2] [3]

Now we can introduce the confounder U 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.

CF-invariant representation learning The goal is to learn a representation that's both predictive of Y, and has the "invariance" property. For CISA domains, the canonical notion of invariance is CF-invariance. Let $\Phi_{\text{cf-inv}}(\mathcal{E})$ denote the set of CF-invariant representations for CISA domains $\mathcal{E}$. Our domain-invariant representation learning objective is:

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

The challenge here is that the spurious factors of variation are unknown and unobserved, so identifying $\Phi_{\text{cf-inv}}(\mathcal{E})$ 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

Our goal 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 $\phi(X) \approx \phi(t(X))$ [Gar+19].

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

[2] Such a variable exists under weak conditions; e.g., Z discrete [Vei+21].

[3] It will be convenient to think of $X_z^\perp$ as a cause of X—this is just shorthand for the part of X that is caused by the non-spurious factors of variation, and avoids introducing another latent variable to explicitly label the non-spurious factors.

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 corresponding to the angle away from vertical. Then, we can understand the action of t as $t(X(w)) = X(w + 30)$. 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. 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. This is nearly correct, with the caveat that things can go wrong (for the naive training approach) if there is a part of X that is causally related to both Z and Y. We follow [Vei+21] in formalizing how to rule out this case:

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

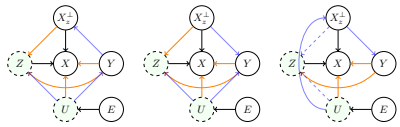
That is, conditioning on the spurious factors, $X_z^\perp$ is sufficient for Y.We can now state the main result connecting data augmentation and domain-invariance:

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, regularized to satisfy $\phi(X) = \phi(t(X)), \forall t \in \mathcal{T}$*

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

Thus for CISA domains, the ideal data augmentation will exactly learn CF-invariant representations, irespective of the true underlying causal structure. Accordingly, such data augmentation would be the gold standard for domain-invariant representation learning. However, in practice the set of pre-determined transformations often is not label-preserving ([GMZ19]) nor exhaustive ([VCS16], [Gei+18]).



(a) anti-causal **(b)** conf-out **(c)** conf-desc

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.

Considering their limitations, 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 [RPH21, e.g.,], or on some distributional relationship between E, Y and X and $t(X)$ [e.g., Goe+20]. 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 approaches to domain-invariant representation learning work by enforcing one of the three forms of distributional invariance as discussed in Section 1. The question now is: when, if ever, are each of these distributional invariances the right approach for domain-invariant representation learning? First, observe that CISA domains can be categorized 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.*

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. *Let ϕ be a CF-invariant representation, if the underlying causal graph is*

1. *anti-causal, then $\phi(X) \perp\!\!\!\perp E | Y$*
2. *confounded-outcome, then $\phi(X) \perp\!\!\!\perp E$*
3. *confounded-descendant, then $Y \perp\!\!\!\perp E | \phi(X)$.* [4]

In words: each of the distributional invariances arises as a particular implication of CF-invariance. Let $\Phi_{\text{DI}}(\mathcal{E})$ be the set of representations matching the causal structure of $\mathcal{E}$.

[4] This theorem looks similar to [Vei+21, Thm. 3.2]. This is deceptive; here we observe the environment E, whereas they assumed observations of the spurious factors Z.

Then $\Phi_{\text{cf-inv}}(\mathcal{E}) \subsetneq \Phi_{\text{DI}}(\mathcal{E}_{\text{train}})$. Therefore enforcing the right distributional invariance partially enforces CF-invariance.

4.3. Invariant Risk Minimization

The Invariant Risk Minimization (IRM) paradigm [Arj+19] restricts the representation to this set:

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

That is, they aims to find representations that elicit a single predictor that has the optimal risk across all domains. 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. For confounded-descendant problems, IRM is a relaxation of $Y \perp\!\!\!\perp E|\phi(X)$. However, for anti-causal and confounded-outcome problems $\Phi_{\text{IRM}}(\mathcal{E}) = \emptyset$ in general. Consider the most simple case with no X (where the two types of graphs collapse into one). Since $P_e(Y)$ can be arbitrarily different, there is no invariant risk minimizer. Fortunately, for anti-causal problems there is a natural generalization of IRM that partially enforces the distributional invariance.

Definition 10. For anti-causal domains, we define the set of representations satisfying g-IRM as:

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

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 11. *Let $\mathcal{E}$ satisfy CISA, then if $\mathcal{E}$ is*

1. *confounded-descendant, then $\Phi_{DI}(\mathcal{E}) \subset \Phi_{IRM}(\mathcal{E})$*
2. *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 example, if $Y \leftarrow f(X, E, \eta)$ where η is noise independent of X, E. There could be disctinct relationships between X and Y in every domain. [5]

[5] 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 [Vei+21].

A. 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 is a relaxation of CF-invariance if it's chosen to match the underlying causal structures of the problem (Theorem 9). However, enforcing full distributional invariance is still hard ([SP20]).

(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 11). 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 [Arj+19].

B. Insights for Robust Prediction

Usually, 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 turn to the implications of our domain-invariant representation learning results for robust prediction.

First, for CISA domains, whether or not robust prediction is even possible will depend 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 [SLD02], which can achieve near-optimal performance efficiently [LP20]. 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.

With this in mind, we can now extract some general insights for robust prediction:

no method dominates in 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 [Vei+21]. 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 can dominate all tasks [Wil+21; Koh+21; GLP20].

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 [Wil+21; Koh+21; GLP20]. For example, [Wil+21] finds that simple augmentations used in [KSH12] 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+21]. [Com+20] 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. [Wil+21] finds that learned data augmentation ([Goe+20]) consistently improves performance in deployment—this method can be viewed as enforcing $\phi(X) \perp E|Y$ and, again, the benchmarks mostly have anti-causal structure.

C. Examples of the three types of CISA domains

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.[6]

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 [PBM16; Arj+19].

D. Experiments

One implication of our work is that we should use methods matching the true causal structure. To evaluate this claim, we conduct experiments on data modified from Color MNIST datasets from [Arj+19]. We can understand the underlying data generating process as anti-causal, but [Arj+19] treated it as confounded-descendant. The two views give the same IRM/gIRM objective when $P_e(Y)$ remains the same,so IRM method gives near-optimal results. When simply re-distributing samples to make $P_e(Y)$ change across domains, we predict gIRM would give superior results to IRM since it matches the underlying causal structure.

More specifically, this is how the (modified) Color MNIST is generated from the original MNIST dataset: first assign label 0 to digits $0-4$ and label 1 to other digits; then the labels are flipped with probability α; assign each label-digit (call it (Y, X_1)) pair to domains by the label so $P_e(Y = 1) = \pi_e$; finally the binary color (call it X_2) is assigned based on the (flipped) label, with flip rate β_e; the final image X is a colored version of the original handwritten

[6] In some cases there is some controversy about this example, since the label Y is often due to human annotators [Lu+21].

digit ($X = g(X_1, X_2)$ with composition function g). So the population parameter for each domain is (α, β_e, π_e). [Arj+19] uses two training domains $(0.25, 0.1, 0.5)$ and $(0.25, 0.2, 0.5)$, and the test domain is $(0.25, 0.9, 0.5)$. In this case, the ideal invariant predictor depends only on the digit X_1 and attains the optimal 75% accuracy. ERM or other robust methods would learn to use X_2 and generalize poorly. IRMv1 (the practical implementation of IRM) is shown to attain the near-optimal test accuracy. We simply change the priors $P_e(Y)$ so that training domains are $(0.25, 0.1, 0.9)$ and $(0.25, 0.1, 0.1)$ and the test domain is $(0.25, 0.9, 0.5)$. The optimal invariant predictor should remain the same and give the same 75% testing accuracy. fig. 3 shows that gIRMv1 penalty forces model to use only X_1 and gets close-to-optimal accuracy; IRMv1 penalty, on the other hand, allows the model use spurious features and get poor test accuracy.

E. Proofs

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

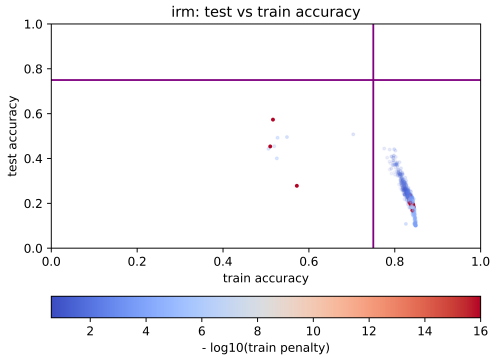
1. *If the model is trained to minimize risk on augmented data, and Z is purely spurious, or*
2. *If the model is trained to minimize risk on original data, regularized to satisfy $\phi(X) = \phi(t(X)), \forall t \in \mathcal{T}$*

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

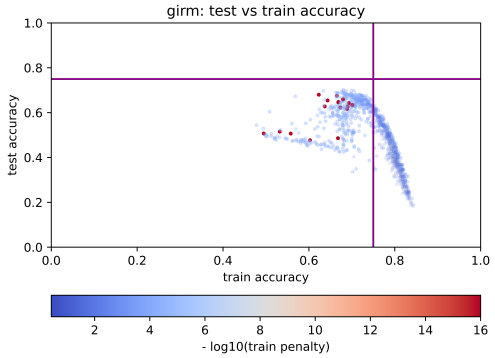
Proof. First, for the convenience of notation let's assume $X = X(z_0)$ $a.e.$ for some $z_0 \in \mathcal{Z}$. Then by the label-preserving $\mathcal{T}$, we have: for each $t \in \mathcal{T}$ we have $t(X)(= t(X(z_0))) = X(z)$ for some $z \in \mathcal{Z}$.

Consider consistency training. Let $\Phi_c(\mathcal{T})$ denote the set of representation functions satisfying consistency requirement under transformation set $\mathcal{T}$, i.e. $\Phi_c(\mathcal{T}) := \{\phi : \phi(X) = \phi(t(X))\ a.e.\ \forall t \in \mathcal{T}\}$. If $\phi \in \Phi_c(\mathcal{T})$, then for any $z, z' \in \mathcal{Z}$, can find $t \in \mathcal{T}$ such that $X(z') = t(X(z))$ since $\mathcal{T}$ enumerates all potential outcomes of Z; therefore $\phi(X(z')) = \phi(t(X(z))) = \phi(X(z))$ $a.e.$ by consistency requirement. Thus $\phi \in \Phi_{\text{cf-inv}}(\mathcal{E})$. On the other hand if $\phi \in \Phi_{\text{cf-inv}}(\mathcal{E})$, then for any $t \in \mathcal{T}$, we have $\phi(t(X)) = \phi(X(z)) = \phi(X)$ for some $z \in \mathcal{Z}$. Thus $\phi \in \Phi_c(\mathcal{T})$. Therefore $\Phi_c(\mathcal{T}) = \Phi_{\text{cf-inv}}(\mathcal{E})$. Therefore, training the model to minimize risk on original data, with hard consistency regularization is equivalent to CF-invariant representation learning, which recovers the optimal CF-invariant predictor on training distribution.

Consider ERM training on augmented data with purely-spurious Z. Let P denote the original distribution, and $\tilde{P}$ denote the distribution after the augmentation. Let T be



(a) IRMv1



(b) gIRMv1

Figure 3: IRMv1 vs gIRMv1 on Color MNIST. For each methods there are 1000 trained models, each represented by a point. The color represents the corresponding penalty. The optimal invariant predictor attains 75% accuracy on both training and test domains, marked by purple lines. The overall trend for gIRMv1 models is: before reaching 75% training accuracy, models improve predictions in both training and test domains; after passing the optimal training accuracy, the models rely on spurious feature to get better training performance, but this is discouraged by gIRMv1 penalty and won't be selected. IRMv1, however, cannot learn a good invariant model — the selected model will only give around 20% testing accuracy

the random variable for transformation operation. First, the generating process of the augmented data is: first sample $T \sim \tilde{P}_T(.)$; then sample $(X, Y)|T = t$ from the distribution of $(t(X), Y)$. Then we have:

$$\begin{aligned}\tilde{P}(X, Y) &= \int P(t(X), Y) d\tilde{P}_T(t) \\ &= \int P(X(z), Y) d\tilde{P}_Z(z) \\ &= \int P(X(z), Y(z)) d\tilde{P}_Z(z) \\ &= \int P(X, Y|do(z)) d\tilde{P}_Z(z)\end{aligned}$$

by the label-preserving of $\mathcal{T}$, and the fact that Y is not a descendant of Z.

Next, observe that $P(y|x, do(z)) = P(y|x_z^\perp)$. This is because: in original probability we have $Y \perp\!\!\!\perp X|X_z^\perp, Z$; the $do(z)$-operation removes the incoming edges of Z and set $Z = z$; as a result $P(y|x, do(z)) = P(y|x_z^\perp, do(z)) = P(y|x_z^\perp)$.

Put together:

$$\begin{aligned}\tilde{P}(X, Y) &= \int P(X, Y|do(z)) d\tilde{P}(z) \\ &= \int P(Y|X, do(z)) P(X|do(z)) d\tilde{P}(z) \\ &= \int P(Y|X_z^\perp) P(X|do(z)) d\tilde{P}(z) \\ &= P(Y|X_z^\perp) \int P(X|do(z)) d\tilde{P}(z) \\ &= P(Y|X_z^\perp) \tilde{P}(X)\end{aligned}$$

Therefore the objective is:

$$E_{\tilde{P}}[L(Y, f(X))] = E_{\tilde{P}(X)}[E_{P(Y|X_z^\perp)}(L(Y, f(X)))]$$

Then for any input x, the the optimal predictor output $f^*(x) = \text{argmin}_{a(x)} \int L(y, a(x)) dP(y|x_z^\perp)$. This is the same as directly restricting predictor to be CF-invariant. □

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.*

Proof. There are a finite number of possible causal DAGs relating the variables $U, Z, X, Y, X_z^\perp$. Moreover, for a DAG to be compatible with CISA it must satisfy some conditions that narrows down the set. In particular, $X_z^\perp$ causes X; Z affects X but not $X_z^\perp$ or Y; U should affect Z and Y but cannot confound $X_z^\perp$ and Y; E only affects U. In Figure 2, we show all possible DAGs on these variables that are compatible with CISA □

Theorem 9. *Let ϕ be a CF-invariant representation, if the underlying causal graph is*

1. *anti-causal, then $\phi(X) \perp\!\!\!\perp E|Y$*
2. *confounded-outcome, then $\phi(X) \perp\!\!\!\perp E$*
3. *confounded-descendant, then $Y \perp\!\!\!\perp E|\phi(X)$.* [7]

Proof. Reading d-separation from the corresponding DAGs, we have $X_z^\perp \perp E|Y$ for anti-causal problems; $X_z^\perp \perp E$ for confounded-outcome problems; $Y \perp E|X_z^\perp$ for confounded-descendant problems. Since ϕ is CF-invariant, that means $\phi(X)$ is $X_z^\perp$-measurable. Thus the claim follows. □

Theorem 11. *Let $\mathcal{E}$ satisfy CISA, then if $\mathcal{E}$ is*

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

Proof. Confounded-descendant case: let $\phi \in \Phi_{\text{DI}}(\mathcal{E})$, i.e. $Y \perp\!\!\!\perp E|\phi(X)$. To show the risk minimizer is the same, it suffices to show $P_e(Y|\phi(X))$ to be the same for all $e \in \mathcal{E}$. This is immediate from the distributional invariance.

Anti-causal case: if the representation $\phi \in \Phi_{\text{DI}}(\mathcal{E})$, i.e. $\phi(X) \perp\!\!\!\perp E|Y$,

$$\begin{aligned}&E_{P_e}[\frac{P_0(Y)}{P_e(Y)} L(Y, (\bar{w} \circ \phi)(X))] \\ &= E_{Y \sim P_e}[\frac{P_0(Y)}{P_e(Y)} [E_{\phi(X) \sim P_e(.|Y)}(L(Y, (\bar{w} \circ \phi)(X))|Y)]] \\ &= E_{Y \sim P_0}[E_{\phi(X) \sim P(.|Y)}(L(Y, (\bar{w} \circ \phi)(X))|Y)]\end{aligned}$$

The second equality is because $\phi(X) \perp\!\!\!\perp E|Y$.

Thus the objective function is the same across domains, so the optimal w is the same. Therefore $\phi \in \Phi_{\text{g-IRM}}(\mathcal{E})$ □

[7] This theorem looks similar to [Vei+21, Thm. 3.2]. This is deceptive; here we observe the environment E, whereas they assumed observations of the spurious factors Z.

References

[Ahu+20] K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar. "Invariant risk minimization games". In: *International Conference on Machine Learning*. PMLR. 2020.

[Alb+20] I. Albuquerque, J. Monteiro, M. Darvishi, T. H. Falk, and I. Mitliagkas. "Adversarial target-invariant representation learning for domain generalization". In: (2020).

[Arj+19] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. "Invariant risk minimization". In: *arXiv preprint arXiv:1907.02893* (2019).

[BCL21] J.-H. Bae, I. Choi, and M. Lee. "Meta-learned invariant risk minimization". In: *arXiv preprint arXiv:2103.12947* (2021).

[BDOR22] E. Ben-David, N. Oved, and R. Reichart. "Pada: example-based prompt learning for on-the-fly adaptation to unseen domains". In: *Transactions of the Association for Computational Linguistics* (2022).

[CB20] Y. Chen and P. Bühlmann. "Domain adaptation under structural causal models". In: *arXiv preprint arXiv:2010.15764* (2020).

[Com+20] R. Tachet des Combes, H. Zhao, Y.-X. Wang, and G. J. Gordon. "Domain adaptation with conditional distribution matching and generalized label shift". In: *Advances in Neural Information Processing Systems* (2020).

[Cub+19] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. "Autoaugment: learning augmentation strategies from data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[Gan+16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. "Domain-adversarial training of neural networks". In: *The journal of machine learning research* 1 (2016).

[Gar+19] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. "Counterfactual fairness in text classification through robustness". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

[Gei+18] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. "Generalisation in humans and deep neural networks". In: *Advances in neural information processing systems* (2018).

[Goe+20] K. Goel, A. Gu, Y. Li, and C. Ré. "Model patching: closing the subgroup performance gap with data augmentation". In: *arXiv preprint arXiv:2008.06775* (2020).

[GLP20] I. Gulrajani and D. Lopez-Paz. "In search of lost domain generalization". In: *arXiv preprint arXiv:2007.01434* (2020).

[GMZ19] H. Guo, Y. Mao, and R. Zhang. "Mixup as locally linear out-of-manifold regularization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 01. 2019.

[HG17] B. Hariharan and R. Girshick. "Low-shot visual recognition by shrinking and hallucinating features". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[Hen+21] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. "The many faces of robustness: a critical analysis of out-of-distribution generalization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[Hen+19] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. "Augmix: a simple data processing method to improve robustness and uncertainty". In: *arXiv preprint arXiv:1912.02781* (2019).

[Kob18] S. Kobayashi. "Contextual augmentation: data augmentation by words with paradigmatic relations". In: *arXiv preprint arXiv:1805.06201* (2018).

[Koh+21] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. "Wilds: a benchmark of in-the-wild distribution shifts". In: *International Conference on Machine Learning*. PMLR. 2021.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* (2012).

[Kru+21] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. "Out-of-distribution generalization via risk extrapolation (rex)". In: *International Conference on Machine Learning*. PMLR. 2021.

[LP20] P. Lemberger and I. Panico. "A primer on domain adaptation". In: *arXiv preprint arXiv:2001.09994* (2020).

[Li+18a] H. Li, S. J. Pan, S. Wang, and A. C. Kot. “Domain generalization with adversarial feature learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[Li+18b] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. “Deep domain generalization via conditional invariant adversarial networks”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[Liu+21] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn. “Just train twice: improving group robustness without training group information”. In: *International Conference on Machine Learning*. PMLR. 2021.

[Lon+18] M. Long, Z. Cao, J. Wang, and M. I. Jordan. “Conditional adversarial domain adaptation”. In: *Advances in neural information processing systems* (2018).

[Lu+21] C. Lu, Y. Wu, J. M. Hernández-Lobato, and B. Schölkopf. “Nonlinear invariant risk minimization: a causal approach”. In: *arXiv preprint arXiv:2102.12353* (2021).

[MH20] T. Matsuura and T. Harada. “Domain generalization using a mixture of multiple latent domains”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 07. 2020.

[MBS13] K. Muandet, D. Balduzzi, and B. Schölkopf. “Domain generalization via invariant feature representation”. In: *International Conference on Machine Learning*. PMLR. 2013.

[Nie+20] Y. Nie, Y. Tian, X. Wan, Y. Song, and B. Dai. “Named entity recognition for social media texts with semantic augmentation”. In: *arXiv preprint arXiv:2010.15458* (2020).

[Pas+19] M. Paschali, W. Simson, A. G. Roy, M. F. Naeem, R. Göbl, C. Wachinger, and N. Navab. “Data augmentation with manifold exploring geometric transformations for increased performance and robustness”. In: *arXiv preprint arXiv:1901.04420* (2019).

[PBM16] J. Peters, P. Bühlmann, and N. Meinshausen. “Causal inference by using invariant prediction: identification and confidence intervals”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 5 (2016).

[RPH21] A. Robey, G. J. Pappas, and H. Hassani. “Model-based domain generalization”. In: *Advances in Neural Information Processing Systems* (2021).

[RC+18] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. “Invariant models for causal transfer learning”. In: *The Journal of Machine Learning Research* 1 (2018).

[SLD02] M. Saerens, P. Latinne, and C. Decaestecker. “Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure”. In: *Neural computation* 1 (2002).

[Sag+19] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. “Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization”. In: *arXiv preprint arXiv:1911.08731* (2019).

[Sch+12] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. “On causal and anticausal learning”. In: *arXiv preprint arXiv:1206.6471* (2012).

[SHB15] R. Sennrich, B. Haddow, and A. Birch. “Improving neural machine translation models with monolingual data”. In: *arXiv preprint arXiv:1511.06709* (2015).

[SP20] R. D. Shah and J. Peters. “The hardness of conditional independence testing and the generalised covariance measure”. In: *The Annals of Statistics* 3 (2020).

[SFS17] B. Sun, J. Feng, and K. Saenko. “Correlation alignment for unsupervised domain adaptation”. In: *Domain Adaptation in Computer Vision Applications*. 2017.

[SS16] B. Sun and K. Saenko. “Deep coral: correlation alignment for deep domain adaptation”. In: *European conference on computer vision*. Springer. 2016.

[VCS16] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich. “Examining the impact of blur on recognition by convolutional networks”. In: *arXiv preprint arXiv:1611.05760* (2016).

[Vei+21] V. Veitch, A. D’Amour, S. Yadlowsky, and J. Eisenstein. “Counterfactual invariance to spurious correlations: why and how to pass stress tests”. In: *arXiv preprint arXiv:2106.00545* (2021).

[Wal+21] Y. Wald, A. Feder, D. Greenfeld, and U. Shalit. “On calibration and out-of-domain generalization”. In: *Advances in Neural Information Processing Systems* (2021).

[WZ19] J. Wei and K. Zou. “Eda: easy data augmentation techniques for boosting

performance on text classification tasks". In: *arXiv preprint arXiv:1901.11196* (2019).

[Wil+21] O. Wiles, S. Gowal, F. Stimberg, S. Alvise-Rebuffi, I. Ktena, T. Cemgil, et al. "A fine-grained analysis on distribution shift". In: *arXiv preprint arXiv:2110.11328* (2021).

[Xie+20] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. "Unsupervised data augmentation for consistency training". In: *Advances in Neural Information Processing Systems* (2020).