Nonlinear Invariant Risk Minimization: A Causal Approach

Chaochao Lu[1,2], Yuhuai Wu[3,5], Jośe Miguel Hernández-Lobato[1,4], and Bernhard Schölkopf[2]

[1]University of Cambridge, Cambridge, UK
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]University of Toronto, Toronto, Canada
[4]Alan Turing Institute, London, UK
[5]Vector Institute, Toronto, Canada

Abstract

Due to spurious correlations, machine learning systems often fail to generalize to environments whose distributions differ from the ones used at training time. Prior work addressing this, either explicitly or implicitly, attempted to find a data representation that has an invariant causal relationship with the target. This is done by leveraging a diverse set of training environments to reduce the effect of spurious features and build an invariant predictor. However, these methods have generalization guarantees only when both data representation and classifiers come from a linear model class. We propose Invariant Causal Representation Learning (ICRL), a learning paradigm that enables out-of-distribution (OOD) generalization in the nonlinear setting (i.e., nonlinear representations and nonlinear classifiers). It builds upon a practical and general assumption: the prior over the data representation factorizes when conditioning on the target and the environment. Based on this, we show identifiability of the data representation up to very simple transformations. We also prove that all direct causes of the target can be fully discovered, which further enables us to obtain generalization guarantees in the nonlinear setting. Extensive experiments on both synthetic and real-world datasets show that our approach significantly outperforms a variety of baseline methods. Finally, in the concluding discussion, we further explore the aforementioned assumption and propose a general view, called *the Agnostic Hypothesis*: there exist a set of hidden causal factors affecting both inputs and outcomes. The Agnostic Hypothesis can provide a unifying view of machine learning, be it supervised, unsupervised, or reinforcement learning, in terms of representation learning. More importantly, it can inspire a new direction to explore the general theory for identifying hidden causal factors, which is key to enabling the OOD generalization guarantees in machine learning.

1 Introduction

Despite impressive success stories, there is still a significant lack of robustness in machine learning algorithms. Specifically, machine learning systems may fail to generalize outside of a specific training distribution, because they learn easier-to-fit spurious correlations which are prone to change between training and testing environments. We illustrate this point by recalling the widely used example of classifying images of camels and cows (Beery et al., 2018). The training dataset has a selection bias, i.e., many pictures of cows are taken on green pastures, while most pictures of camels happen to be in deserts. After training a model on this dataset, it is found that the model builds on spurious correlations, i.e., it relates green pastures with cows and deserts with camels, thus classifying green pastures as cows and deserts as camels, and failing to recognize images of cows when they are taken on sandy beaches.

To address this problem, a natural idea is to identify which features of the training data present domain-varying spurious correlations with labels and which features describe true correlations of interest that are stable across domains. In the example above, the former are the features describing the context (e.g., pastures and deserts), whilst the latter are the features describing animals (e.g., animal shape). By exploiting the varying degrees of spurious correlation naturally present in training data collected from multiple environments one can try to identify stable features and build invariant predictors. Invariant risk minimization (IRM) seeks to find data representations (Arjovsky et al., 2019) or features (Rojas-Carulla et al., 2018) for which the optimal predictor is invariant across all environments. The general formulation of IRM is a challenging bi-leveled optimization problem, and theoretical guarantees require constraining both data representations and classifiers to be linear (Arjovsky et al., 2019, Theorem 9), or considering the special case of feature selection (Rojas-Carulla et al., 2018, Theorem 4). Ahuja et al. (2020) study the problem from the perspective of game theory, with an approach termed invariant risk minimization games (IRMG). They show that the set of Nash equilibria for a proposed game are equivalent to the set of invariant predictors for any finite number of environments, even with nonlinear data representations and nonlinear classifiers. However, these theoretical results in the nonlinear setting only guarantee that one can learn invariant predictors from training environments, but do not guarantee that the learned invariant predictors can generalize well across all environments including unseen testing environments. In other words, the generalization of IRMG is still limited to the linear case.

We propose a novel approach, referred to as Invariant Causal Representation Learning (ICRL), which enables out-of-distribution (OOD) generalization in the nonlinear setting (i.e., nonlinear representations and nonlinear classifiers[1]). We first introduce a practical and general assumption: the prior over the data representation has a factorized distribution when conditioning on the target (e.g., labels) and the environment (represented as an index). This lets us tackle

[1]In fact, we are not restricted to the classification case and allow the target to be either continuous or categorical, which will be formally defined in Section 2.2.

the problem of finding causal invariant representations by a combination of recent results from representation learning and techniques from graphical causal discovery, as follows. Utilizing identifiability results of identifiable variational autoencoders (Khemakhem et al., 2020), this assumption leads to a guarantee that the data representation can be identified up to certain simple transformations. Subject to this assumption, we then theoretically show that all the direct causes of the target can be fully discovered, by analyzing all possible graphs in a structural equation model setting, and building on conditional-independence based causal discovery methods. Once they are discovered, the challenging bi-leveled optimization problem in IRM and IRMG can be reduced to two simpler independent optimization problems, that is, learning the data representation and learning the optimal classifier can be performed separately. This leads to a practical algorithm and enables us to obtain generalization guarantees in the nonlinear setting.

It is worth noting that in the concluding discussion, we further explore the assumption introduced above and propose a general view, called *the Agnostic Hypothesis*: there exist a set of hidden causal factors affecting both inputs and outcomes. The Agnostic Hypothesis can provide a unifying view of machine learning, be it supervised, unsupervised, or reinforcement learning, in terms of representation learning. More importantly, it can inspire a new direction to explore a general theory for identifying hidden causal factors, which is key to enabling the OOD generalization guarantees in machine learning.

2 Preliminaries

2.1 Identifiable Variational Autoencoders

A general issue with variational autoencoders (VAEs)[2] (Kingma and Welling, 2013; Rezende et al., 2014) is the lack of identifiability guarantees of the deep latent variable model. That is, it is generally impossible to approximate the true joint distribution over observed and latent variables, including the true prior and posterior distributions over latent variables. Consider a simple latent variable model where $\boldsymbol{O} \in \mathbb{R}^d$ stands for an observed variable (random vector) and $\boldsymbol{X} \in \mathbb{R}^n$ for a latent variable. Khemakhem et al. (2020) showed that any model with unconditional latent distribution $p_{\boldsymbol{\theta}}(\boldsymbol{X})$ is unidentifiable. That is, we can always find transformations $t(\boldsymbol{X})$ such that $t(\boldsymbol{X})$ and $\boldsymbol{X}$ differ as random variables yet they are equal in distribution, $t(\boldsymbol{X}) \stackrel{d}{=} \boldsymbol{X}$. Hence, the primary assumption that they make to obtain an identifiability result is to posit a conditionally factorized prior distribution over the latent variables $p_{\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{U})$, where $\boldsymbol{U} \in \mathbb{R}^m$ is an additional observed variable (Hyvarinen et al., 2019). More specifically, let $\boldsymbol{\theta} = (\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda}) \in \Theta$ be the parameters of the conditional generative model

$$p_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{X}|\boldsymbol{U}) = p_{\boldsymbol{f}}(\boldsymbol{O}|\boldsymbol{X}) p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{X}|\boldsymbol{U}), \tag{1}$$

[2] A brief description of VAEs is given in Appendix A.

where $p_{\boldsymbol{f}}(\boldsymbol{O}|\boldsymbol{X}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{O}-\boldsymbol{f}(\boldsymbol{X}))$ in which $\boldsymbol{\epsilon}$ is an independent noise variable with probability density function $p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$, and the prior probability density function is assumed to belong to the exponential family and given by

$$p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{X}|\boldsymbol{U}) = \prod_i \mathcal{Q}_i(\boldsymbol{X}_i)/\mathcal{Z}_i(\boldsymbol{U}) \cdot \exp\left[\sum_{j=1}^{k} T_{i,j}(\boldsymbol{X}_i)\lambda_{i,j}(\boldsymbol{U})\right], \tag{2}$$

where $\mathcal{Q}_i$ is the base measure, $\boldsymbol{X}_i$ the i-th dimension of $\boldsymbol{X}$, $\mathcal{Z}_i(\boldsymbol{U})$ the normalizing constant, $\boldsymbol{T}_i = (T_{i,1}, \ldots, T_{i,k})$ the sufficient statistics, $\boldsymbol{\lambda}_i(\boldsymbol{U}) = (\lambda_{i,1}(\boldsymbol{U}), \ldots, \lambda_{i,k}(\boldsymbol{U}))$ the corresponding natural parameters depending on $\boldsymbol{U}$, and k the dimension of each sufficient statistic that is fixed in advance. It is worth noting that this assumption is not very restrictive as exponential families have universal approximation capabilities (Sriperumbudur et al., 2017). As in VAEs, we maximize the corresponding evidence lower bound,

$$\mathcal{L}_{\text{iVAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}) := \mathbb{E}_{p_D}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{U})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{X}|\boldsymbol{U}) - \log q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{U})\right]\right], \tag{3}$$

where we denote by p_D the empirical data distribution given by dataset $\mathcal{D} = \left\{\left(\boldsymbol{O}^{(i)}, \boldsymbol{U}^{(i)}\right)\right\}_{i=1}^{N}$. This approach is called identifiable VAE (iVAE). Most importantly, it can be proved that under the conditions stated in Theorem 2 of (Khemakhem et al., 2020), iVAE can identify the latent variables $\boldsymbol{X}$ up to a permutation and pointwise transformation defined as below,

Definition 1 (Permutation and Pointwise Transformation, Khemakhem et al. (2020))**.** *We say that the latent variables $\boldsymbol{X}$ are identified up to a permutation and pointwise transformation if $\boldsymbol{X}$ are related to their original $\boldsymbol{X}^*$ as follows:*

$$(\boldsymbol{T}_1^*(\boldsymbol{X}_1^*), \ldots, \boldsymbol{T}_n^*(\boldsymbol{X}_n^*)) = A\left(\boldsymbol{T}_1(\boldsymbol{X}_1), \ldots, \boldsymbol{T}_n(\boldsymbol{X}_n)\right),$$

where $\boldsymbol{T}_i^*$ and $\boldsymbol{T}_i$ are the sufficient statistics and A is a permutation matrix.

2.2 Invariant Risk Minimization

Arjovsky et al. (2019) introduced invariant risk minimization (IRM), whose goal is to construct an **invariant predictor** f that performs well across all environments $\mathcal{E}_{all}$ by exploiting data collected from multiple environments $\mathcal{E}_{tr}$, where $\mathcal{E}_{tr} \subseteq \mathcal{E}_{all}$. Technically, they consider datasets $D_e := \{(\boldsymbol{o}_i^e, \boldsymbol{y}_i^e)\}_{i=1}^{n_e}$ from multiple training environments $e \in \mathcal{E}_{tr}$, where $\boldsymbol{o}_i^e \in \mathcal{O} \subseteq \mathbb{R}^d$ is the input observation and its corresponding label is $\boldsymbol{y}_i^e \in \mathcal{Y} \subseteq \mathbb{R}^s$.[3] The dataset D_e, collected from environment e, consists of examples identically and independently distributed according to some probability distribution $P(\boldsymbol{O}^e, \boldsymbol{Y}^e)$. The goal of IRM is to use these multiple datasets to learn a predictor $\boldsymbol{Y} = f(\boldsymbol{O})$ that achieves the minimum risk for all the environments. Here we define the risk reached by f in environment e as $R^e(f) = \mathbb{E}_{\boldsymbol{O}^e, \boldsymbol{Y}^e}\left[\ell(f(\boldsymbol{O}^e), \boldsymbol{Y}^e)\right]$. Then, the invariant predictor can be formally defined as below.

[3] The setup applies to both continuous and categorical data. If any observation or label is categorical, we one-hot encode it.

Definition 2 (Invariant Predictor, Arjovsky et al. (2019)). *We say that a data representation $\Phi \in \mathcal{H}_\Phi : \mathcal{O} \to \mathcal{C}$ elicits an invariant predictor $w \circ \Phi$ across environments $\mathcal{E}$ if there is a classifier $w \in \mathcal{H}_w : \mathcal{C} \to \mathcal{Y}$ simultaneously optimal for all environments, that is, $w \in \arg\min_{\bar{w} \in \mathcal{H}_w} R^e(\bar{w} \circ \Phi)$ for all $e \in \mathcal{E}$.*

Mathematically, IRM can be phrased as the following constrained optimization problem:

$$\min_{\Phi \in \mathcal{H}_\Phi, w \in \mathcal{H}_w} \sum\nolimits_{e \in \mathcal{E}_{tr}} R^e(w \circ \Phi) \quad \text{s.t. } w \in \arg\min_{\bar{w} \in \mathcal{H}_w} R^e(\bar{w} \circ \Phi), \forall e \in \mathcal{E}_{tr}. \tag{4}$$

Since this is a generally infeasible bi-leveled optimization problem, Arjovsky et al. (2019) rephrased it as a tractable penalized optimization problem by transfering the inner optimization routine to a penalty term. The main generalization result (Theorem 9 in Arjovsky et al. (2019)) states that if both Φ and w come from the class of linear models (i.e., $\mathcal{H}_\Phi = \mathbb{R}^{n \times n}$ and $\mathcal{H}_w = \mathbb{R}^{n \times 1}$), under certain conditions on the diversity of training environments (Assumption 8 in Arjovsky et al. (2019)) and the data generation, the invariant predictor $w \circ \Phi$ across $\mathcal{E}_{tr}$ obtained by solving Eq. (4) remains invariant in $\mathcal{E}_{all}$.

3 Problem Setup

3.1 A Motivating Example

In this section, we extend the example which was introduced by Wright (1921) and discussed by Arjovsky et al. (2019), and provide a further in-depth analysis.

Model 1. *Consider a structural equation model (SEM) with a discrete environment variable E whose role is to modulate the noises in the structural assignments connecting the other variables (cf. Fig. 1a below):*

$$\begin{aligned} X_1 &\leftarrow \mathit{Gaussian}(0, \sigma_1(E)), \\ Y &\leftarrow X_1 + \mathit{Gaussian}(0, \sigma_2(E)), \\ X_2 &\leftarrow Y + \mathit{Gaussian}(0, \sigma_3(E)), \end{aligned}$$

where $\mathit{Gaussian}(0, \sigma)$ denotes a Gaussian random variable with zero mean and standard deviation $\sigma \geq 0$, and σ is a function of the value $e \in \mathcal{E}_{all}$ taken by the environment variable E, with $\mathcal{E}_{all}$ the set of all environments.

To ease exposition, here we consider the simple scenario in which $\mathcal{E}_{\text{all}}$ only contains all modifications varying the noises of X_1, X_2 and Y within a finite range, i.e., $\sigma_i(e) \in [0, \sigma^2_{\max}]$. Then, to predict Y from (X_1, X_2) using a least-square predictor $\hat{Y}^e = \hat{\alpha}_1 X_1^e + \hat{\alpha}_2 X_2^e$ for environment e, we can

- Case 1: regress from X_1^e, to obtain $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 0$,
- Case 2: regress from X_2^e, to obtain $\hat{\alpha}_1 = 0$ and $\hat{\alpha}_2 = \frac{\sigma_1(e)+\sigma_2(e)}{\sigma_1(e)+\sigma_2(e)+\sigma_3(e)}$,

- Case 3: regress from (X_1^e, X_2^e), to obtain $\hat{\alpha}_1 = \frac{\sigma_3(e)}{\sigma_2(e)+\sigma_3(e)}$ and $\hat{\alpha}_2 = \frac{\sigma_2(e)}{\sigma_2(e)+\sigma_3(e)}$.

In the generic scenario (i.e., $\sigma_1(e) \neq 0$, $\sigma_2(e) \neq 0$, and $\sigma_3(e) \neq 0$), the regression using X_1 in Case 1 is an invariant correlation: it is the only regression whose coefficients do not vary with the environment e. In contrast, the regressions in both Case 2 and Case 3 have coefficients that depend on the environment e. Therefore, only the invariant correlation in Case 1 will generalize well to new test environments.

From a practical perspective, let us take a closer look at Case 3. As we do not know in advance that regressing on X_1 alone will lead to an invariant predictor, in practice we may do the regression on all the available data (X_1^e, X_2^e). As explained, when $\sigma_i(e) \neq 0$ for $i = 1, 2, 3$, this fails to generalize. Indeed, no empirical risk minimization (ERM) algorithm (i.e., purely minimizing training error) (Vapnik, 1995) would work in this setting. Invariant Causal Prediction (ICP) methods (Peters et al., 2015) also do not work, since as argued by Arjovsky et al. (2019), the noise variance in Y may change across environments. To this end, Arjovsky et al. (2019) proposed IRM. As aforementioned, however, IRM and IRMG guarantee generalization to unseen environments only in the linear setting, while we consider the case where both Φ and w are from the class of nonlinear models.

A more straightforward way to understand the motivating example is in its corresponding graphical representation[4], as shown in Fig. 1a. Following Peters et al. (2015), we treat the environment as a random variable E, where E could be any information specific to the environment. For simplicity, we let E be the environment index, i.e., $E \in \{1, \ldots, N\}$ and N is the number of training environments. Note that here we consider E as a surrogate variable (Zhang et al., 2017; Huang et al., 2020). In fact, a more practical version appearing in real world problems is present in Fig. 1b, where the true variables $\{X_1, X_2\}$ are unobserved and we only can observe their transformation $\boldsymbol{O}$, which is a function of $\{X_1, X_2\}$. In this case, even if Y is not affected by E (i.e., remove the edge $E \to Y$), applying ICP to $\boldsymbol{O}$ still fails, since each variable (i.e., each dimension) of $\boldsymbol{O}$ is jointly influenced by both X_1 and X_2. By contrast, both IRM and IRMG work, as long as the transformation is linear. These findings are also empirically illustrated in Section 5.1.

3.2 Assumptions on Causal Graph

Our approach to addressing the problem presented in Fig. 1b when the transformation is nonlinear is to first identify both X_1 and X_2 from $\boldsymbol{O}$, then determine the direct cause X_1 of Y, and finally learn an invariant predictor based on X_1 alone. In this approach, the first step plays a fundamental role, since without it we are unable to perform the latter two steps. As discussed in Section 2.1,

[4] The relation between SEM and its graphical representation is formally defined in Appendix B.

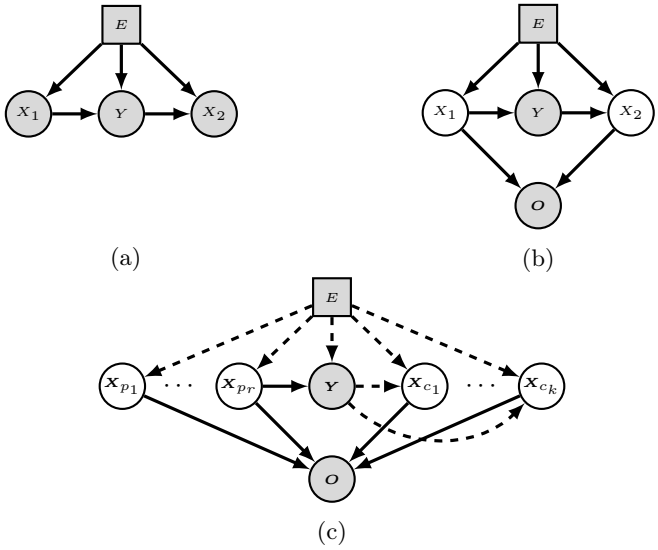


Figure 1: (a) Causal structure of Model 1. (b) A more practical extension of Model 1, where X_1 and X_2 are not directly observed and $\boldsymbol{O}$ is their observation. (c) A general version of (b), where we assume there exist multiple unobserved variables. Each of them could be either a parent, a child of $\boldsymbol{Y}$, or has no connection with $\boldsymbol{Y}$. Grey nodes denote observed variables and white nodes represent unobserved variables. Dashed lines denote the edges which might be absent in some scenarios.

however, it is generally impossible to identify X_1 and X_2 from $\boldsymbol{O}$ without additional assumptions. To obtain an identifiability result, we follow Khemakhem et al. (2020) and assume that the prior over X_1 and X_2 has a distribution that factorizes conditional on additional observed variables. In other words, the primary assumption leading to identifiability here is that each X_i is statistically dependent on some additional observed variables, but conditionally independent of the other X_j. In the example shown in Fig. 1b, apart from $\boldsymbol{O}$, both Y and E are observed, and thus are naturally treated as such additional observed variables. Hence, we formally have the following assumptions leading to identifiability:

Assumption 1. *$\boldsymbol{X}_i$ depends on one or both of $\boldsymbol{Y}$ and E for any i.*

Assumption 2. *$\boldsymbol{X}_i \perp\!\!\!\perp \boldsymbol{X}_j | \boldsymbol{Y}, E$ for any $i \neq j$.*

To allow for multi-dimensional variables, we will replace X_i and Y with $\boldsymbol{X}_i$ and $\boldsymbol{Y}$, respectively. The causal graph in Fig. 1b clearly satisfies Assumptions 1 and 2. As a result, when data are generated according to a data-generating process consistent with the causal graph in Fig. 1b, X_1 and X_2 can be identified from data.[5]

[5] This will be formally stated in Section 4.1, where some additional assumptions are required

While Fig. 1b is sufficient for Assumptions 1 and 2 to hold, it turns out that it is fortunately not necessary.

Indeed, the causal graph in Fig. 1b is not the only scenario satisfying Assumption 1&2, and there exist a large number of other scenarios. For example, if we remove the arrow from E to either or all of X_1, Y, and X_2, or if we remove the arrow from X_1 to Y or from X_1 to Y or both, all the resulting causal graphs still satisfy Assumption 1&2. In all these examples, X_1 and X_2 can be also identified from the data generated according to them. Thus, graphically, Assumption 1&2 only impose relatively weak constraints on the causal graph (causal relationships) over $\{X_1, X_2, Y, E\}$,[6] such that X_1 and X_2 can be identified from the data generated according to it.

When taking a closer look at Assumption 1, it is apparent that this assumption rules out all the useless X_i in the task of predicting Y. This is because, if Assumption 1 is violated, meaning that X_i is independent of Y and E and has no influence in predicting Y, then such X_i should be viewed as noises and thus eliminated during learning.

Assumption 2 can cover more scenarios than the common assumption that $X_i \perp\!\!\!\perp X_j$ for $i \neq j$ in latent variable models (e.g., disentanglement (Bengio et al., 2013; Locatello et al., 2019), autoencoders (Kingma and Welling, 2013; Rezende et al., 2014), ICA (Comon, 1994; Hyvärinen and Oja, 2000), etc.). If Y is caused by at most one of X_i and X_j, and no matter whether X_i and X_j are caused by E or not, then $X_i \perp\!\!\!\perp X_j | Y, E$ holds, but we may well have $X_i \not\perp\!\!\!\perp X_j$ (e.g., if Y causes both X_i and X_j, or if there is a chain $X_i \to Y \to X_j$). If Y causes or is caused by at most one of X_i and X_j, and at most one of X_i and X_j is caused by E, then both $X_i \perp\!\!\!\perp X_j$ and $X_i \perp\!\!\!\perp X_j | Y, E$ hold. The only corner case is when $\{Y, X_i, X_j\}$ form a collider $X_i \to Y \leftarrow X_j$ (no matter whether X_i and X_j are caused by E or not). In this case, we have $X_i \not\perp\!\!\!\perp X_j | Y, E$. In other words, if both X_i and X_j are causes of Y, then $X_i \perp\!\!\!\perp X_j | Y, E$ does not hold. However, this issue caused by Assumption 2 can be addressed by Assumption 2, too. Specifically, if Y has two causes X_i and X_j, then X_j would be absorbed into X_i[7] because they would be forced to satisfy Assumption 2 during learning. Accordingly, to allow for the corresponding causal graph to satisfy Assumption 2, we usually only retain X_i in the causal graph to represent all causes (i.e., X_i and X_j) of Y. The same applies to the setting in which Y has more than two causes, which we will discuss in the next section.

It is important to note that Assumption 2 holds true as long as not all X_i are the causes of Y. In reality, there exist many such scenarios[8]. One example is that the labels could be the causes of some features (i.e., $Y \to X_i$), which has been argued to be the case for digit and object recognition (Zhang et al.,

for the identifiability results.

[6] Note that the arrow from X_i to $\boldsymbol{O}$ must exist, because the observation $\boldsymbol{O}$ is generated as a function of each X_i.

[7] X_i could be a multi-dimensional variable so that it has enough capacity to absorb all the other causes of Y, which will be discussed in the next section.

[8] In fact, the case where all X_i are causes of Y is irrelevant for our algorithm, since in that case it is trivial to build an invariance predictor.

2013; Peters et al., 2017). Another is that some features are affected only by environments (i.e., $X_i \leftarrow E$), where E could be domain index (Muandet et al., 2013), time index (Hyvarinen and Morioka, 2016), previous data points in a time series (Khemakhem et al., 2020), related audio or text (Arandjelovic and Zisserman, 2017), etc.

3.3 The General Setting

As we have seen that the above analysis is not restricted to the two dimensional example (i.e., $\boldsymbol{X} = (X_1, X_2) \in \mathbb{R}^2$), here we explicitly extend it to a more general multi-dimensional setting[9] as shown in Fig. 1c. In particular, we now have $\boldsymbol{O} \in \mathcal{O} \subseteq \mathbb{R}^d$, $\boldsymbol{Y} \in \mathcal{Y} \subseteq \mathbb{R}^s$, $\boldsymbol{X} = (\boldsymbol{X}_{p_1}, \ldots, \boldsymbol{X}_{p_r}, \boldsymbol{X}_{c_1}, \ldots, \boldsymbol{X}_{c_k}) \in \mathcal{X} \subseteq \mathbb{R}^{n(r+k)}$, where we have assumed each $\boldsymbol{X}_i$ to be n-dimensional for simplicity. We also assume that $\boldsymbol{X}$ is of lower dimension than $\boldsymbol{O}$, that is, $n(r+k) \leq d$.

Note that to consider as many underlying causal graphs as possible in Fig. 1c, we use dashed arrows to indicate the possibility of causal connections, the strength of which remains to be determined from data. We only assume here that the underlying causal graph behind the data satisfies Assumption 1&2. This, as discussed in Section 3.2, is the key assumption to identify $\boldsymbol{X}$ from the data, which allows us to further determine the underlying causal graph behind the data and thus discover the causes of $\boldsymbol{Y}$.

Besides, as discussed in Section 3.2, if $\boldsymbol{Y}$ has multiple causes, all the causes will be absorbed into one of them during learning. Without loss of generality, we assume that $\boldsymbol{X}_{p_r}$ is such a cause absorbing all the other causes of $\boldsymbol{Y}$. Thus, as shown in Fig. 1c, except $\boldsymbol{X}_{p_r}$, none of $\{\boldsymbol{X}_{p_1}, \ldots, \boldsymbol{X}_{p_{r-1}}\}$ has an arrow pointing to $\boldsymbol{Y}$, meaning that all the parents of $\boldsymbol{Y}$ are absorbed into $\boldsymbol{X}_{p_r}$. It is worth emphasising that $\boldsymbol{X}_{p_r}$ is a multi-dimensional variable so it has enough capacity to absorb and represent all the parents of $\boldsymbol{Y}$.

As it covers many scenarios (e.g., the one of von Kügelgen et al. (2020)), the causal graph in Fig. 1c can be interpreted as a broad model for causally analysing the prediction of $\boldsymbol{Y}$ from $\boldsymbol{X}$. For instance, $\boldsymbol{X}_{c_1}$ might not connect to either of $\boldsymbol{Y}$ and E; $\boldsymbol{X}_{p_r}$ might have no connection with E; $\boldsymbol{Y}$ could be disconnected from E; etc. As such, it can account for settings in which (1) $\boldsymbol{X}$ can contain causal features (i.e., $\boldsymbol{X}_{p_r}$), spurious features (i.e., all the other $\boldsymbol{X}_i$ having connections to either or both of E and $\boldsymbol{Y}$) or both; (2) either, neither or both of $p(\boldsymbol{X})$ and $p(\boldsymbol{Y}|\boldsymbol{X})$ can change as a function of the environment E. The same applies to $p(\boldsymbol{Y})$ and $p(\boldsymbol{X}|\boldsymbol{Y})$.

[9] For simplicity, we do not explicitly consider unobserved confounders in this paper. In particular, we assume that there are no unobserved confounders between $X_i, \boldsymbol{Y}, \boldsymbol{O}$, and E. However, even if there were unobserved confounders, our approach would not be affected in some specific cases. For example, if there were an unobserved confounder between X_i and $\boldsymbol{O}$, it would be absorbed into X_i and our approach would not be affected.

Algorithm 1: Invariant Causal Representation Learning

Phase 1: We first learn the iVAE model, including the generative model and its corresponding inference model, by optimizing the evidence lower bound in Eq. (8) on the data $\{\boldsymbol{O}, \boldsymbol{Y}, E\}$. Then, we use the learned iVAE model to infer the corresponding latent variable $\boldsymbol{X}$ from $\{\boldsymbol{O}, \boldsymbol{Y}, E\}$, which is guaranteed to be identified up to a permutation and pointwise transformation.

Phase 2: Having obtained $\boldsymbol{X}$, according to Proposition 1, we can discover from them which are the direct causes (parents) $\text{Pa}(\boldsymbol{Y})$ of $\boldsymbol{Y}$ by only performing **Rule 1.2**, **Rule 1.6**, **Rule 2.1**, and **Rule 3.1** described in Appendix E.3.

Phase 3: Having obtained $\text{Pa}(\boldsymbol{Y})$, we can separately optimize Eq. (9) and Eq. (10) to learn the invariant data representation Φ and the invariant classifier w.

4 Our Approach

In this section, we formally introduce our algorithm Invariant Causal Representation Learning (ICRL) to address the nonlinear IRM problem in the general setting. It consists of three phases summarized in Algorithm 1. The basic idea is that we first identify the latent variables $\boldsymbol{X}$ by leveraging iVAE under Assumptions 1&2 (Phase 4.1), then discover direct causes of $\boldsymbol{Y}$ (Phase 4.2), and finally learn an invariant predictor based on the identified direct causes (Phase 4.3).

4.1 Phase 1: Identifying Latent Variables Using iVAE

Under Assumptions 1&2, it is straightforward to identify the true hidden factors $\boldsymbol{X}$ from $\boldsymbol{O}$ with the help of $\boldsymbol{Y}$ and E by leveraging iVAE. We can directly substitute $\boldsymbol{U}$ with $(\boldsymbol{Y}, E)$ in Eq. (1), and obtain the generative model

$$p_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{X}|\boldsymbol{Y}, E) = p_{\boldsymbol{f}}(\boldsymbol{O}|\boldsymbol{X}) p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{X}|\boldsymbol{Y}, E), \tag{5}$$

$$p_{\boldsymbol{f}}(\boldsymbol{O}|\boldsymbol{X}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{O} - \boldsymbol{f}(\boldsymbol{X})). \tag{6}$$

Likewise, we obtain its corresponding prior distribution and lower bound

$$p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{X}|\boldsymbol{Y}, E) = \prod\nolimits_i \mathcal{Q}_i(\boldsymbol{X}_i)/\mathcal{Z}_i(\boldsymbol{Y}, E) \cdot \exp\left[\sum\nolimits_{j=1}^{k} T_{i,j}(\boldsymbol{X}_i)\lambda_{i,j}(\boldsymbol{Y}, E)\right], \tag{7}$$

$$\mathcal{L}_{\text{phase1}}(\boldsymbol{\theta}, \boldsymbol{\phi}) := \mathbb{E}_{p_D}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{X}|\boldsymbol{Y}, E) - \log q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{Y}, E)\right]\right]. \tag{8}$$

This bound can be further derived for computational convenience, cf. Appendix C.

We next rephrase the identifiability result of Khemakhem et al. (2020), replacing $\boldsymbol{U}$ with $(\boldsymbol{Y}, E)$:

Theorem 1. *Assume that we observe data sampled from a generative model defined according to Eqs. (5-7), with parameters* $\boldsymbol{\theta} := (\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda})$ *and* $k \geq 2$. *Assume the following holds: (i) The set* $\{\boldsymbol{O} \in \mathcal{O} | \varphi_{\boldsymbol{\epsilon}}(\boldsymbol{O}) = 0\}$ *has measure zero, where* $\varphi_{\boldsymbol{\epsilon}}$ *is the characteristic function of the density* $p_{\boldsymbol{\epsilon}}$ *defined in Eq. (6). (ii) The mixing function* $\boldsymbol{f}$ *in Eq. (6) is injective, and has all second order cross derivatives. (iii) The sufficient statistics* $T_{i,j}$ *in Eq. (7) are twice differentiable, and* $(T_{i,j})_{1 \leq j \leq k}$ *are linearly independent on any subset of* $\mathcal{X}$ *of measure greater than zero. (iv) There exist* $nk + 1$ *distinct points* $(\boldsymbol{Y}, E)^0, \ldots, (\boldsymbol{Y}, E)^{nk}$ *such that the matrix* $L = \big(\boldsymbol{\lambda}((\boldsymbol{Y}, E)^1) - \boldsymbol{\lambda}((\boldsymbol{Y}, E)^0), \ldots, \boldsymbol{\lambda}((\boldsymbol{Y}, E)^{nk}) - \boldsymbol{\lambda}((\boldsymbol{Y}, E)^0)\big)$ *of size* $nk \times nk$ *is invertible. Then the parameters* $\boldsymbol{\theta}$ *are identifiable up to a permutation and pointwise transformation.*

Theorem 1 deals with the case $k \geq 2$. In Appendix D, we provide a result dealing with the case $k = 1$. We further have the following consistency result for the estimation.

Theorem 2. *Assume the following holds: (i) The family of distributions* $q_\phi(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{Y}, E)$ *contains* $p_\phi(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{Y}, E)$. *(ii) We maximize* $\mathcal{L}_{phase1}(\boldsymbol{\theta}, \boldsymbol{\phi})$ *with respect to both* $\boldsymbol{\theta}$ *and* $\boldsymbol{\phi}$. *Then in the limit of infinite data, iVAE learns the true parameters* $\boldsymbol{\theta}^*$ *up to a permutation and pointwise transformation.*

As a consequence of Theorem 1&2, we have:

Theorem 3. *Assume the hypotheses of Theorem 1 and Theorem 2 hold, then in the limit of infinite data, iVAE learns the true latent variables* $\boldsymbol{X}^*$ *up to a permutation and pointwise transformation.*

The proofs of Theorem 2 and Theorem 3 are given in Appendix E. Theorem 3 states that we can leverage iVAE to learn the true conditionally factorized latent variables up to a permutation and pointwise transformation, which achieves our goal as required by Assumptions 1&2.

4.2 Phase 2: Discovering Direct Causes

After identifying all the conditionally independent latent variables $\boldsymbol{X}$ from $\boldsymbol{O}$, the next question is how to determine which component(s) of $\boldsymbol{X}$ is a direct cause of $\boldsymbol{Y}$, denoted by $\mathrm{Pa}(\boldsymbol{Y})$. Assumption 2 allows us to assess in parallel whether or not $\boldsymbol{X}_i$ is a direct cause of $\boldsymbol{Y}$. Fig. 1c shows that there exist only five possible connections between $\boldsymbol{X}_i$, $\boldsymbol{Y}$, E, and $\boldsymbol{O}$, as illustrated in Fig. 2a. Among them, only the arrow from $\boldsymbol{X}_i$ to $\boldsymbol{O}$ must exist, because $\boldsymbol{O}$ is generated as a function of the $\boldsymbol{X}_i$. The other four might be not necessarily, but with the constraint that there must exist one connection between $\boldsymbol{X}_i$ and $(\boldsymbol{Y}, E)$ (Assumption 1). This leaves us with ten possible types of structures, shown in Figs. 2b-2k. It turns out that these structures can be efficiently identified, and for each $\boldsymbol{X}_i$ we can assess in parallel whether or not it is the direct cause of $\boldsymbol{Y}$. Note that only in four of them (i.e., Figs. 2b, 2e, 2g, and 2j) does $\boldsymbol{X}_i$ serve as a parent of $\boldsymbol{Y}$. Here, the reason that we also include Figs. 2g and 2j is that we follow Arjovsky et al. (2019) and consider a more general definition of invariance which allows for

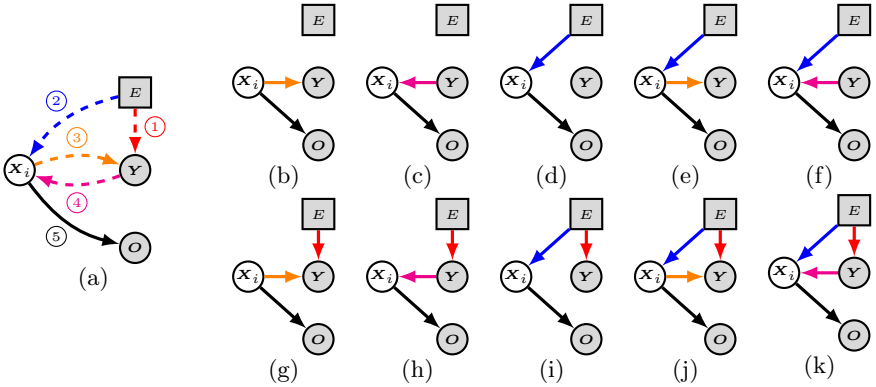


Figure 2: (a) General causal structure over $\{\boldsymbol{X}_i, \boldsymbol{Y}, \boldsymbol{O}, E\}$, where the arrow from $\boldsymbol{X}_i$ to $\boldsymbol{O}$ is a must-have connection and the other four might not be necessarily. (b) Ten possible causal structures from (a) under Assumptions 1&2.

changes in the noise variance of $\boldsymbol{Y}$. For convenience, we include the definition in Appendix B. Given the data $\{\boldsymbol{X}_i, \boldsymbol{Y}, E, \boldsymbol{O}\}$, we are able to distinguish all ten structures by performing conditional independence tests (Spirtes et al., 2000; Zhang et al., 2012), leveraging causal discovery algorithms (Peters et al., 2017; Zhang et al., 2017; Huang et al., 2020). This is summarized in Proposition 1, proven in Appendix E.

Proposition 1. *The ten structures shown in Figs. 2b-2k can be identified in parallel using causal discovery methods consistent in the infinite sample limit.*

Note that for each $\boldsymbol{X}_i$, we only need to check if they are one of the four structures (i.e., Figs. 2b, 2e, 2g, and 2j), and this check can be also performed in parallel.

4.3 Phase 3: Learning an Invariant Predictor

Having obtained the invariant causal representation $\text{Pa}(\boldsymbol{Y})$ for $\boldsymbol{Y}$ across the training environments, learning an invariant predictor $w \circ \Phi$ in Eq. (4) is then reduced to two simpler independent optimization problems: (i) learning the invariant data representation Φ from $\boldsymbol{O}$ to $\text{Pa}(\boldsymbol{Y})$, and (ii) learning the invariant classifier w from $\text{Pa}(\boldsymbol{Y})$ to $\boldsymbol{Y}$. Mathematically, these two optimization problems can be respectively phrased as

$$\min_{\Phi \in \mathcal{H}_\Phi} \sum\nolimits_{e \in \mathcal{E}_{tr}} R^e(\Phi) = \min_{\Phi \in \mathcal{H}_\Phi} \sum\nolimits_{e \in \mathcal{E}_{tr}} \mathbb{E}_{\boldsymbol{O}^e, \text{Pa}(\boldsymbol{Y}^e)} \left[\ell(\Phi(\boldsymbol{O}^e), \text{Pa}(\boldsymbol{Y}^e)) \right], \tag{9}$$

$$\min_{w \in \mathcal{H}_w} \sum\nolimits_{e \in \mathcal{E}_{tr}} R^e(w) = \min_{w \in \mathcal{H}_w} \sum\nolimits_{e \in \mathcal{E}_{tr}} \mathbb{E}_{\text{Pa}(\boldsymbol{Y}^e), \boldsymbol{Y}^e} \left[\ell(w(\text{Pa}(\boldsymbol{Y}^e)), \boldsymbol{Y}^e) \right]. \tag{10}$$

The expressions (9) and (10) ensure that ICRL can achieve low error across $\mathcal{E}_{tr}$. Also, in Phase 1&2, we showed that ICRL can enforce invariance across $\mathcal{E}_{tr}$. Now this brings us to the question: how to enable the OOD generalization?

In other words, how does ICRL achieve low error across $\mathcal{E}_{all}$? As Arjovsky et al. (2019) pointed out, low error across $\mathcal{E}_{tr}$ and invariance across $\mathcal{E}_{all}$ lead to low error across $\mathcal{E}_{all}$ since the generalization error of $w \circ \Phi$ follows standard error bounds once the data representation Φ eliciting an invariant predictor $w \circ \Phi$ across $\mathcal{E}_{all}$ is estimated. Thus, securing OOD generalization finally reduces to the following question: under which conditions does invariance across $\mathcal{E}_{tr}$ imply invariance across $\mathcal{E}_{all}$? Not surprisingly, $\mathcal{E}_{tr}$ must contain sufficient diversity to satisfy an underlying invariance across $\mathcal{E}_{all}$. Fortunately, the hypotheses of Theorem 1 automatically provides such a guarantee, and we therefore have the following result whose proof is in Appendix E.

Proposition 2. *Subject to the assumptions of Theorem 1 and Theorem 2, if ICRL learns an invariant predictor $w \circ \Phi$ across $\mathcal{E}_{tr}$ in the limit of infinite data, then such a predictor $w \circ \Phi$ is invariant across $\mathcal{E}_{all}$.*

5 Experiments

We compare our approach with a variety of methods on both synthetic and real-world datasets. The supplement contains a detailed description of the datasets (Appendix F) and model architectures (Appendix G). In all comparisons, unless stated otherwise, we averaged the performance of the different methods over ten runs.

5.1 Synthetic data

We first conduct a series of experiments on synthetic data generated according to an extension of the SEM in Model 1. The extension consists of increasing the dimensionality of the two input features $\boldsymbol{X} := (X_1, X_2)$ to 10 dimensions through a linear or nonlinear transformation, as illustrated in Fig. 1b. Technically, the goal is to predict Y from $\boldsymbol{O}$, where $\boldsymbol{O} = g(\boldsymbol{X})$ and $g(\cdot)$ is called $\boldsymbol{X}$ Transformer. We consider three types of transformations:

(a) *Identity*: $g(\cdot)$ is the identity matrix $\boldsymbol{I} \in \mathbb{R}^{2\times 2}$, i.e., $\boldsymbol{O} = g(\boldsymbol{X}) = \boldsymbol{X}$.

(b) *Linear*: $g(\cdot)$ is a random matrix $\boldsymbol{S} \in \mathbb{R}^{2\times 10}$, i.e., $\boldsymbol{O} = g(\boldsymbol{X}) = \boldsymbol{X} \cdot \boldsymbol{S}$.

(c) *Nonlinear*: $g(\cdot)$ is implemented by a multilayer perceptron with 2-dimensional input and 10-dimensional output, whose parameters are randomly set in advance.

For simplicity, here we consider the regression task, in which the mean squared error (MSE) is used as a metric.

We consider a simple scenario in which we fix $\sigma_1 = 1$ and $\sigma_2 = 0$ for all environments and only allow σ_3 to vary across environments. In this case, σ_3 controls how much the representation depends on the variable X_2, which is responsible for the spurious correlations. Each experiment draws 1000 samples from each of the three environments $\sigma_3 = \{0.2, 2, 100\}$, where the first two are

Table 1: Regression on synthetic data: Comparison of methods in terms of MSE (mean ± std deviation).

X TRANSFORMER	ALGORITHM	TRAIN MSE ($\sigma_3 = \{0.2, 2\}$)	TEST MSE ($\sigma_3 = 100$)
Identity	ERM	0.00 ± 0.00	0.00 ± 0.00
	IRM	0.00 ± 0.00	0.00 ± 0.00
	F-IRM GAME	0.81 ± 0.37	9.58 ± 17.10
	V-IRM GAME	0.80 ± 0.24	241.01 ± 301.52
	ICRL (ours)	0.01 ± 0.03	1.00 ± 0.02
Linear	ERM	0.00 ± 0.00	0.00 ± 0.00
	IRM	0.00 ± 0.00	0.00 ± 0.00
	F-IRM GAME	0.99 ± 0.01	1.11 ± 0.16
	V-IRM GAME	0.89 ± 0.22	380.88 ± 759.94
	ICRL (ours)	0.01 ± 0.03	1.02 ± 0.03
Nonlinear	ERM	0.06 ± 0.01	220.79 ± 229.97
	IRM	0.08 ± 0.01	149.60 ± 104.85
	F-IRM GAME	64.48 ± 42.18	360114.72 ± 241206.47
	V-IRM GAME	0.90 ± 0.21	780.67 ± 729.68
	ICRL (ours)	$\mathbf{0.31 \pm 0.03}$	$\mathbf{30.38 \pm 3.67}$

Table 2: Results on synthetic Data: Comparison of iVAE and VAE used in Phase 1 in terms of MSE (mean ± std deviation).

X TRANSFORMER	ALGORITHM	TRAIN MSE ($\sigma_3 = \{0.2, 2\}$)	TEST MSE ($\sigma_3 = 100$)
Nonlinear	ICRL-VAE	0.26 ± 0.09	1174.96 ± 1385.81
	ICRL-iVAE	$\mathbf{0.31 \pm 0.03}$	$\mathbf{30.38 \pm 3.67}$

for training and the third for testing. We compare with several baselines:[10] ERM, and two variants of IRMG: F-IRM Game (with Φ fixed to the identity) and V-IRM Game (with variable Φ).

As shown Table 1, in the cases of *Identity* and *Linear*, our approach is better than IRMG but only comparable with ERM and IRM. This might be because the identifiability result up to a pointwise nonlinear transformation renders the problem more difficult by converting the original identity or linear problem to a nonlinear problem. In the *Nonlinear* case, it is clear that the advantage of our approach becomes more obvious as the spurious correlation becomes stronger.

We also perform a series of experiments to further analyze the three phases of our approach, including an analysis of the importance of Assumptions 1&2 and of the necessity of iVAE in Phase 1, how accurately the direct causes can be recovered in Phase 2, and how well the two optimization problems can be

[10] We also tried ICP, but ICP was unable to find any parent of Y even in the identity case.

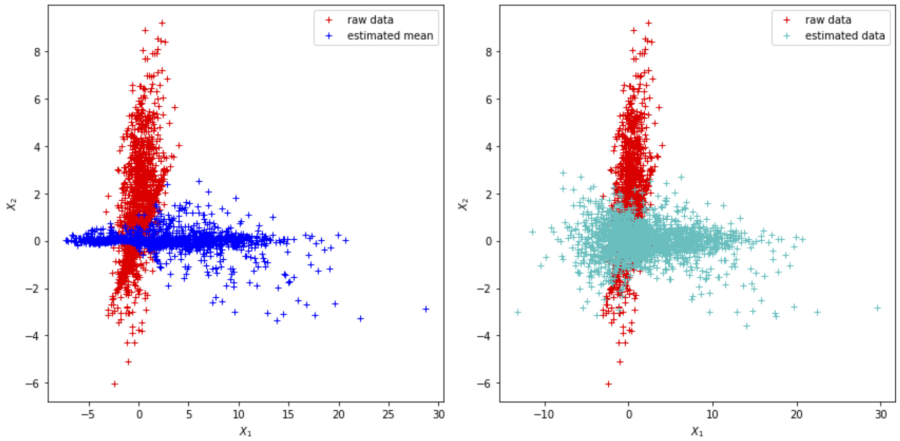


Figure 3: Left: Comparison of the raw data $\boldsymbol{X}$ and the mean of $\hat{\boldsymbol{X}}$ inferred through the learned inference model in the *Nonlinear* case. Right: Comparison of the raw data $\boldsymbol{X}$ and the sampled points $\hat{\boldsymbol{X}}$ using the reparameterization trick in the *Nonlinear* case. The comparisons clearly show that the inferred $\hat{\boldsymbol{X}}$ is equal to $\boldsymbol{X}$ up to a permutation and pointwise transformation.

addressed in Phase 3.

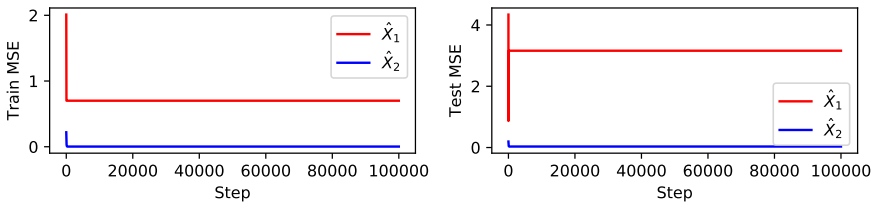
5.1.1 Analyzing Phase 1

Theorem 3 tells us that we can leverage iVAE to learn the true conditionally factorized latent variables up to a permutation and pointwise transformation. We empirically verify this point by comparing the raw data $\boldsymbol{X}$ with the corresponding $\hat{\boldsymbol{X}}$ inferred through the learned inference model in iVAE. Fig. 3 clearly shows that the inferred $\hat{\boldsymbol{X}}$ is equal to $\boldsymbol{X}$ up to a permutation and pointwise transformation.

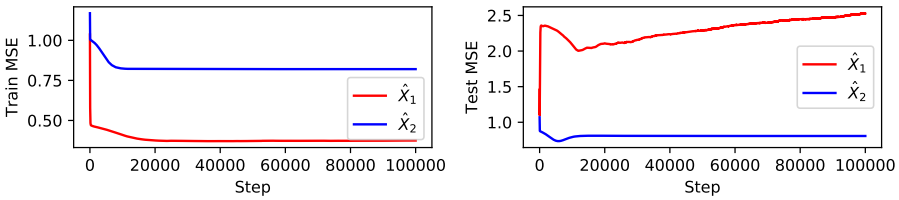
To show the importance of iVAE, we conduct an experiment in which we replace iVAE with the original VAE in Phase 1. As shown in Table 2, the performance of ICRL based on iVAE significantly outperforms the one based on VAE. It is worth noting that when VAE is instead used in Phase 1, it usually occurs in Phase 2 that either all the dimensions or no dimension of $\hat{\boldsymbol{X}}$ are identified as the parents of $\boldsymbol{Y}$. This is because all components of $\hat{\boldsymbol{X}}$ are mixed together and will influence one another even when conditioning on $\boldsymbol{Y}$ and E.

5.1.2 Analyzing Phase 2

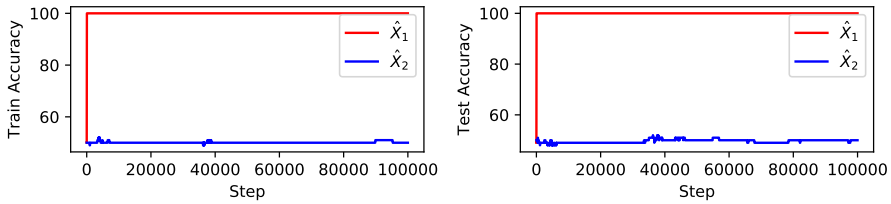
To show how well our method can identify the direct causes of $\boldsymbol{Y}$ in Phase 2, we compare the final performance when the identified direct cause (i.e., X_1) and the identified non-cause (i.e., X_2) are respectively used in Phase 3 to learn the predictor. Note that, there might exist a permutation between the inferred



(a) Regression results in the *Linear* case in terms of MSE, where the inferred $\hat{X}_2$ is the identified cause.



(b) Regression results in the *Nonlinear* case in terms of MSE, where the inferred $\hat{X}_2$ is the identified cause.



(c) Classification results on Synthetic Data in terms of accuracy, where the inferred $\hat{X}_1$ is the identified cause.

Figure 4: Comparison of the inferred $\hat{X}_1$ and $\hat{X}_2$ in terms of their final performance.

$\{\hat{X}_1, \hat{X}_2\}$ and the true $\{X_1, X_2\}$. Fig. 4a and Fig. 4b show the results on the regression task, from which we can obviously see that the predictor elicited by the identified cause has a much better generalization performance. The classification result (i.e., Y is binarized) in Fig. 4c further demonstrates this point.

5.1.3 Analyzing Phase 3

In this experiment, we want to verify how well the data representation Φ can be learned by optimizing the loss in Eq. (9). The main idea is to check how well the learned $\hat{\Phi}$ can purely extract the cause X_1 from $\boldsymbol{O}$. For this, we first learn

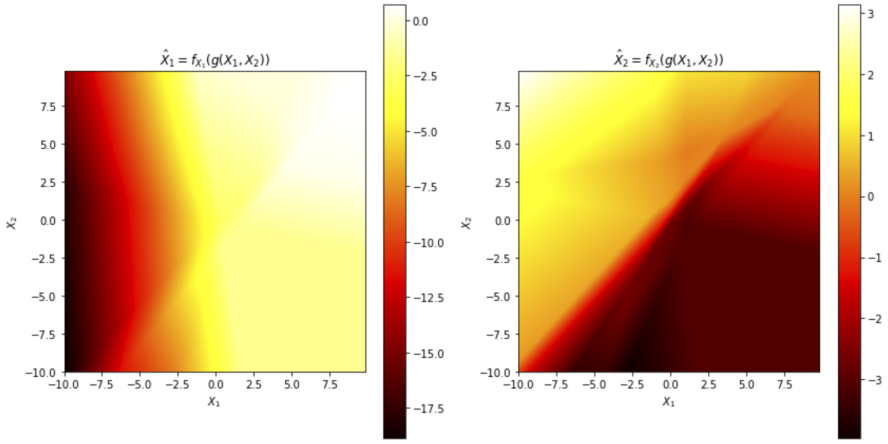


Figure 5: Left: Energy plot of $\hat{X}_1 = f_{X_1}(g(\boldsymbol{X}))$. Right: Energy plot of $\hat{X}_2 = f_{X_2}(g(\boldsymbol{X}))$. Note that, here $\hat{\Phi}_{X_i}$ is denoted by f_{X_i}.

$\hat{\Phi}_{X_i}$ for each X_i. Formally, we have

$$\hat{X}_i = \hat{\Phi}_{X_i}(\underbrace{g(\boldsymbol{X})}_{\boldsymbol{O}}), \text{ where } \boldsymbol{X} = (X_1, X_2).$$

Then, we observe how $\hat{X}_i$ will change while tuning X_1 and X_2 respectively. Fig. 5 shows the energy plots when $\hat{X}_1$ is the identified cause of Y and $\hat{X}_2$ the child of Y. Note that, in the plots, $\hat{\Phi}_{X_i}$ is denoted by f_{X_i}. In theory, we are able to learn an invariant data representation Φ_{X_1} for X_1 from $\boldsymbol{O}$, because there is no spurious correlation between X_1 and $\boldsymbol{O}$. By contrast, we cannot learn an invariant data representation Φ_{X_2} for X_2 from $\boldsymbol{O}$, because there exist spurious correlations between X_2 and $\boldsymbol{O}$. The results shown in Fig. 5 clearly verify our theory. Specifically, in the left plot, $\hat{X}_1$ remains approximately unchanged when changing X_2 but it changes when changing X_1. However, in the right plot, $\hat{X}_2$ changes whether we change X_1 or X_2.

5.2 Colored MNIST and Colored Fashion MNIST

In this section, we report experiments on two datasets used in IRM and IRMG: Colored MNIST and Colored Fashion MNIST. We follow the setting of Ahuja et al. (2020) to create these two datasets. The task is to predict a binary label assigned to each image which is originally grayscale but artificially colored in a way that correlated strongly but spuriously with the class label. We add noise to the preliminary label (i.e., $Y = \{0, 1\}$) by flipping it with 25 percent probability to construct the final labels. We sample the color id by flipping the final labels with probability p_e, where p_e is 0.2 in the first environment, 0.1 in the second environment, and 0.9 in the third environment. The third environment is the testing environment. We compare with 1) IRM, 2) two variants of

Table 3: Colored Fashion MNIST: Comparison of methods in terms of accuracy (%) (mean ± std deviation).

ALGORITHM	TRAIN ACCURACY	TEST ACCURACY
ERM	83.17 ± 1.01	22.46 ± 0.68
ERM 1	81.33 ± 1.35	33.34 ± 8.85
ERM 2	84.39 ± 1.89	13.16 ± 0.82
ROBUST MIN MAX	82.81 ± 0.11	29.22 ± 8.56
F-IRM GAME	62.31 ± 2.35	69.25 ± 5.82
V-IRM GAME	68.96 ± 0.95	70.19 ± 1.47
IRM	75.01 ± 0.25	55.25 ± 12.42
ICRL (ours)	$\mathbf{74.32 \pm 0.43}$	$\mathbf{73.14 \pm 0.56}$
ERM GRAYSCALE	74.79 ± 0.37	74.67 ± 0.48
OPTIMAL	75	75

Table 4: Colored MNIST: Comparison of methods in terms of accuracy (%) (mean ± std deviation).

ALGORITHM	TRAIN ACCURACY	TEST ACCURACY
ERM	84.88 ± 0.16	10.45 ± 0.66
ERM 1	84.84 ± 0.21	10.86 ± 0.52
ERM 2	84.95 ± 0.20	10.05 ± 0.23
ROBUST MIN MAX	84.25 ± 0.43	15.24 ± 2.45
F-IRM GAME	63.37 ± 1.14	59.91 ± 2.69
V-IRM GAME	63.97 ± 1.03	49.06 ± 3.43
IRM	59.27 ± 4.39	62.75 ± 9.59
ICRL (ours)	$\mathbf{70.34 \pm 0.29}$	$\mathbf{66.21 \pm 1.42}$
ERM GRAYSCALE	71.81 ± 0.47	71.36 ± 0.65
OPTIMAL	75	75

IRMG: F-IRM Game (with Φ fixed to the identity) and V-IRM Game (with a variable Φ), 3) three variants of ERM: ERM (on entire training data), ERM e (on each environment e), and ERM GRAYSCALE (on data with no spurious correlations), and 4) ROBUST MIN MAX (minimizing the maximum loss across the multiple environments). Table 3 shows that our approach outperforms all others on Colored Fashion MNIST. It is worth emphasising that both train and test accuracies of our method closely approach the ones of ERM GRAYSCALE and OPTIMAL, implying that it does approximately learn the true invariant causal representation with almost no correlation with the color. We can draw a similar conclusion from the results on Colored MNIST (Table 4). However, this dataset seems more difficult than the fashion version, because even ERM GRAYSCALE, where the spurious correlation with color is removed, falls well short of the optimum. In this case, two training environments might be not enough to eliminate all the spurious correlations.

6 Related Work

Invariant Prediction Peters et al. (2015) introduced the method of Invariant Causal Prediction (ICP), aiming to find the *causal feature set* (i.e., all direct causes of a target variable of interest) by exploiting the invariance property in causality which has been widely discussed under the term "autonomy", "modularity", and "stability" (Haavelmo, 1944; Aldrich, 1989; Hoover, 1990; Pearl, 2009; Dawid et al., 2010; Schölkopf et al., 2012). This invariance property assumed in ICP and its nonlinear extension (Heinze-Deml et al., 2018) is limited, because no intervention is allowed on the target variable $\boldsymbol{Y}$. Besides, ICP methods implicitly assume that the variables of interest $\boldsymbol{X}$ are given. Magliacane et al. (2018) and Subbaswamy et al. (2019) attempt to find invariant predictors that are maximally predictive using conditional independence tests and other graph-theoretic tools, both of which also assume that the $\boldsymbol{X}$ are given and further assume that additional information about the structure over $\boldsymbol{X}$ is known. Arjovsky et al. (2019) reformulate this invariance as an optimization-based problem, allowing us to learn an invariant data representation from $\boldsymbol{O}$ constrained to be a linear transformation of $\boldsymbol{X}$. Ahuja et al. (2020) extend IRM to the nonlinear setting from the perspective of game theory, but their nonlinear theory holds only in training environments. Rosenfeld et al. (2020) demonstrate that IRM and subsequent works have significant under-explored risks and issues with their formulation. Nagarajan et al. (2020) identify that spurious correlations during training can induce two distinct skews in the training set, one geometric and another statistical, which could result in two complementary ways by which gradient-descent-trained ERM is guaranteed to rely on those spurious correlations.

Domain Generalization Domain generalization emphasizes the ability to transfer acquired knowledge to domains unseen during training. A wide range of methods has been proposed for learning domain-invariant representations. Ben-David et al. (2007) formalize the intuition that a good representation is the key to effective domain adaptation, theoretically with a generalization bound. Khosla et al. (2012) develop a max-margin classifier that explicitly exploits the effect of dataset bias and improves generalization ability to unseen domains. Fang et al. (2013) propose a metric learning approach based on structural SVMs such that the neighbors of each training sample consist of examples from both the same and different domains. Muandet et al. (2013) propose a kernel-based optimization algorithm called Domain-Invariant Component Analysis (DICA), which aims to both minimize the discrepancy among domains and prevent the loss of relationship between input and output features. Ghifary et al. (2015) train a multi-task autoencoder that recognizes invariances among domains by learning to reconstruct analogs of original inputs from different domains. Motiian et al. (2017) learn an embedding subspace where samples from different domains are close if they have the same class labels, and far apart otherwise. Li et al. (2018b) minimize the differences in joint distributions to achieve target domain generalization through a conditional invariant adversarial network. Li

et al. (2018a) build on adversarial autoencoders by considering maximum mean discrepancy regularization and aligning the domains' distributions.

7 Looking Forward: The Agnostic Hypothesis

In this paper, we developed a novel framework to learn invariant predictors from a diverse set of training environments. As we have seen, the key to our approach is how to identify all the direct causes of the outcome, both theoretically and practically. Our implementation is predicated on Assumptions 1&2, that the data representation can be factorized when conditioning on the outcome and the environment, as shown in Fig. 1c. More importantly, when taking a closer look at Fig. 1c, one more fundamental underlying assumption is that there exist a set of hidden causal factors ($\boldsymbol{X}_{p_r}$) affecting both input images ($\boldsymbol{O}$) and class labels ($\boldsymbol{Y}$), that is,

$$\boldsymbol{O} \leftarrow \boldsymbol{X}_{p_r} \rightarrow \boldsymbol{Y}. \tag{11}$$

In other words, $\boldsymbol{X}_{p_r}$ is the common cause of both $\boldsymbol{O}$ and $\boldsymbol{Y}$. We call this underlying assumption ***the Agnostic Hypothesis***, which is originally presented in Lu (2020a,b) and inspired by the discussion on the example of *optical character recognition* in Peters et al. (2017). The Agnostic Hypothesis provides a more generally reasonable way to look at image classification, in comparison to two existing popular but mutually exclusive opinions: **Causal** (i.e., labels are viewed as an effect of images) and **Anticausal** (i.e., images are viewed as an effect of labels). In fact, the Agnostic Hypothesis is more highly beneficial to identifying hidden causal factors $\boldsymbol{X}_{p_r}$. Before diving into the reason behind it, for completeness we first still use the previous digit recognition example to briefly summarise the two exiting opinions.

7.1 Opinion 1: Anticausal

One dominant viewpoint is that digit recognition is an anticausal problem, i.e., predicting cause from effect, where the observed digit image is viewed as an effect of its label which is encoded as an high level concept of interest in human brain (Schölkopf et al., 2012; Peters et al., 2017). In this case, the process of generating an digit image can be described as follows: the writer first has a high level concept of interest in his mind (i.e., which digit to draw?), and then writes the digit through a complex sequence of biochemical reactions and sensorimotor movements. As such, when trying to predict its labels from an image, we are actually inverting the generating process, with the aim to predict cause (label) from effect (image), as shown in Fig. 6A. A supporting evidence is that in semi-supervised learning (i.e., an approach to machine learning that integrates a small amount of labeled data with a large amount of unlabeled data during training), an additional set of images usually contribute to the improvement of classification performance. Technically, given training points from $P(X, Y)$ and an additional set of inputs sampled from $P(X)$ where X are images and Y

labels, our goal is to estimate $P(Y|X)$. In light of $P(X,Y) = P(Y|X)P(X)$, let us consider two scenarios: (a) If $X \to Y$, by independence of the mechanism, $P(X)$ is independent of $P(Y|X)$. In other words, $P(X)$ contains no information about $P(Y|X)$, and therefore the additional sample from $P(X)$ will not give a better estimate of $P(Y|X)$. (b) If $Y \to X$, then $P(X)$ is no longer independent of $P(Y|X)$, that is, $P(X)$ provides some information about $P(Y|X)$. In this case, estimating $P(Y|X)$ will definitely benefit from the addition data sampled from $P(X)$. A large number of semi-supervised image classification experiments favor the latter scenario in which Y is a cause and X its effect.

7.2 Opinion 2: Causal

An opposite viewpoint was proposed in the concluding dialogue of Arjovsky et al. (2019). They claimed that digit recognition is a causal problem, i.e., predicting effect from cause, which attempts to predict human annotations Y from images X in order to imitate the cognitive process (i.e., humans produce labels by following a causal and cognitive process after observing images), as shown on the RHS of Fig. 6B. From this perspective, in supervised learning problems about predicting annotations, $P(Y|X)$ should be stable across environments or domains. Hence, the ubiquitous Empirical Risk Minimization principle (Vapnik, 1992) should work quite well in this setting and we should not worry too much about it. On the LHS of Fig. 6B, we can interpret the process from two directions. In the causal direction, Nature Variables (e.g., colour, light, angle, etc.) produce images through Nature causal mechanisms. In the anticausal direction, we attempt to disentangle the underlying causal factors (i.e., Nature Variables) of variation behind images. In the anticausal process, inference might be a more accurate term than disentanglement, because Nature Variables could be dependent on one another in some cases (e.g., the concept of predicates proposed in Vapnik (2019) can be viewed in some form of dependencies between Nature Variables, not Nature Variables alone, which will be discussed later).

7.3 Opinion 3: The Agnostic Hypothesis

Although it seems that both opinions above make sense to some extent, viewing either images or labels as the cause is not accurate. Technically, in the anticausal view, the learned mapping from images to labels is unstable because it is affected by the image distribution. In the causal view, viewing images as the cause has a conflict with various practical observations that the learned mapping from images to labels is not quite stable across domains and also can be easily fooled or attacked (Goodfellow et al., 2015) in a large number of real world applications.

According to the analysis above, we propose to look at the digit recognition problem from the agnostic viewpoint (i.e., absence of evidence is not evidence of absence). Precisely, we would like to believe in ***the Agnostic Hypothesis*** that there must be a third party of hidden causal factors (Z), namely *Nature Variables* (e.g., $\boldsymbol{X}_{p_r}$ in our case), affecting both images (X) and labels (Y). That is, both images and labels are effects of Nature Variables, as shown in Fig. 6C. It

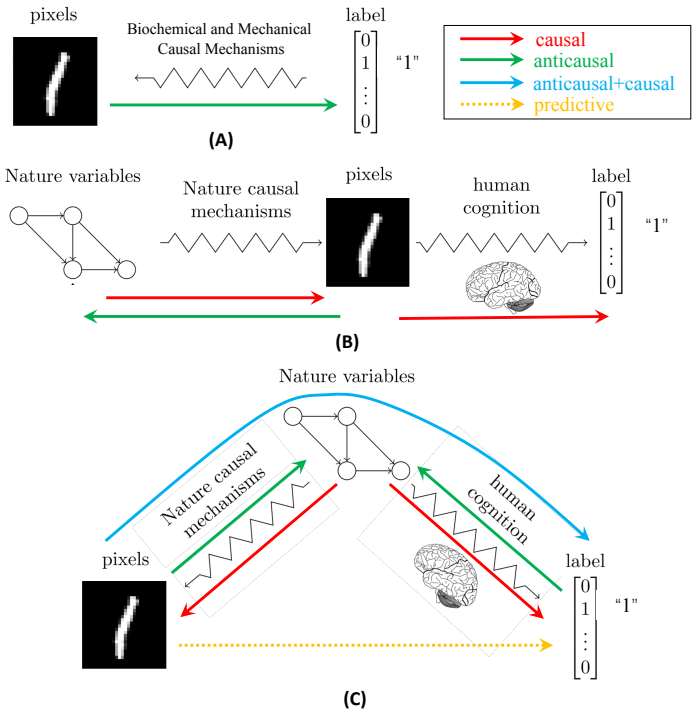


Figure 6: Three different opinions on digit recognition. (A) Digit recognition is an anticausal problem, i.e., predicting cause from effect, where images are thought of as the effect of labels. (B) Digit recognition is a causal problem, i.e., predicting effect form cause, where images are thought of as the cause of labels. (C) The Agnostic Hypothesis: there exist a third party of hidden variables (Nature Variables) affecting both images and labels. Note that, for a straightforward comparison, we accordingly modified the figure presented in Arjovsky et al. (2019) for each opinion.

is apparent to see that the Agnostic Hypothesis is so general as to include both causal and anticausal cases discussed above. Specifically, if the error in labelling is small (i.e., the error between Z and Y is small), then we can use the labels as a proxy for Nature Variables, in which case the Agnostic Hypothesis is reduced to the anticausal case. Similarly, when the error between Z and X is small, the Agnostic Hypothesis is reduced to the causal case. In general, however, these two extreme cases do not hold true, because even those abstract labels are incapable of accurately capturing the true meaning of Nature Variables due to the unsayable property in language (Agamben, 1999; Wittgenstein, 2009).

Historically, the Agnostic Hypothesis can be traced at least to the theories below.

- **Theory of Forms.** Plato's Theory of Forms (Edelstein, 1966) asserts

that there are two realms: the physical realm, which is a changing and imperfect world we see and interact with on a daily basis, and the realm of Forms, which exists beyond the physical realm and which is stable and perfect. This theory asserts that the physical realm is only a projection of the realm of Forms. Plato further claimed that each being has five things: the name, the definition, the image, the knowledge, and the Form, the first four of which can be viewed as the projections of the fifth. In this sense, the Agnostic Hypothesis is explicitly consistent with Plato's theory.

- **Manipulability Theory.** It provides a paradigmatic assertion in causal relationships that manipulation of a cause will result in the manipulation of an effect. In other words, causation implies that by varying one factor, we can make another vary (Cook and Campbell, 1979). In digit recognition, it makes sense that image causes label because changes in image will apparently lead to changes in label. Nevertheless, the converse that label causes image is also reasonable in that changes in label will definitely result in changes in images. Since it is clear that there is no causal loop, be it temporal or not, between images and labels, there must exist something hidden behind. Otherwise, it will violate the manipulability theory. Hence, the theory indirectly supports the Agnostic Hypothesis.

- **Principle of Common Cause.** The relation between association and causation was formalized in Reichenbach (1991): If two random variables X and Y are statistically dependent, then one of the following causal explanations must hold: a) X causes Y; b) Y causes X; c) there exists a random variable Z that is the common cause of both X and Y. This principle is directly in favor of the Agnostic Hypothesis. Because the previous analysis shows that there is no convincing evidence in full support of either image causes label or label causes image, there must be the third case in which there exists a hidden variable affecting both image and label.

Now let us investigate whether or not the Agnostic Hypothesis can give a better explanation.

Firstly, we can think of image and label as two different representation spaces projected from Nature Variables via Nature causal mechanism and human cognitive mechanism, respectively. In other words, image is the way Nature interprets Nature Variables whilst label is the way humans interpret them. Hence, both are the effects of Nature Variables, but on different levels (i.e., label is more abstract than image).

Secondly, in light of the fork junction (i.e., $X \leftarrow Z \rightarrow Y$), we can claim two things: (1) Given Nature Variables, image is independent of label; (2) If Nature Variable is hidden (i.e., not given), image is NOT independent of label. Both indeed make sense. The former states that image can be perfectly explained by Nature Variables without the help of label, vice versa (see the red arrows in Fig. 6C). The latter explains why it is meaningful to predict label from image, and actually the reason holds in most supervised learning applications (see the yellow dashed arrow in Fig. 6C).

Thirdly, almost all the existing approaches to digit recognition, or more general image classification, consist of two parts: a feature extractor and a classifier whether explicitly or implicitly. For example, traditional approaches are usually explicitly made up of a handcrafted feature extractor and a well-chosen classifier. By contrast, deep learning approaches appear more implicit, where deep neural networks can always be split into two components: the first of which can be viewed as a feature extractor and the second as a classifier. In effect, this setting can be well explained by the blue arrow in Fig. 6C that is composed of an anticausal process from image to Nature Variables and a causal process from Nature Variables to label. We can think of the anticausal part as a feature extractor and the causal part as a classifier.

Inspired by the points above, we argue that Nature Variables are the key to the OOD generalization in invariant risk minimization. Here we conceptually explain why this argument universally makes sense. Without loss of generality, we continue taking digit/image classification for example. Ideally, if we have a perfect feature extractor, then we can infer Nature Variables from images, which will, to the largest extent, reduce the negative influence from irrelevant noisy information on the input. In other words, the perfect feature extractor can defeat the attack on the input so as to guarantee the OOD generalization of the systems in terms of the input information. Furthermore, because Nature Variables are causal parents of labels, the learned classifier based on Nature Variables should be invariant across environments or domains as discussed in Arjovsky et al. (2019) and Peters et al. (2016), which guarantees the OOD generalization of the systems in terms of the output information. More importantly, once Nature Variables are obtained, we would know how they influence one another and how they affect both images and labels. It would thus render the systems' behaviours more interpretable.

7.4 A Unifying View of Machine Learning

All the explanations above are not limited to digit recognition and also apply to general machine learning problems. In fact, the Agnostic Hypothesis provides a unifying view of machine learning as shown in Fig. 7, which paves the way for inspiring both new algorithm designs and a new theory of machine learning. In this section, we first re-interpret machine learning, which are widely categorised as supervised, unsupervised, or reinforcement learning, under the Agnostic Hypothesis. Then, we discuss the implications of such a unifying view and its practical consequences, which are substantiated by the experimental results found in the literature.

7.4.1 Supervised Learning

When the training data contains examples of what the correct output should be for given inputs, as aforementioned, under the Agnostic Hypothesis the inputs and the correct outputs can be thought of as two different representation spaces

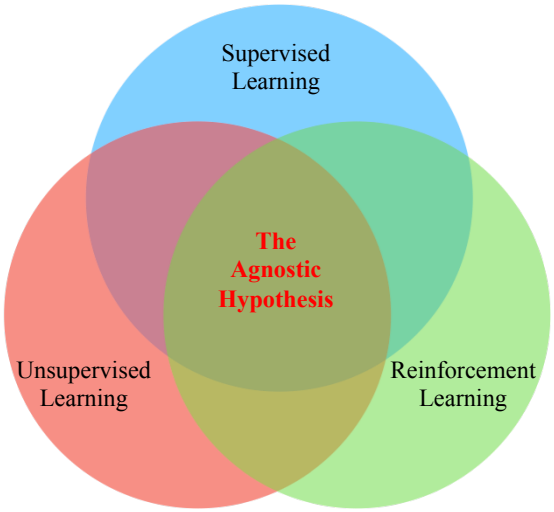


Figure 7: The Agnostic Hypothesis provides a unifying view of machine learning.

projected from Nature Variables via two different mechanisms. In other words, inputs and outputs are two distinct interpretations or views of Nature Variables.

7.4.2 Unsupervised Learning

When the training data just contains inputs without any output information, there are several scenarios. If the input data come only from one domain, under the Agnostic Hypothesis they are interpreted as just one projection or view of Nature Variables. If they are from multiple transforms or variants of the same domain (Chen et al., 2020), they are naturally viewed as multiple projections or views of Nature Variables.

7.4.3 Reinforcement Learning

A reinforcement learning agent is learning what to do, i.e., how to map situations to actions, so as to maximise a numerical reward signal (Sutton and Barto, 2018). Rather than being directly told which action to take, the agent must learn from sequences of observed information (i.e., states/observations in a fully/partially observable environment, actions, and rewards). Under the Agnostic Hypothesis, each component of the observed information reflects one aspect of the environment and can be thus viewed as one projection or view of Nature Variables [11].

[11] Although uncovering structure in an agent's experience itself, whose perfect representation should be undoubtedly in some form of Nature Variables, does not address the reinforcement learning problem of maximising a reward signal, it plays a vital role in real world reinforcement learning applications with strong demand for the OOD generalization (Garcıa and Fernández, 2015; Gros et al., 2020).

7.4.4 Implications of a Unifying View

The reason why we should look at machine learning from the unifying view of the Agnostic Hypothesis is that compared to the traditional view, it offers a promising way to explore the requirements of machine learning for the OOD generalization in the real world applications. As stated in Russell and Norvig (2002), the representation of the learned information plays a pivotal role in determining how the learning algorithm must work. Hence, from the unifying view, all the required properties of machine learning are rooted in how accurately we can estimate Nature Variables from data, because Nature Variables are, without doubt, the optimal representation of the data.

Now the question comes to whether or not it is possible to identify Nature Variables from data. Considering that under the Agnostic Hypothesis the data in reality are viewed as projections or views of Nature Variables, one naturally has an intuition that the more diverse views of Nature Variables we have, the more accurate estimate of Nature Variables we can attain. For convenience, we call this intuition ***Multiview Universal Approximation*** (**MUA**). Not surprisingly, some theoretical works have demonstrated MUA to some extent by showing that under some assumptions multiple views will lead to identifiability of Nature Variables up to some affine ambiguity (Gresele et al., 2019). Actually, under some stronger assumptions only two views are even capable of identifying Nature Variables up to some unavoidable indeterminacy (Hyvarinen et al., 2019; Gresele et al., 2019). Although these initial works are predicated on strict assumptions on Nature Variables (e.g., Nature Variables are assumed to be independent or conditionally independent of each other, which, as aforementioned, is unnecessary in reality, etc.), they are a good starting point on the road towards the general theory of MUA.

From this point of view, the unifying view based on the Agnostic Hypothesis has at least four practical implications.

a) It can help identify the issues in current machine learning algorithms. For example, in the research of adversarial attacks (Goodfellow et al., 2015), the reason why it is far too easy to fool convolutional networks with an imperceivable but carefully constructed noise in the input is that the feature extractor part of the networks cannot accurately infer the Nature Variables in the anticausal direction. Hence, the learned predictive link between image and label is so unstable that a small disturbance on the input image will lead to wrong Nature Variables misguiding the classifier part. This issue widely exists in supervised learning, because at the testing time only one view is used to infer Nature Variables. This issue could be mitigated if the input data involve multiple views in some scenarios, such as in time series prediction problems that take as input multiple time step data and each time step can be viewed as one view of Nature Variables.

b) It can help understand why algorithms work. For instance, it is reasonable that multitask learning (Caruana, 1997) has been used successfully across all applications of machine learning, because multitask data provide multiple views of their shared features (Nature Variables), making inferring them more accurate

as suggested in MUA. Another example is that in reinforcement learning, one widely leveraged multiview data to discover the invariant part of states (Lu et al., 2018; Zhang et al., 2020).

c) It can help inspire a new algorithm design. It is worth noting that an unsupervised learning approach, proposed in a very recent work (Chen et al., 2020), leverages data augmentation to considerably outperform previous methods for self-supervised and semi-supervised learning and even to be on a par with supervised learning methods on ImageNet. It indeed makes sense, because data augmentation created multiple views of latent features (Nature Variables), leading to identifying Nature Variables more accurately as stated in MUA.

d) It can help inspire a new theory of machine learning. As mentioned previously, current machine learning theories neither satisfy the demands for the OOD generalization in real world applications nor answer the key question of how to discover Nature Variables from data (Vapnik, 2019). The unifying view provides a promising way to address both issues by developing a general theory of MUA with more relaxed assumptions, because under the Agnostic Hypothesis MUA is a natural and feasible way to identify Nature Variables as mentioned previously. Another possible thread of discovering them is through intervention (Pearl, 2009) if it is allowed to interact with environments.

7.5 Broader Discussion

The Agnostic Hypothesis can be viewed as a kind of description of the relationship between invariants and variants, where invariants are reflected by Nature Variables and variants by their corresponding projections or views. In this sense, the Agnostic Hypothesis has a long history in philosophy and science. It can be dated back to the time of Plato, whose Theory of Forms described the relationship as aforementioned. Hegel argued that there exists a higher level of cognition commonly taken as capable of having purportedly eternal contents (i.e., invariants) which come from the changing contents (i.e., variants) based in everyday perceptual experience. Furthermore, Wigner summarised physics as a process of discovering the laws of inanimate nature, that is, recognising invariance from the world of baffling complexity around us (Wigner, 1990). Recently, Vapnik was motivated to propose a statistical theory of learning based on statistical invariants constructed using training data and given predicates (Vapnik, 2019). Roughly speaking, a predicate is a function revealing some invariant property of the world of interest, like the Form in Plato's theory.

Although a number of philosophers agree that there exist such invariants beyond the variants, there is no consensus on whether or not they are apprehensible. For example, Kant thought that there is some unknowable invariant, called thing-in-itself, outside of all possible human experience (Kant, 1998). However, Schopenhauer believed that the supreme invariant principle of the universe is likewise apprehensible through introspection, and that we can understand the world as various manifestations of this general principle (Schopenhauer, 2012). Despite the controversy in philosophy, Vapnik still provided two examples in art to show that it might be possible to comprehend those invariants to some ex-

tent (Vapnik, 2020). One is that Bach's music is full of repeated patterns. The other is that Vladimir Propp, a Soviet formalist scholar, analyzed the basic plot components of Russian folk tales to identify 31 simplest irreducible narrative elements which are so general that they also apply to many other stories and movies. Both seemingly demonstrate that there exist such invariants at least in some form. It is thus natural to ask how we can identify them from data, which is so important that Vapnik thought that the essence of intelligence is the discovery of good predicates (Vapnik, 2020). Schmidhuber once expressed a similar opinion that "all the history of science is the history of compression progress", where obviously the optimal compression should be in the form of Nature Variables (Schmidhuber, 2018). Bengio proposed a "consciousness prior" for learning representations of high-level concepts (Bengio, 2017). We hope that the Agnostic Hypothesis can provide inspiration to explore the general theory of MUA for identifying Nature Variables both theoretically and practically, which is key to enabling OOD generalization guarantees in machine learning.

References

Agamben, G. (1999). *Potentialities: Collected essays in philosophy*. Stanford University Press.

Ahuja, K., Shanmugam, K., Varshney, K., and Dhurandhar, A. (2020). Invariant risk minimization games. *arXiv preprint arXiv:2002.04692*.

Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, 41(1):15–34.

Arandjelovic, R. and Zisserman, A. (2017). Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. (2007). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137.

Bengio, Y. (2017). The consciousness prior. *arXiv preprint arXiv:1709.08568*.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

Cook, T. and Campbell, D. (1979). Experimental and quasi-experimental designs for research. *Chicago und IL: Rand McNally*.

Dawid, A. P., Didelez, V., et al. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231.

Edelstein, L. (1966). *Plato's seventh letter*, volume 14. Brill.

Fang, C., Xu, Y., and Rockmore, D. N. (2013). Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, pages 1657–1664. IEEE Computer Society.

Garcıa, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. (2015). Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, pages 2551–2559. IEEE Computer Society.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations*.

Gresele, L., Rubenstein, P., Mehrjou, A., Locatello, F., and Schölkopf, B. (2019). The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *35th Conference on Uncertainty in Artificial Intelligence (UAI 2019)*, pages 296–313. Curran.

Gros, S., Zanon, M., and Bemporad, A. (2020). Safe reinforcement learning via projection on a safe set: How to achieve optimality? *IFAC*.

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

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115.

Heinze-Deml, C., Peters, J., and Meinshausen, N. (2018). Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2).

Hoover, K. D. (1990). The logic of causal inference: Econometrics and the conditional analysis of causation. *Economics & Philosophy*, 6(2):207–234.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53.

Hyvarinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *arXiv preprint arXiv:1605.06336*.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868.

Kant, I. (1998). *Critique of Pure Reason*. Cambridge University Press.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. (2012). Undoing the damage of dataset bias. In *ECCV (1)*, volume 7572 of *Lecture Notes in Computer Science*, pages 158–171. Springer.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, H., Jialin Pan, S., Wang, S., and Kot, A. C. (2018a). Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. (2018b). Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124.

Lu, C. (2020a). The agnostic hypothesis: A unifying view of machine learning. Blogpost at causallu.com. https://causallu.com/2020/10/24/the-agnostic-hypothesis-a-unifying-view-of-machine-learning/.

Lu, C. (2020b). Is image classification a causal problem? Blog-post at causallu.com. https://causallu.com/2020/04/05/is-image-classification-a-causal-problem/.

Lu, C., Schölkopf, B., and Hernández-Lobato, J. M. (2018). Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. (2017). Unified deep supervised domain adaptation and generalization. In *ICCV*, pages 5716–5726. IEEE Computer Society.

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *ICML (1)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 10–18. JMLR.org.

Nagarajan, V., Andreassen, A., and Neyshabur, B. (2020). Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peters, J., Bühlmann, P., and Meinshausen, N. (2015). Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference*. The MIT Press.

Reichenbach, H. (1991). *The direction of time*, volume 65. Univ of California Press.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018). Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34.

Rosenfeld, E., Ravikumar, P., and Risteski, A. (2020). The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*.

Russell, S. and Norvig, P. (2002). *Artificial intelligence: a modern approach*.

Schmidhuber, J. (2018). Godel machines, meta-learning, and lstms. https://youtu.be/3FIo6evmweo.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. (2012). On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262.

Schopenhauer, A. (2012). *The world as will and representation*, volume 1. Courier Corporation.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017). Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1):1830–1888.

Subbaswamy, A., Chen, B., and Saria, S. (2019). A universal hierarchy of shift-stable distributions and the tradeoff between stability and performance. *arXiv preprint arXiv:1905.11374*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, NY.

Vapnik, V. (2020). Predicates, invariants, and the essence of intelligence. https://youtu.be/bQa7hpUpMzM.

Vapnik, V. N. (2019). Complete statistical theory of learning. *Automation and Remote Control*, 80(11):1949–1975.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027.

von Kügelgen, J., Mey, A., Loog, M., and Schölkopf, B. (2020). Semi-supervised learning, causality and the conditional cluster assumption. *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Wigner, E. P. (1990). The unreasonable effectiveness of mathematics in the natural sciences. In *Mathematics and Science*, pages 291–306. World Scientific.

Wittgenstein, L. (2009). *Philosophical investigations*. John Wiley & Sons.

Wright, S. (1921). Correlation and causation. *J. agric. Res.*, 20:557–580.

Zhang, A., Lyle, C., Sodhani, S., Filos, A., Kwiatkowska, M., Pineau, J., Gal, Y., and Precup, D. (2020). Invariant causal prediction for block mdps. *arXiv preprint arXiv:2003.06016*.

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, page 1347. NIH Public Access.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827.

A Variational Autoencoders

We briefly describe the identifiable variational autoencoders (iVAEs) proposed by Khemakhem et al. (2020). As we know, the framework of variational autoencoders (VAEs) (Kingma and Welling, 2013; Rezende et al., 2014) allows us to efficiently learn deep latent-variable models and their corresponding inference models. Consider a simple latent variable model where $\boldsymbol{O} \in \mathbb{R}^d$ stands for an observed variable (random vector) and $\boldsymbol{X} \in \mathbb{R}^n$ for a latent variable. The VAE model actually learns a full generative model $p_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{X}) = p_{\boldsymbol{\theta}}(\boldsymbol{O}|\boldsymbol{X})p_{\boldsymbol{\theta}}(\boldsymbol{X})$ and an inference model $q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O})$ that approximates its posterior $p_{\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{O})$, where $\boldsymbol{\theta}$ is a vector of parameters of the generative model, $\boldsymbol{\phi}$ a vector of parameters of the inference model, and $p_{\boldsymbol{\theta}}(\boldsymbol{X})$ is a prior distribution over the latent variables. Instead of maximizing the data log-likelihood, we maximize its lower bound $\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{O}) \geq \mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}) := \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{O}|\boldsymbol{X})\right] - \text{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O})||p_{\boldsymbol{\theta}}(\boldsymbol{X})\right),$$

where we have used Jensen's inequality, and $\text{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence between two distributions.

B Definitions

For convenience, we restates some definitions here and please refer to the original papers (Arjovsky et al., 2019; Peters et al., 2017) for more details.

Definition 3. *A structural equation model (SEM) $\mathcal{C} := (\mathcal{S}, N)$ governing the random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ is a set of structural equations:*

$$\mathcal{S}_i : X_i \leftarrow f_i(\text{Pa}(X_i), N_i),$$

where $\text{Pa}(X_i) \subseteq \{X_1, \ldots, X_d\} \setminus \{X_i\}$ are called the parents of X_i, and the N_i are independent noise random variables. We say that "X_i causes X_j" if $X_i \in \text{Pa}(X_j)$. We call causal graph of $\boldsymbol{X}$ to the graph obtained by drawing i) one node for each X_i, and ii) one edge from X_i to X_j if $X_i \in \text{Pa}(X_j)$. We assume acyclic causal graphs.

Definition 4. *Consider a SEM $\mathcal{C} := (\mathcal{S}, N)$. An intervention e on $\mathcal{C}$ consists of replacing one or several of its structural equations to obtain an intervened SEM $\mathcal{C}^e := (\mathcal{S}^e, N^e)$, with structural equations:*

$$\mathcal{S}_i^e : X_i^e \leftarrow f_i^e(\text{Pa}^e(X_i^e), N_i^e),$$

The variable $\boldsymbol{X}^e$ is intervened if $\mathcal{S}_i \neq \mathcal{S}_i^e$ or $N_i \neq N_i^e$.

Definition 5. *Consider a structural equation model (SEM) $\mathcal{S}$ governing the random vector $(X_1, \ldots, X_n, \boldsymbol{Y})$, and the learning goal of predicting $\boldsymbol{Y}$ from $\boldsymbol{X}$. Then, the set of all environments $\mathcal{E}_{all}(\mathcal{S})$ indexes all the interventional distributions $P(\boldsymbol{X}^e, \boldsymbol{Y}^e)$ obtainable by valid interventions e. An intervention $e \in \mathcal{E}_{all}(\mathcal{S})$ is valid as long as (i) the causal graph remains acyclic, (ii) $\mathbb{E}\left[\boldsymbol{Y}^e|\text{Pa}(\boldsymbol{Y})\right] = \mathbb{E}\left[\boldsymbol{Y}|\text{Pa}(\boldsymbol{Y})\right]$, and (iii) $\mathbb{V}\left[\boldsymbol{Y}^e|\text{Pa}(\boldsymbol{Y})\right]$ remains within a finite range.*

C Derivation

In Phase 1, the lower bound is defined by

$$\begin{aligned}\mathcal{L}_{\text{phase1}}(\boldsymbol{\theta},\boldsymbol{\phi}) :=& \mathbb{E}_{p_D}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{O},\boldsymbol{X}|\boldsymbol{Y},E) - \log q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)\right]\right] \\ =& \mathbb{E}_{p_D}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\log p_{\boldsymbol{f}}(\boldsymbol{O}|\boldsymbol{X}) + \log p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{X}|\boldsymbol{Y},E)\right.\right. \\ & \left.\left. - \log q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)\right]\right] \\ =& \mathbb{E}_{p_D}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\log p_{\boldsymbol{f}}(\boldsymbol{O}|\boldsymbol{X})\right]\right. \\ & + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\log p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{X}|\boldsymbol{Y},E)\right] \\ & \left. - \mathbb{E}_{q_{\phi}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\log q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)\right]\right].\end{aligned}$$

The first term is the data log-likelihood and the third term has a closed-form solution,

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\log q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)\right] = -\frac{J}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{J}\left(1+\log\sigma_j^2\right).$$

where σ_j is simply denote the j-th element of the variational s.d. (σ) evaluated at datapoint i that is simply a function of $(\boldsymbol{O},\boldsymbol{Y},E)$ and the variational parameters ϕ.

Now let us look at the second term,

$$\begin{aligned}& \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\log p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{X}|\boldsymbol{Y},E)\right] \\ =& \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\log\left(\prod_i \frac{\mathcal{Q}_i(\boldsymbol{X}_i)}{\mathcal{Z}_i(\boldsymbol{Y},E)}\exp\left[\sum_{j=1}^{k} T_{i,j}(\boldsymbol{X}_i)\lambda_{i,j}(\boldsymbol{Y},E)\right]\right)\right] \\ =& \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\sum_i \log\left(\frac{\mathcal{Q}_i(\boldsymbol{X}_i)}{\mathcal{Z}_i(\boldsymbol{Y},E)}\exp\left[\sum_{j=1}^{k} T_{i,j}(\boldsymbol{X}_i)\lambda_{i,j}(\boldsymbol{Y},E)\right]\right)\right] \\ \propto& \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)}\left[\sum_i\sum_{j=1}^{k} T_{i,j}(\boldsymbol{X}_i)\lambda_{i,j}(\boldsymbol{Y},E)\right] \\ \approx& \frac{1}{L}\sum_l\sum_i\sum_{j=1}^{k} T_{i,j}(\boldsymbol{X}_i^l)\lambda_{i,j}(\boldsymbol{Y},E^l),\end{aligned}$$

where we let the base measure $\mathcal{Q}_i(\boldsymbol{X}_i) = 1$ and L is the sample size.

D Theorems

Theorem 4. *Assume that we observe data sampled from a generative model defined according to Eqs. (5-7), with parameters $\boldsymbol{\theta} = (\boldsymbol{f},\boldsymbol{T},\boldsymbol{\lambda})$ and $k = 1$. Assume the following holds: (i) The set $\{\boldsymbol{O} \in \mathcal{O}|\varphi_{\boldsymbol{\epsilon}}(\boldsymbol{O}) = 0\}$ has measure*

zero, where $\varphi_{\boldsymbol{\epsilon}}$ is the characteristic function of the density $p_{\boldsymbol{\epsilon}}$ defined in Eq. (6). (ii) The mixing function $\boldsymbol{f}$ in Eq. (6) is injective, and all partial derivatives of $\boldsymbol{f}$ are continuous. (iii) The sufficient statistics $T_{i,j}$ in Eq. (7) are differentiable almost everywhere and not monotonic, and $(T_{i,j})_{1\leq j\leq k}$ are linearly independent on any subset of $\mathcal{X}$ of measure greater than zero. (iv) There exist $nk+1$ distinct points $(\boldsymbol{Y},E)^0,\ldots,(\boldsymbol{Y},E)^{nk}$ such that the matrix $L = \big(\boldsymbol{\lambda}((\boldsymbol{Y},E)^1)-\boldsymbol{\lambda}((\boldsymbol{Y},E)^0),\ldots,\boldsymbol{\lambda}((\boldsymbol{Y},E)^{nk})-\boldsymbol{\lambda}((\boldsymbol{Y},E)^0)\big)$ of size $nk\times nk$ is invertible. Then the parameters $\boldsymbol{\theta}=(\boldsymbol{f},\boldsymbol{T},\boldsymbol{\lambda})$ are identifiable up to a permutation and pointwise transformation.

E Proofs

E.1 Proof of Theorem 2

This proof is similar to that of Theorem 4 in Khemakhem et al. (2020)

Proof. The loss function in Eq. 3 can be rephrased as follows:

$$\mathcal{L}_{\text{phase1}}(\boldsymbol{\theta},\boldsymbol{\phi}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{O}|\boldsymbol{Y},E) - KL\left(q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)||p_{\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)\right).$$

If the family of $q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)$ is flexible enough to contain $p_{\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{Y},E)$, then by optimizing the loss over its parameter $\boldsymbol{\phi}$, we will minimize the KL term which will eventually reach zero, and the loss will be equal to the log-likelihood. Under this circumstance, the iVAE inherits all the properties of maximum likelihood estimation. In this particular case, since our identifiability is guaranteed up to a permutation and pointwise transformation, the consistency of MLE means that we converge to the true parameter $\boldsymbol{\theta}^*$ up to a permutation and pointwise transformation in the limit of infinite data. Because true identifiability is one of the assumptions for MLE consistency, replacing it by identifiability up to a permutation and pointwise transformation does not change the proof but only the conclusion. □

E.2 Proof of Theorem 3

Proof. Theorem 1 and Theorem 2 guarantee that in the limit of infinite data, iVAE can learn the true parameters $\boldsymbol{\theta}^* := (\boldsymbol{f}^*,\boldsymbol{T}^*,\boldsymbol{\lambda}^*)$ up to a permutation and pointwise transformation. Let $(\hat{\boldsymbol{f}},\hat{\boldsymbol{T}},\hat{\boldsymbol{\lambda}})$ be the parameters obtained by iVAE, and we therefore have $(\hat{\boldsymbol{f}},\hat{\boldsymbol{T}},\hat{\boldsymbol{\lambda}}) \sim_P (\boldsymbol{f}^*,\boldsymbol{T}^*,\boldsymbol{\lambda}^*)$, where $\sim_P$ denotes the equivalence up to a permutation and pointwise transformation. If there were no noise, this would mean that the learned $\hat{\boldsymbol{f}}$ transforms $\boldsymbol{O}$ into $\hat{\boldsymbol{X}} = \hat{\boldsymbol{f}}^{-1}(\boldsymbol{O})$ that are equal to $\boldsymbol{X}^* = (\boldsymbol{f}^*)^{-1}(\boldsymbol{O})$ up to a permutation and sighed scaling. If with noise, we obtain the posteriors of the latents up to an analogous indeterminacy. □

E.3 Proof of Proposition 1

Proof. The following rules can be independently performed to distinguish all the 12 possible structures shown in Figs. 2b-2k. For clarity, we divide them into three groups.

Group 1 All the six structures in this group can be discovered only by performing conditional independence tests.

- **Rule 1.1** If $\boldsymbol{X}_i \perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E$, and $E \perp\!\!\!\perp \boldsymbol{Y}$, then Fig. 2d is discovered.
- **Rule 1.2** If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \perp\!\!\!\perp E$, and $E \not\perp\!\!\!\perp \boldsymbol{Y}$, then Fig. 2g is discovered.
- **Rule 1.3** If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E$, and $E \perp\!\!\!\perp \boldsymbol{Y}$, then Fig. 2f is discovered.
- **Rule 1.4** If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E$, $E \not\perp\!\!\!\perp \boldsymbol{Y}$, and $\boldsymbol{X}_i \perp\!\!\!\perp \boldsymbol{Y}|E$, then Fig. 2i is discovered.
- **Rule 1.5** If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E$, $E \not\perp\!\!\!\perp \boldsymbol{Y}$, and $\boldsymbol{X}_i \perp\!\!\!\perp E|\boldsymbol{Y}$, then Fig. 2h is discovered.
- **Rule 1.6** If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E$, $E \not\perp\!\!\!\perp \boldsymbol{Y}$, and $\boldsymbol{Y} \perp\!\!\!\perp E|\boldsymbol{X}_i$, then Fig. 2e is discovered.

Group 2 If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \perp\!\!\!\perp E$, and $E \perp\!\!\!\perp \boldsymbol{Y}$, then we can discover both Fig. 2b and Fig. 2c. These two structures cannot be further distinguished only by conditional independence tests, because they come from the same Markov equivalence class. Fortunately, we can further distinguish them by running binary causal discovery algorithms (Peters et al., 2017), e.g., ANM (Hoyer et al., 2009) or the bivariate fit model that is based on a best-fit criterion relying on a Gaussian Process regressor.

- **Rule 2.1** If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \perp\!\!\!\perp E$, and $E \perp\!\!\!\perp \boldsymbol{Y}$, and a chosen binary causal discovery algorithm prefers $\boldsymbol{X}_i \to \boldsymbol{Y}$ to $\boldsymbol{X}_i \leftarrow \boldsymbol{Y}$, then Fig. 2b is discovered.
- **Rule 2.2** If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \perp\!\!\!\perp E$, and $E \perp\!\!\!\perp \boldsymbol{Y}$, and a chosen binary causal discovery algorithm prefers $\boldsymbol{X}_i \leftarrow \boldsymbol{Y}$ to $\boldsymbol{X}_i \to \boldsymbol{Y}$, then Fig. 2c is discovered.

Group 3 If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E$, $E \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}|E$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E|\boldsymbol{Y}$, and $\boldsymbol{Y} \not\perp\!\!\!\perp E|\boldsymbol{X}_i$, then we can discover both Fig. 2j and Fig. 2k. These two structures cannot be further distinguished only by conditional independence tests, because they come from the same Markov equivalence class. They also cannot be distinguished by any binary causal discovery algorithm, since both $\boldsymbol{X}_i$ and $\boldsymbol{Y}$ are affected by E. Fortunately, Zhang et al. (2017) provided a heuristic solution to this issue based on the invariance of causal mechanisms, i.e., $P(\text{cause})$ and $P(\text{effect}|\text{cause})$ change independently. The detailed description of their method

is given in Section 4.2 of Zhang et al. (2017). For convenience, here we directly borrow their final result. Zhang et al. (2017) states that determining the causal direction between $\boldsymbol{X}_i$ and $\boldsymbol{Y}$ in Fig. 2j and Fig. 2k is finally reduced to calculating the following term:

$$\Delta_{\boldsymbol{X}_i \to \boldsymbol{Y}} = \left\langle \log \frac{\bar{P}(\boldsymbol{Y}|\boldsymbol{X}_i)}{\langle \hat{P}(\boldsymbol{Y}|\boldsymbol{X}_i) \rangle} \right\rangle, \tag{12}$$

where $\langle \cdot \rangle$ denotes the sample average, $\bar{P}(\boldsymbol{Y}|\boldsymbol{X}_i)$ is the empirical estimate of $P(\boldsymbol{Y}|\boldsymbol{X}_i)$ on all data points, and $\langle \hat{P}(\boldsymbol{Y}|\boldsymbol{X}_i) \rangle$ denotes the sample average of $\hat{P}(\boldsymbol{Y}|\boldsymbol{X}_i)$, which is the estimate of $P(\boldsymbol{Y}|\boldsymbol{X}_i)$ in each environment. We take the direction for which Δ is smaller to be the causal direction.

- **Rule 3.1** If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E$, $E \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}|E$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E|\boldsymbol{Y}$, $\boldsymbol{Y} \not\perp\!\!\!\perp E|\boldsymbol{X}_i$, and $\Delta_{\boldsymbol{X}_i \to \boldsymbol{Y}}$ is smaller than $\Delta_{\boldsymbol{Y} \to \boldsymbol{X}_i}$, then Fig. 2j is discovered.
- **Rule 3.2** If $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E$, $E \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{X}_i \not\perp\!\!\!\perp \boldsymbol{Y}|E$, $\boldsymbol{X}_i \not\perp\!\!\!\perp E|\boldsymbol{Y}$, $\boldsymbol{Y} \not\perp\!\!\!\perp E|\boldsymbol{X}_i$, and $\Delta_{\boldsymbol{Y} \to \boldsymbol{X}_i}$ is smaller than $\Delta_{\boldsymbol{X}_i \to \boldsymbol{Y}}$, then Fig. 2k is discovered.

□

E.4 Proof of Proposition 2

Proof. Firstly, assumption (iii) and assumption (iv) in Theorem 1 are the requirements of the set of training environments containing sufficient diversity and satisfying an underlying invariance which holds across all the environments. Interestingly, assumption (iii) elicits Lemma 4 of Khemakhem et al. (2020), which is closely similar to the *linear general position* in Assumption 8 of Arjovsky et al. (2019). Thus, Lemma 4 can be similarly called the *nonlinear general position* in our generalization theory, whose proof can be found in Arjovsky et al. (2019). Secondly, when the set of training environments lie in this nonlinear general position and the other hypotheses of Theorem 1&2 hold, it is guaranteed in Theorem 3 that all the latent factors $\boldsymbol{X}$ can be identified up to a permutation and pointwise transformation. Since this identifiability result holds under the assumptions guaranteeing that training environments contain sufficient diversity and satisfy an underlying invariance which holds across all the environments, it also holds across all the environments. Thirdly, Proposition 1 suggests that all the direct causes $\mathrm{Pa}(\boldsymbol{Y})$ of $\boldsymbol{Y}$ can be fully discovered, which also holds across all the environments due to the same reason above. Finally, the challenging bi-leveled optimization problem in both IRM and IRMG now can be reduced to two simpler independent optimization problems: (i) learning the invariant data representation Φ from $\boldsymbol{O}$ to $\mathrm{Pa}(\boldsymbol{Y})$, and (ii) learning the invariant classifier w from $\mathrm{Pa}(\boldsymbol{Y})$ to $\boldsymbol{Y}$, as described in Eq. (9) and Eq. (10). For both (i) and (ii), since there exist no spurious correlations between $\boldsymbol{O}$ and $\mathrm{Pa}(\boldsymbol{Y})$ and between $\mathrm{Pa}(\boldsymbol{Y})$ and $\boldsymbol{Y}$, learning theory guarantees that in the limit of infinite data, we will converge to the true invariant data representation Φ and the true invariant classifier w. □

It is worth noting that although assumption (iii) and assumption (iv) in Theorem 1 require complicated conditions to satisfy the diversity across training environments for generalization guarantees, it is not the case in practice. As we will observe in our experiments, it is often the case that two environments are sufficient to recover invariances.

F Datasets

For convenience and completeness, we provide the descriptions of Colored MNIST Digits and Colored Fashion MNIST here. Please refer to the original papers (Arjovsky et al., 2019; Ahuja et al., 2020; Gulrajani and Lopez-Paz, 2020; Venkateswara et al., 2017) for more details.

F.1 Synthetic Data

For the nonlinear transformation, we use the MLP:

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Output layer: Fully connected layer, output size = 10

F.2 Colored MNIST Digits

We use the exact same environment as in Arjovsky et al. (2019). Arjovsky et al. (2019) propose to create an environment for training to classify digits in MNIST digits data[12], where the images in MNIST are now colored in such a way that the colors spuriously correlate with the labels. The task is to classify whether the digit is less than 5 (not including 5) or more than 5. There are three environments (two training containing 30,000 points each, one test containing 10,000 points) We add noise to the preliminary label ($\tilde{y} = 0$ if digit is between 0-4 and $\tilde{y} = 0$ if the digit is between 5-9) by flipping it with 25 percent probability to construct the final labels. We sample the color id z by flipping the final labels with probability p_e, where p_e is 0.2 in the first environment, 0.1 in the second environment, and 0.9 in the third environment. The third environment is the testing environment. We color the digit red if $z = 1$ or green if $z = 0$.

F.3 Colored Fashion MNIST

We modify the fashion MNIST dataset[13] in a manner similar to the MNIST digits dataset. Fashion MNIST data has images from different categories: "t-shirt", "trouser", "pullover", "dress", "coat", "sandal", "shirt", "sneaker", "bag",

[12] https://www.tensorflow.org/api_docs/python/tf/keras/datasets/mnist/load_data

[13] https://www.tensorflow.org/api_docs/python/tf/keras/datasets/fashion_mnist/load_data

"ankle boots". We add colors to the images in such a way that the colors correlate with the labels. The task is to classify whether the image is that of foot wear or a clothing item. There are three environments (two training, one test) We add noise to the preliminary label ($\tilde{y} = 0$: "t-shirt", "trouser", "pullover", "dress", "coat", "shirt" and $\tilde{y} = 1$: "sandal", "sneaker", "ankle boots") by flipping it with 25 percent probability to construct the final label. We sample the color id z by flipping the noisy label with probability p_e, where p_e is 0.2 in the first environment, 0.1 in the second environment, and 0.9 in the third environment, which is the test environment. We color the object red if $z = 1$ or green if $z = 0$.

G Model Architectures

In this section, we describe the architectures of different models used in different experiments.

G.1 Synthetic Data

G.1.1 ERM

Linear ERM

- Input layer: Input batch *(batch size, input dimension)*
- Output layer: Fully connected layer, output size = 1

Nonlinear ERM

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Output layer: Fully connected layer, output size = 1

G.1.2 IRM

Linear Data Representation Φ

- Input layer: Input batch *(batch size, input dimension)*
- Output layer: Fully connected layer, output size = 1

Nonlinear Data Representation Φ

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Output layer: Fully connected layer, output size = 1

G.1.3 F-IRM GAME

Linear Classifier w

- Input layer: Input batch *(batch size, input dimension)*
- Output layer: Fully connected layer, output size = 1

Nonlinear Classifier w

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Output layer: Fully connected layer, output size = 1

G.1.4 V-IRM GAME

Linear Data Representation Φ

- Input layer: Input batch *(batch size, input dimension)*
- Output layer: Fully connected layer, output size = 2

Nonlinear Data Representation Φ

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Output layer: Fully connected layer, output size = 2

Linear Classifier w

- Input layer: Input batch *(batch size, 2)*
- Output layer: Fully connected layer, output size = 1

Nonlinear Classifier w

- Input layer: Input batch *(batch size, 2)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Output layer: Fully connected layer, output size = 1

G.1.5 ICRL

iVAE Linear Prior

- Input layer: Input batch *(batch size, input dimension)*
- Mean Output layer: $\mathbf{0}$, which is a vector full of 0 with the length 2
- Log Variance Output layer: Fully connected layer, output size = 2

iVAE Nonlinear Prior

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Mean Output layer: $\mathbf{0}$, which is a vector full of 0 with the length 2
- Log Variance Output layer: Fully connected layer, output size = 2

iVAE Linear Encoder

- Input layer: Input batch *(batch size, input dimension)*
- Mean Output layer: Fully connected layer, output size = 2
- Log Variance Output layer: Fully connected layer, output size = 2

iVAE Nonlinear Encoder

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Mean Output layer: Fully connected layer, output size = 2
- Log Variance Output layer: Fully connected layer, output size = 2

iVAE Linear Decoder

- Input layer: Input batch *(batch size, 2)*
- Mean Output layer: Fully connected layer, output size = output dimension
- Variance Output layer: $0.01 \times \mathbf{1}$, where $\mathbf{1}$ is a vector full of 1 with the length of output dimension

iVAE Nonlinear Decoder

- Input layer: Input batch *(batch size, 2)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Mean Output layer: Fully connected layer, output size = output dimension
- Variance Output layer: $0.01 \times \mathbf{1}$, where $\mathbf{1}$ is a vector full of 1 with the length of output dimension

Linear Data Representation Φ

- Input layer: Input batch *(batch size, input dimension)*
- Output layer: Fully connected layer, output size = 1

Nonlinear Data Representation Φ

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Output layer: Fully connected layer, output size = 1

Linear Classifier w

- Input layer: Input batch *(batch size, 1)*
- Output layer: Fully connected layer, output size = 1

Nonlinear Classifier w

- Input layer: Input batch *(batch size, 1)*
- Layer 1: Fully connected layer, output size = 6, activation = ReLU
- Output layer: Fully connected layer, output size = 1

G.2 Demo Architectures for Colored MNIST Digits and Colored Fashion MNIST

iVAE Prior

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 100, activation = ReLU
- Mean Output layer: $\mathbf{0}$, which is a vector full of 0 with the length 100
- Log Variance Output layer: Fully connected layer, output size = 100

iVAE *O*-Encoder

- Input layer: Input batch *(batch size, 2, 28, 28)*
- Layer 1: Convolutional layer, output channels = 32, kernel size = 3, stride = 2, padding = 1, activation = ReLU
- Layer 2: Convolutional layer, output channels = 32, kernel size = 3, stride = 2, padding = 1, activation = ReLU
- Layer 3: Convolutional layer, output channels = 32, kernel size = 3, stride = 2, padding = 1, activation = ReLU
- Output layer: Flatten

iVAE $(\boldsymbol{Y}, E)$-Encoder

- Input layer: Input batch *(batch size, input dimension)*
- Output layer: Fully connected layer, output size = 100, activation = ReLU

iVAE $(\boldsymbol{O}, \boldsymbol{Y}, E)$-Merger/Encoder

- Input layer: Input batch *(batch size, input dimension)*
- Layer 1: Fully connected layer, output size = 100, activation = ReLU
- Mean Output layer: Fully connected layer, output size = 100
- Log Variance Output layer: Fully connected layer, output size = 100

iVAE Decoder

- Input layer: Input batch *(batch size, 100)*
- Layer 1: Fully connected layer, output size = $32 \times 4 \times 4$, activation = ReLU
- Layer 2: Reshape to *(batch size, 32, 4, 4)*
- Layer 3: Deconvolutional layer, output channels = 32, kernel size = 3, stride = 2, padding = 1, outpadding = 0, activation = ReLU
- Layer 4: Deconvolutional layer, output channels = 32, kernel size = 3, stride = 2, padding = 1, outpadding = 1, activation = ReLU
- Layer 5: Deconvolutional layer, output channels = 2, kernel size = 3, stride = 2, padding = 1, outpadding = 1
- Mean Output layer: activation = Sigmoid
- Variance Output layer: $0.01 \times \mathbf{1}$, where $\mathbf{1}$ is a matrix full of 1 with the size of $2 \times 28 \times 28$.

Data Representation Φ

- Input layer: Input batch *(batch size, 2, 28, 28)*
- Layer 1: Convolutional layer, output channels = 32, kernel size = 3, stride = 2, padding = 1, activation = ReLU
- Layer 2: Convolutional layer, output channels = 32, kernel size = 3, stride = 2, padding = 1, activation = ReLU
- Layer 3: Convolutional layer, output channels = 32, kernel size = 3, stride = 2, padding = 1, activation = ReLU

- Layer 4: Flatten
- Mean Output layer: Fully connected layer, output size = 100
- Log Variance Output layer: Fully connected layer, output size = 100

Classifier w

- Input layer: Input batch *(batch size, 100)*
- Layer 1: Fully connected layer, output size = 100, activation = ReLU
- Output layer: Fully connected layer, output size = 1, activation = Sigmoid