OUT-OF-DISTRIBUTION GENERALIZATION WITH MAXIMAL INVARIANT PREDICTOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-Distribution (OOD) generalization is a problem of seeking the predictor function whose performance in the worst environment is optimal. This paper makes both theoretical and algorithmic contributions to OOD problem. We consider a set of all invariant features conditioned to which the target variable and the environment variable becomes independent, and theoretically prove that one can seek an OOD optimal predictor by looking for the mutual-information maximizing feature amongst the invariant features. We establish this result as *Maximal Invariant Predictor condition*. Our theoretical work is closely related to approaches like Invariant Risk Minimization and Invariant Rationalization. We also derive from our theory the *Inter Gradient Alignment*(IGA) algorithm that uses a parametrization trick to conduct *feature searching* and *predictor training* at once. We develop an extension of the Colored-MNIST that can more accurately represent the pathological OOD situation than the original version, and demonstrate the superiority of IGA over previous methods on both the original and the extended version of Colored-MNIST.

1 INTRODUCTION

In general, most machine learning algorithms today make an inherent assumption that all members of all datasets in concern are independently and identically sampled from the same distribution (IID). Unfortunately, this assumption is not always valid (Bengio et al., 2020). In reality, train-dataset and test-dataset can be coming from different distributions, on which the input-output relations are different because of the presence of environmental factors such as those related to the way the data were collected and where the data were obtained (Shen et al., 2018; Storkey, 2009). This is why most machine learning algorithms often fail when challenged to make a prediction on the dataset sampled from "yet-unseen" distribution ("out of distribution (OOD)" dataset) (Arjovsky et al., 2019).

To address this problem, we need to consider the set of distributions that can be produced by all possible environmental factors, and look for the model whose performance on the worst-case environment is optimal; that is, we want to find a solution of

$$\arg\min_{f} \max_{\epsilon \in \mathcal{A}} \mathrm{Cost}(f|\epsilon) \tag{1}$$

where $\mathcal{A}$ is the set of all possible values for the environmental factor, and $\mathrm{Cost}(f|\epsilon)$ is the cost of the model f in the presence of the environmental factor ϵ. The problem (1) is often referred to as OOD generalization problem (Búhlmann, 2018; Arjovsky et al., 2019). We would say that f is OOD-optimal if it solves (1). Unfortunately, as mentioned in Arjovsky et al. (2019), the problem (1) often cannot be directly solved because we cannot observe datasets in all environments of $\mathcal{A}$. This problem is both interesting and difficult because not all distributions can be produced by the members of $\mathcal{A}$; indeed, if *any* distribution can be produced by some $\epsilon \in \mathcal{A}$, then $\max_{\epsilon \in \mathcal{A}} \mathrm{Cost}(f|\epsilon)$ can be equally maximal for any choice of f because the distribution that is most adversarial to f would also be realizable by some ϵ in $\mathcal{A}$.

Several studies in the past tackled this OOD problem from both theoretical and algorithmic direction, with the aforementioned work of Arjovsky et al. (2019) being one of them. While different in approach, studies like Arjovsky et al. (2019) and Peters et al. (2016); Búhlmann (2018); Subbaswamy et al. (2019) all generally advocate that, in order to solve (1), one shall exploit some *invariance* that is shared across the observed set of environments, such as causal relations or a *feature* that is equally important in all environments. These approaches are different from distributional robustness-type methods (Ben-Tal et al., 2013; Hu et al., 2018; Najafi et al., 2019) that aim to optimize the worst-performance of the model over a compact set of environments embedded in some *metric* space. While distributional robustness is a powerful family of methods, it may not necessarily be the best option when we can't foresee how far the new environment $\tilde{\epsilon}$ would be from the training environments.

The invariance-based OOD studies mentioned in the previous paragraph are, however, case-specific results in the sense that they prove the OOD optimality of their methods on situations in which the relationships amongst the observable variables can be described by some particular graphical model or a certain form of linear models. Therefore, the successes of these studies make us wonder the following two fundamental questions in a more general setting : "When does the invariant feature help us in solving OOD problem?" "What type of invariant feature is useful in OOD problem?" To the best of our knowledge, there has not been a study that has succeeded in answering the first question in the scope beyond a 'case study'. There is still much room left for further investigation. As for the second question, Ferenc (2019) recently investigated an information theoretic generalization of Arjovsky et al. (2019), and suggested an objective

$$\underset{P(Y|\Phi(X),\epsilon)=P(Y|\Phi(X))\forall\epsilon}{\arg\max} \quad I(Y, \Phi(X)) \tag{2}$$

where Y is the target variable, X is the input variable, ϵ is the environment factor and Φ is searched over a general space of nonlinear measurable functions. This objective claims that one needs not only *some* invariant feature, but an invariant feature that maximizes the mutual information with the target variable. Chang et al. (2020) uses a similar approach in their particular graphical model setting.

In this study, we will discuss *when* we can use an invariant feature to solve OOD problem as well as *when* we can solve (1) using (2). More particularly, we use the theory of basic probability and the classic decomposition result used in Darmois (1953); Peters et al. (2012); Achille & Soatto (2017) to investigate when we can use the invariance of form $P(Y|\Phi, \epsilon) = P(Y|\Phi)$ to solve OOD problem. We also name the solution of (2) as **Maximal Invariant Predictor**(MIP), and fill the gap between the OOD objective (1) and the objective (2). To find a MIP, we derive **Inter-Gradient-Alignment(IGA)** algorithm directly from (2), which uses a parametrization trick to conduct *invariant-feature searching* and *predictor training* simultaneously. We demonstrate the efficacy of our algorithm on Colored MNIST(C-MNIST, Arjovsky et al. (2019)) and **Extended Colored-MNIST**(EC-MNIST), a slight generalization of the original that can be used to more accurately construct the pathological OOD situation described in Arjovsky et al. (2019). We summarize our contributions below.

1. We prove a sufficient condition required for an invariant Φ whose corresponding $\mathbb{E}[Y|\Phi(X)]$ is OOD optimal, and justify the use of Maximal Invariant Predictor(MIP) for to solve OOD problem.
2. From MIP we derive Inter-Gradient-Alignment(IGA), a novel OOD algorithm.
3. We present Extended Colored MNIST, a generalization of the original Colored MNIST that can be used to represent an important pathological case that cannot be realized by the original.
4. We demonstrate the efficacy of IGA on Colored MNIST and Extended Colored MNIST.

The paper is structured as follows. We first describe the notations + problem setting in section 2. We present our theoretical results in section 3, and describe IGA in section 4. Finally, we demonstrate the efficacy of our algorithm on C-MNIST and EC-MNIST in section 5.

2 Notations and Problem Setting

In this section, we introduce the set of notations we use throughout the paper, and describe our theoretical claims in more formal language. We use r.v to abbreviate *random variable*.

Notation We follow the rules of notations used in a standard probability text like Durrett (2019). For any set of r.vs $\mathcal{M} = \{M_1, M_2, .., \}$ on the probability triple $(\Omega, \mathcal{F}, P)$, we say $Z \in \sigma(\mathcal{M})$ or *Z is measurable with respect to $\sigma(\mathcal{M})$* whenever Z can be written as a measurable function of Ms. **We may use Z and $Z(\mathcal{M})$ interchangeably in this case**[1]. We say $A \perp B$ when A and B are

[1] For those not familiar with this notation, please identify $\sigma(M)$ as $\{Z; Z = s(M)$ for some function $s\}$ in the main part of this manuscript.

independent. Let us also use a lower case letter to denote a realization of the corresponding r.v (e.g., m is a realization of M, and ϵ is a realization of $\mathcal{E}$). We always use $\mathbb{E}$ to represent the expectation with respect to the probability distribution P. Inside $\mathbb{E}$, variables in upper case are the only variables that are integrated with respect to P. Also, for any probability distribution Q on $\mathcal{F}$, we use its lower case letter q to denote its density.

Formal Problem Statement Let X,Y,$\mathcal{E}$ respectively represent the input r.v, the target r.v, and the environmental r.v. At the time of the training, only X, Y are observable, and the observations are grouped by the realizations of corresponding $\mathcal{E}$. We will measure the performance of the predictor $f(X)$ by some Bregman divergence loss D, which generalizes popular losses like KL divergence and Mean Square Error (Banerjee et al., 2003). That is, for all $\epsilon \in \text{supp}(\mathcal{E})$, we compute the loss on the environment ϵ by $L_\epsilon(f) := \mathbb{E}[D(f(X), Y)|\epsilon]$. OOD problem is to seek the minimizer of the Out-of-Distribution (OOD) loss, given by

$$\arg\min_f \max_{\epsilon \in \text{supp}(\mathcal{E})} L_\epsilon(f), \tag{3}$$

We say that f is OOD optimal if it is a solution of (3). That being said, we also assume the following for the underlying distributions.

Invariance assumption We assume that

$$\mathcal{I} := \{\Phi \in \sigma(X); Y \perp \mathcal{E}|\Phi\} = \{\Phi \in \sigma(X); I(Y; \mathcal{E}|\Phi) = 0\} \tag{4}$$

is non-empty. We refer to $\mathcal{I}$ as the set of ***invariant features*** from now on. In the graphical model represented in Figure 1, for example, $\mathcal{I}$ is the set of all random variables that can be generated by some subset of $\{X_1, X_3, X_4\}$ that contains X_1. This fact can be verified with d-separation theorem (Bishop, 2007). However, in our study, we do not develop our theory from an explicit graphical models like the one we just explained. Unlike in Peters et al. (2016); Subbaswamy et al. (2019); Chang et al. (2020), we also do not necessarily assume that some subset of X's *coordinates* can represent something meaningful in the system (e.g., node representation in DAG). Instead, we study in the setting in which the data is generated from some abstract distribution $P(X, Y, \mathcal{E})$. Our invariance assumption is not too far-fetched in this general setting as well. For example, consider the problem of animal-image recognition. If Y, X, $\mathcal{E}$ are respectively the animal label, the image, and the person who takes the picture, an animal's biological feature captured in the image can be a member of $\mathcal{I}$. Indeed, such Φ is not unique all the time as well; for instance, *face of animal* and *posture of animal* can both be in $\mathcal{I}$.

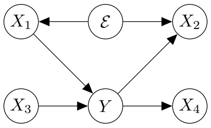


Figure 1: A graphical model with nonempty $\mathcal{I}$ (4). The target variable is Y and the covariate is $X := [X_1, X_2, X_3, X_4]$.

3 THEORY

3.1 RELATED THEORETICAL WORKS

In order to highlight the contributions of our work, we would like to mention several theoretical studies of OOD optimal predictors that are related to our work. Arjovsky et al. (2019) is one of the pioneers that proposed a construction of an OOD optimal solution from an invariant feature. In their work, they theoretically investigated a specific case in which Φ is linear in X and $g(\Phi(X)) = \mathbb{E}[Y|\Phi(X), \epsilon]$ for an invariant Φ is also linear in Φ for all ϵ, and proved the condition under which their algorithm can obtain an OOD optimal solution in their situation. Chang et al. (2020) also showed that, when the underlining graphical model is of a specific form, they can find an OOD optimal solution by seeking $\arg\max_{Y \perp \mathcal{E}|M \odot X} I(Y; M \odot X)$ amongst all binary mask M for which $Y \perp \mathcal{E}|M \odot X$. Rojas-Carulla et al. (2018) claims that one can find an OOD optimal solution in the form $\mathbb{E}[Y|M \odot X]$ for a binary mask M. Rojas-Carulla et al. (2018), however, makes an unstated assumption that for any ϵ, there exists another environment ϵ' such that $P(Y, X|\epsilon') = P(M \odot X, Y|\epsilon)P(M^c \odot X|\epsilon)$ with M^c being the complementary mask of M. While the cases studied in these works are all important on their own, many of their assumptions do not hold in general. The proofs of the OOD optimality in many of these works above also assume that the observed coordinate decomposition of X is aligned in a such a special way that the invariant feature can be expressed using a subset of X's coordinates or a linear function of X.

In this paper, we aim to make a more general statement than these predecessors. First, we will describe when we can use the invariant feature in $\mathcal{I}$ (4) to find a solution to (3). Second, we will describe when we can use MIP (2) to solve (3), thereby giving support to the idea in Ferenc (2019). In the rest of this section, we provide supports to these two claims using the theories of basic probability and the result of Darmois (1953) which claims that, for any pair of random variables X, Y, we can find a noise function $N_Y \perp X$ such that $Y = f(X, N_Y)$ when the cumulative distribution is sufficiently regular.

3.2 Maximal Invariant Predictor

In this section we provide brief definitions and theoretical results. For their details, please see Appendix B. Let us assume that the invariant set $\mathcal{I}$ in (4) is non-empty, and suppose $\Phi \in \mathcal{I}$. Then applying the aforementioned result of Darmois (1953) to the r.v pair $(\Phi, \mathcal{E})$, we can say that there exists $\mathcal{E}_\psi \perp \Phi$ such that $\sigma(\Phi, \mathcal{E}) = \sigma(\Phi, \mathcal{E}_\psi)$. Applying the same argument again to $(\mathcal{E}_\psi, \mathcal{E})$, we can say that there is $\mathcal{E}_\phi \perp \mathcal{E}_\psi$ such that $\sigma(\mathcal{E}, \mathcal{E}_\psi) = \sigma(\mathcal{E}_\phi, \mathcal{E}_\psi)$. For any $\Phi \in \mathcal{I}$, we can thus decompose ϵ into the pair $(\epsilon_\phi, \epsilon_\psi)$. In summary, we have the following;

- $\mathcal{E}_\psi \in \sigma(\mathcal{E})$, a part of $\mathcal{E}$ that is independent of Φ
- $\mathcal{E}_\phi = \mathcal{E}_\psi^c$; that is, $\mathcal{E}_\phi \perp \mathcal{E}_\psi$ and $\mathcal{E} \in \sigma(\mathcal{E}_\phi, \mathcal{E}_\psi)$.

Put in still other words, this is a decomposition of $\mathcal{E}$ into the component that can affect Φ and the component that cannot. Also, let us apply Darmois (1953) to (X, Φ) to obtain $\Psi \perp \Phi$ with $X \in \sigma(\Phi, \Psi)$.We emphasize that we are assuming a setup that is at least as general as the works of (Darmois, 1953; Peters et al., 2012; Achille & Soatto, 2017) that use this decomposition result. For a more intuitive scenario, consider another image recognition problem of predicting the animal label Y from the image X in the set of natural images collected with a loose directive of "take any pictures of a given list of animals". Suppose that Φ is the physical appearance of the animal. Then the part of the environmental factor that is able to influence Φ can be the biological state of the animal ($\mathcal{E}_\phi$) (e.g., physical state, genetic feature). The part of the environmental factor that can *not* influence Φ can be, for example, the preference of the photographer ($\mathcal{E}_\psi$). Clearly, $\mathcal{E}_\psi \perp \Phi$. Meanwhile, Ψ, the complement of Φ, can be the background of the animal. Since the photographer does not have much choice for the individual animal to encounter, Ψ is independent of Φ if he/she is not so picky.

The following is our first claim about the case in which an invariant feature can be used to find an OOD optimal solution.

Proposition 3.1. *Let $\Phi \in \mathcal{I}$. Also, put $f^*(X) = \mathbb{E}[Y|\Phi]$ and suppose that, for all ϵ_ϕ there exists some $\tilde{\epsilon}_\psi$ such that*[2] *$X \perp Y|(\Phi, \epsilon_\phi, \tilde{\epsilon}_\psi)$. Then $f^* = \arg\min_f \max_{\epsilon \in \mathrm{supp}(\mathcal{E})} L_e(f)$.*

For the proof, see Appendix B. Because the proposition 3.1 is pivotal in our theory, we would like to give a name to the condition used therein;

Definition 3.2. *We say $\Phi \in \mathcal{I}$ satisfies a controllability condition if for all ϵ_ϕ, there exists at least one corresponding ϵ_ψ such that $X \perp Y|\epsilon_\phi, \epsilon_\psi$.*

In other words, this is a condition in which there is always a way to twig a Φ-irrelevant part of the environment to sever the relation between X and Y using Φ and the environment. From the proposition 3.1, we can also say the following.

Corollary 3.3. *If there exists one ϵ_ψ for which $\Psi \perp Y|\epsilon_\psi$, then $\mathbb{E}[Y|\Phi]$ is OOD optimal.*

Figure 8 in Appendix B is an example of a system in which the corollary 3.3 can hold. The corollary 3.3 may be helpful in determining when the OOD optimality statement in Rojas-Carulla et al. (2018) is valid. In their proof, Rojas-Carulla et al. (2018) claims their solution to be OOD-optimal under an

[2] As we describe in Appendix B, we can increase the number cases in which this holds if we choose $\mathcal{E}_\psi$ to be larger in the sense of sigma algebra.

unstated assumption that, for any ϵ, there exists some ϵ' such that $P(Y, X|\epsilon') = P(\Phi, Y|\epsilon)P(\Psi|\epsilon)$. We can guarantee the validity of this assumption if $\Psi \perp \mathcal{E}$ and if Φ satisfies the condition of the corollary 3.3. It also turns out that, if ϵ and Φ can sever the relation between Y and X, then this Φ is optimal in the environment ϵ.

Proposition 3.4. *Suppose $\Phi \in \mathcal{I}$ and suppose that ϵ satisfies $X \perp Y|\Phi, \epsilon$. Then for this particular ϵ, $\Phi = \arg\max_{Z\in\sigma(X)} I(Y; Z|\epsilon)$.*

In other words, the controllability condition for a feature Φ also guarantees that we can twig the Φ-irrelevant part of ϵ to produce $\tilde{\epsilon}$ in which Φ is the best feature in the environment $\tilde{\epsilon}$. The situation considered here is not too unrealistic. In the case of the animal-recognition example we mentioned above, we can choose $\tilde{\epsilon}_\psi$ to be a person who wants to take a picture of every animal in front of an all-green background for the Chroma-key purpose. With such $\tilde{\epsilon}$, the biological feature of an animal would always be the best feature.

At this point yet, we still have not made an explicit connection between (1) and (2). The following lemma would help us to fill the gap.

Lemma 3.5. *Suppose $\Phi \in \mathcal{I}$ and suppose that for some ϵ^*, $\Phi \perp Y|\epsilon^*$ Then there exists no $\tilde{\Phi} \in \mathcal{I}$ with $\Phi \in \sigma(\tilde{\Phi})$ such that $I(Y; \Phi) < I(Y; \tilde{\Phi})$.*

The lemma 3.5 states that, unless Φ is a maximal element of $\mathcal{I}$, there is no hope in finding an environment in which $\mathbb{E}[Y|\Phi]$ is OOD optimal. This motivates to find $\Phi \in \mathcal{I}$ satisfying

$$\Phi^* \in \arg\max_{\Phi\in\mathcal{I}} I(Y; \Phi) = \arg\max_{I(Y;\mathcal{E}|\Phi)=0} I(Y; \Phi). \tag{5}$$

The equation (5) is indeed the very *Maximal Invariant Predictor(MIP)* condition mentioned in the introduction. When we restrict the search space $\mathcal{I}$ to its linear subset $\{\Phi \in \mathcal{I}; \Phi = M \odot X \textit{ for some binary mask } M\}$, we retrieve the objective in Chang et al. (2020). We can justify this objective more formally when appropriate conditions are met. Note that, in cases like Figure 1, all invariant features are generated by a common invariant feature vector $[X_1, X_3, X_4]$. This type of case may also arise when the data is generated by an undirected graphical model that is a perfect map (see Appedix C). In these cases, MIP is unique and the proposition 3.5 would imply that Φ cannot be OOD optimal unless it achieves $\max_{\Phi\in\mathcal{I}} I(Y; \Phi)$. If we use these facts, we can also claim the following:

Theorem 3.6. *Suppose that $\mathcal{I}$ can be generated by one invariant subset of r.vs, and that there exists at least one $\Phi \in \mathcal{I}$ that satisfies the controllability condition. Then $\mathbb{E}[Y|\Phi^*]$ is OOD optimal if Φ^* is MIP.*

4 Inter-environmental Gradient Alignment Algorithm

To seek a solution to the MIP objective (5), we propose *Inter Gradient Alignment*(IGA) algorithm, which optimizes the following objective;

$$\arg\min_{\theta} \mathbb{E}[L_{\mathcal{E}}(\theta)] + \lambda \operatorname{trace}(\operatorname{Var}(\nabla_\theta L_{\mathcal{E}}(\theta))). \tag{6}$$

where L_ϵ is the loss of the θ-parameterized prediction model computed on the environment ϵ. We are using the upper case $\mathcal{E}$ above to indicate that $\mathcal{E}$ is the random variable with respect to which the expectation and the variance are taken. We call this Inter Gradient Alignment algorithm, because it encourages the gradient computed at each environment to align with those of other environments. Note that (6) does not require us to separately compute the invariant feature Φ. Table 2 is a summary of our algorithm. We use $\hat{\mathbb{E}}$ to denote the empirical expectation. Let $\mathcal{G}_{train} = \{\epsilon_1, \epsilon_2, ...\epsilon_k\} \subset \operatorname{supp}(\mathcal{E})$ denote a size$=k$ set of environments from which the user can collect the dataset. IGA assumes that the user is given a set of datasets $\{D_\epsilon; \epsilon \in \mathcal{G}_{train}\}$, with each D_ϵ being a set of samples from $p(y, x|\epsilon)$. Because we assume that neither the identities nor the effects of ϵ are disclosed to the user in any way whatsoever, each ϵ is only observable as an integer index of datasets.

4.1 Derivation of IGA

We derive (6) from the MIP objective (5). To evaluate $I(Y; \Phi)$ and $I(Y; \mathcal{E}|\Phi)$, we need to be able to evaluate both $p(y|\phi, \epsilon)$ and $p(y|\phi)$ for arbitrary realizations y, ϕ and ϵ. However, the complexity of

```
Algorithm 1 Inter-environmental Gradient Alignment Algorithm
Input: $q_\theta(y|x), \{D_\epsilon; \epsilon \in \mathcal{G}_{train}\}$
Return: $q_\theta$
for each iteration do
    for $\epsilon_i$ in $\mathcal{G}_{train}$ do
        compute $L_{\epsilon_i}(\theta) = \hat{\mathbb{E}}[\log q_\theta(Y|X)|\epsilon_i]$
        compute $\nabla_\theta L_{\epsilon_i}(\theta)$
    end for
    compute $\hat{\mathbb{E}}[L_\mathcal{E}] := \frac{1}{|\mathcal{G}_{train}|} \sum_{\epsilon_i \in \mathcal{G}_{train}} L_{\epsilon_i}(\theta)$
    compute trace($\hat{\text{Var}}(\nabla_\theta L_\mathcal{E}(\theta))) := \sum_{\epsilon_i \in \mathcal{G}_{train}} ||\nabla_\theta L_{\epsilon_i}(\theta) - \nabla_\theta \hat{\mathbb{E}}[L_\mathcal{E}]||^2$
    update $\theta$ by the gradient descent using the eq (6)
end for
```

Figure 2: IGA algorithm

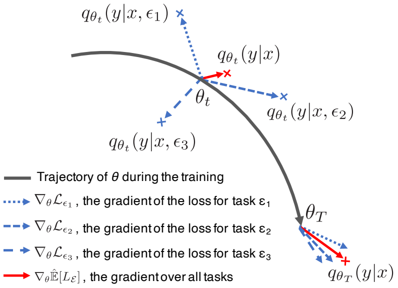

$q_{\theta_t}(y|x,\epsilon_1)$
$q_{\theta_t}(y|x)$
θ_t
$q_{\theta_t}(y|x,\epsilon_2)$
$q_{\theta_t}(y|x,\epsilon_3)$
Trajectory of θ during the training
$\nabla_\theta \mathcal{L}_{\epsilon_1}$, the gradient of the loss for task ε₁
$\nabla_\theta \mathcal{L}_{\epsilon_2}$, the gradient of the loss for task ε₂
$\nabla_\theta \mathcal{L}_{\epsilon_3}$, the gradient of the loss for task ε₃
$\nabla_\theta \hat{\mathbb{E}}[L_\mathcal{E}]$, the gradient over all tasks
θ_T
$q_{\theta_T}(y|x)$

Figure 3: The relation of the θ-updates in IGA to the gradient computed at each task.

$p(y|\phi,\epsilon)$ is usually unknown. The relationship between $p(y|\phi,\epsilon)$ is $p(y|\phi)$ is unknown as well, and it depends on the choice of $h : X \to \Phi$. We therefore use a specific parametrization to represent these distributions. Note that because $\Phi \in \sigma(X)$ (so that Φ can be written as a function of X), $p(y|\phi)$ can be written as $q(y|x)$ for some h-related distribution q. Representing q with a θ-parametrized family of functions $f(\cdot|\theta)$, we may write $p(y|\phi,\epsilon)$ and $p(y|\phi)$ as

$$p(y|\phi) \cong q_\theta(y|x) = f(y|x; \theta - \alpha\nabla_\theta \mathbb{E}[L_\mathcal{E}(\theta)]) \tag{7}$$

$$p(y|\phi,\epsilon) \cong q_\theta(y|x,\epsilon) = f(y|x; \theta - \alpha\nabla_\theta L_\epsilon(\theta)). \tag{8}$$

The advantage of this parametrization is multi-fold. First, when we evaluate these approximations empirically, the ϵ in our approximations only appears as an index in the empirical expectation. We do not have to use the *real* identity of ϵ in the evaluation of the approximation. Second, if the model is regular enough, we can expect $\mathbb{E}[q_\theta(y|x,\mathcal{E})] \cong q_\theta(y|x)$ for small enough α. Finally, by its design, $q_\theta(y|x,\epsilon)$ is closer to $p(y|x,\epsilon)$ than $q_\theta(y|x)$ is at all time, which is required by the property of Kullback Leibler Divergence. See Appendix C.1 for more discussion about our parametrization. Now, $I(Y;\Phi)$ in (6) can be evaluated with $\mathbb{E}[L_\mathcal{E}(\theta)]$. The regularization term $I(Y;\mathcal{E}|\Phi)$ can be approximated as

$$\begin{aligned} I(Y;\mathcal{E}|\Phi) &\cong \mathbb{E}[d_{KL}(q_\theta(Y|X,\mathcal{E}) \| q_\theta(Y|X))] \\ &\cong \mathbb{E}[L_\mathcal{E}(\theta - \alpha\nabla_\theta \mathbb{E}[L_\mathcal{E}(\theta)]) - L_\mathcal{E}(\theta - \alpha\nabla_\theta L_\mathcal{E}(\theta))] \cong \alpha \, \text{trace}(\text{Var}(\nabla_\theta L_\mathcal{E}(\theta))). \end{aligned} \tag{9}$$

Thus, $q_\theta(y|x,\epsilon)$ is close to $q_\theta(y|x)$ if α is small. Figure 3 is an illustration each ∇L_ϵ relative to the actual direction of the update. For more detailed computation, see Appendix C.2.

4.2 Two phase learning vs one phase learning

Previous deep learning OOD algorithms like Arjovsky et al. (2019) and Chang et al. (2020) aim to learn an OOD optimal predictor in two phases: (i) the phase of learning a *invariant feature* $\Phi \in \sigma(X)$ via a function $h : X \to \Phi$ and (ii) the phase of learning a predictor $g : \Phi \to Y$. When the loss is of Bregman divergence type, the optimal g^* for any given h is of the form $\mathbb{E}[Y|\Phi]$ or $P(Y|\Phi)$, and g^* itself depends on the choice of h; we shall therefore write g_h^* for g^*. Because g_h^* and h are dependent on one another, allowing a large model space for either g_h^* or h would make the training difficult. The algorithm of Arjovsky et al. (2019) took the approach of using a small model space for g_h^* and large model space for h, and assumed g_h^* to be always linear in black box Φ. Meanwhile, because the complexity of g_h^* may vary with h, it may not be possible to find g_h^* if the model space is too small. Chang et al. (2020) took an approach of assuming a possibly large model space for g_h^*, and instead sought h from those that can be expressed as $M \odot X$ with a binary mask random variable M. However, if the model space of g_h^* is large, optimizing g_h^* with respect to function h can be a daunting task. In fact, Chang et al. (2020) is giving up the computation of the gradient of g_h^* with respect to the parameter of h. As we discuss in Appendix E.2, it was empirically difficult to train these inter-related functions in two separate phases when the model space was large. IGA is different from previous approaches in that it implicitly trains g and h in one phase.

5 Experiment

We evaluated the efficacy of our method on *Colored-MNIST* and *Extended Colored-MNIST*. For more detail of the setting and the implementation, please see Appendix D.

5.1 Experimental setup

Extended Colored-MNIST (EC-MNIST) Because our extension generalizes the original *Colored-MNIST* (*C-MNIST*) in Arjovsky et al. (2019), we describe our *Extended Colored-MNIST* (*EC-MNIST*) first. Unlike the original C-MNIST, the invariant predictor set $\mathcal{I}$ (4) in our extended version contains more than one variable. Our extended version is particularly different from the original one in that it can describe a case in which the domain invariant feature (i.e, features Φ that is independent to $\mathcal{E}$) alone is not necessarily sufficient for the optimal environment-agnostic prediction. Each instance of our EC-MNIST dataset is constructed as follows;

1. Set x_{ch2} to 1 with probability ϵ_{ch2}. Set it to 0 with probability $1 - \epsilon_{ch2}$.
2. Generate a binary label $\hat{y}_{obs}$ from y with the following rule: $\hat{y}_{obs} = 0$ if $y \in \{0 \sim 4\}$ and $\hat{y}_{obs} = 1$ otherwise. If $x_{ch2} = k$, construct y_{obs} by flipping $\hat{y}_{obs}$ with probability $p_k (k \in \{0, 1\})$.
3. Put $y_{obs} = \hat{x}_{ch0}$, and construct x_{ch0} from $\hat{x}_{ch0}$ by flipping $\hat{x}_{ch0}$ with probability ϵ_{ch0}.
4. Construct x_{obs} as $x_{fig} \times [x_{ch0}, (1 - x_{ch0}), x_{ch2}]$. As an RGB image, this will come out as an image in which the red scale is *turned on* and the green scale is *turned off* if $x_{ch0} = 1$, and other-way around if $x_{ch0} = 0$. Blue scale is *turned on* only if $x_{ch2} = 1$.

Figure 5 is the graphical model for the generation of EC-MNIST. In the experiment on this dataset, only (Y_{obs}, X_{obs}) are assumed observable. At training times, the machine learner will be given a set of datasets $\mathcal{D}_{train} = \{D_\epsilon; \epsilon \in R_{train}\}$ in which D_ϵ is a set of observations gathered when $\mathcal{E} = \epsilon$. In the test time, the learner will be challenged to make an inference of Y_{obs} from X_{obs} in the presence of an unknown environment ϵ^*. If ϵ^* is far away from any members of R_{train}, a model overfitted to $\mathcal{D}_{train}$ can fail catastrophically on ϵ^*.

For our EC-MNIST, Y_{obs} can be predicted at maximal probability of $\epsilon_{ch2} \max\{p_0, 1 - p_0\} + (1 - \epsilon_{ch2}) \max\{p_1, 1 - p_1\}$. In this problem, X_{fig} is a $\mathcal{E}$ independent factor. At the same time, X_{ch2} is a member of $\mathcal{I}$, and X_{fig} together with X_{ch2} can create a better predictor than X_{fig} alone. In fact, the oracle prediction by X_{fig} alone can attain an average value of at most $\max\{\epsilon_{ch2} p_0 + (1 - \epsilon_{ch2}) p_1, \epsilon_{ch2}(1 - p_0) + (1 - \epsilon_{ch2})(1 - p_1)\}$, which is lower than the that of the $[X_{fig}, X_{ch2}]$ oracle. This follows from Fatou's lemma (Folland, 2013). In EC-MNIST, X_{fig} is the maximal random variable that is independent of environmental factors. The oracle with X_{fig} alone coincides with the upper bound of Adversarial Domain Adaptation(ADA) (Li et al., 2018; Zhang et al., 2018) which looks for a feature Φ with which it is difficult to identify which environment the X is coming from. Because ADA essentially looks for a feature Φ that is independent to $\mathcal{E}$, ADA cannot provide an optimal solution in this case. IRM (Arjovsky et al., 2019) also discusses such a case in their work.

Colored MNIST (C-MNIST): The original *C-MNIST* in Arjovsky et al. (2019) is a special case of our EC-MNIST in which the distribution of x_{ch2} does not vary with ϵ. Figure 4 is the graphical model of C-MNIST. Notice that if one uses X_{fig} alone, one can make the predictions at the optimal accuracy of $\max(1 - p, p)$. Thus, ADA may attain the optimal solution in C-MNIST. In this sense, our EC-MNIST is more suited for the investigation of pathological cases in OOD study.

5.2 Result

We compared our algorithm against Invariant Risk Minimization (IRM)(Arjovsky et al., 2019), Empirical Risk Minimization (ERM), and the aforementioned oracle(s) in terms of OOD accuracy, which is the minimum accuracy over all environments. For the comparison against other two-phase training method akin to Chang et al. (2020), see Appendix E.2.

Colored-MNIST The left column in table 6 compares the results of the algorithms in terms of the OOD accuracy (3). As we see in the table, our method outperforms both ERM and IRM. Figure 7(a)(b) plots the OOD accuracy against the regularization parameter. In general, larger regularization parameter promotes the performance. The OOD accuracy plateaus around $\lambda \sim 10^4$.

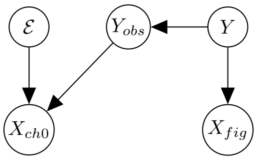


Figure 4: The graphical model of *C-MNIST*

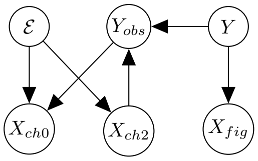
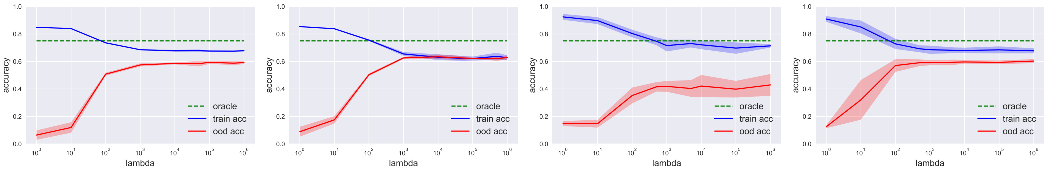


Figure 5: The graphical model of *EC-MNIST*

Model	OOD accuracy *C-MNIST*	 *EC-MNIST*
Oracle	0.75	0.75
X_{fig}	0.75	0.25
ERM	$0.172 \pm .029$	$0.176 \pm .029$
IRM	$0.592 \pm .011$	$0.430 \pm .080$
Ours	$\mathbf{0.620} \pm \mathbf{.015}$	$\mathbf{0.594} \pm \mathbf{.013}$
IRM†	$0.593 \pm .044$	

Figure 6: Numerical performance of OOD algorithms. X_{fig} designates the *figure-only* oracle (i.e. the upper bound of ADA). † is the OOD accuracy of IRM reported in Ahuja et al. (2020). For the result of IRM in Arjovsky et al. (2019), please see Appendix E.1.



(a) IRM on *C-MNIST* (b) Ours on *C-MNIST* (c) IRM on *EC-MNIST* (d) Ours on *EC-MNIST*

Figure 7: The plot of regularization parameter λ against the accuracies on *C-MNIST* ($p = 0.25$) and on *EC-MNIST* ($p_0 = 0.25, p_1 = 0.75$).

Extended Colored-MNIST The right column in the table 6 compares the results of the algorithms in terms of the OOD performance (3). Again in this set of experiments, we perform better than ERM and IRM. We also perform better than the X_{fig} oracle. Because X_{ch2} is necessary in order to outperform the X_{fig} oracle (see section 5.1), our result suggests that we are actually using the feature X_{ch2} in making the prediction. Figure 7(c)(d) plots the OOD accuracy against the regularization parameter. Again, a larger regularization parameter generally promotes the OOD performance.

6 Other Related Works

As mentioned in the introduction, many studies in the past advocated the importance of invariance in OOD problem, and they differ in the form of invariance they propose. Because the feature that is easy to learn differs with the chosen model-architecture and the environment, some human-imposed measure often seems necessary to encourage the learner to focus on the environment-agnostic relations (Geirhos et al., 2020). The aforementioned ADA (Li et al., 2018; Zhang et al., 2018) is practically a method to look for good ϵ-independent feature. We may say that such a method is looking for a good feature that is invariant of the choice of domain, or equivalently the choice of ϵ to which the input X is conditioned to. Covariate shift (Ben-David et al., 2007; Johansson et al., 2019; Shimodaira, 2000) works in the setting that $P(Y|X, \epsilon)$ is invariant with respect to ϵ. As discussed in Búhlmann (2018), DAG causal relation is another powerful representation of *invariance* with respect to the environment, and much research has been done to better utilize the causal relations for OOD problem (Peters et al., 2016; Subbaswamy et al., 2019; Rojas-Carulla et al., 2018; Búhlmann, 2018; Chang et al., 2020). Most of these works look for a part of causal relations that remain invariant with respect to environments. At the same time, identifying causal relations are difficult in practice (Búhlmann, 2018), and these theories are not always readily applicable. We may say that recent studies of IRM and its variants (Arjovsky et al., 2019; Ahuja et al., 2020) are, in part, and an effort to investigate the invariance of form $\mathbb{E}[Y|\Phi, \epsilon] = \mathbb{E}[Y|\Phi]$ where Φ is a black-box non-linear feature. Chang et al. (2020) seeks an invariant feature of form $P(Y|M \odot X, \epsilon) = P(Y|M \odot X)$ with M being a binary mask variable. Our study is an approach that claims the usefulness of the invariance of form $P(Y|\Phi, \epsilon) = P(Y|\Phi)$ with black-box Φ, and we used an information-theoretic mean to justify the usage of such invariant feature in OOD problem.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *arXiv preprint arXiv:1706.01350*, 2017.

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor, 2003.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

A. Ben-Tal, D. den Hertog, A.M.B. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013. ISSN 0025-1909. Appeared earlier as CentER Discussion Paper 2011-061.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2020.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 2007.

Peter Búhlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

Enrique Castillo, Jose M Gutierrez, and Ali S Hadi. *Expert systems and probabilistic network models*. Springer Science & Business Media, 2012.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S Jaakkola. Invariant rationalization. *arXiv preprint arXiv:2003.09772*, 2020.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

G. Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 21(1/2):2–8, 1953. ISSN 03731138.

Rick Durrett. *Probability: Theory and Examples*. Thomson, 2019.

Huszár Ferenc. Invariant risk minimization: An information theoretic view. 2019. URL `https://www.inference.vc/invariant-risk-minimization/`.

Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013.

Maxime Gasse. *Probabilistic Graphical Model Structure Learning: Application to Multi-Label Classification*. PhD thesis, 2017.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2029–2037, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 125–136. Curran Associates, Inc., 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Fredrik D. Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 527–536. PMLR, 16–18 Apr 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018.

Peter Lucas, José A Gámez, and Antonio Salmerón Cerdan. *Advances in Probabilistic Graphical Models*, volume 213. Springer, 2007.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems 32*, pp. 5542–5552. Curran Associates, Inc., 2019.

Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

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

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1576–1584. Curran Associates, Inc., 2015.

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally regularized learning with agnostic data selection bias. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 411–419, 2018.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000.

Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pp. 3–28, 2009.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3118–3127. PMLR, 16–18 Apr 2019.

Yexun Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. Domain-invariant adversarial learning for unsupervised domain adaption. *arXiv preprint arXiv:1811.12751*, 2018.

Appendix

A Why invariant features?

We give more explanation about our motivation in solving the OOD problem. As mentioned in the introduction, OOD problem arises when one has to deal with sets of datasets coming from possibly different distributions. For example, when we are to conduct image-recognition from a dataset of animal pictures collected by a group of photographers, we might have to consider the fact that some photographer might favor a picture of a given animal in a certain posture at a certain set of locations at a certain time of the day, and some do not. If *camel* happens to be always found in desert in the set of pictures taken by *Jerry* the photographer, an *animal classifier* trained on his dataset may characterize *camel* as *a horse-like figure in desert*. Meanwhile, it might be the case that *Tom* the photographer prefers to take a picture of any animal at a zoo. Indeed, the *animal classifier* trained on *Jerry*'s dataset will most likely fail to recognize *Tom*'s picture of camel as *camel*, because a picture of a *camel in a cage* does not agree with the *description of the camel* learned by the classifier.

Such behavior shall not be considered as a *bug*, because *a horse-like figure in desert* is definitely a characterizing description of *camel* on the distribution that underlies the set of natural images taken by *Jerry*. This type of problem can arise in a more subtle way. Some studies report that in practice, the machine learner may associate the label with even more subtle features like a texture or a pixel-pattern that cannot be discerned by humans (Ilyas et al., 2019; Geirhos et al., 2019).

Then how shall we formalize what is "reasonable" and what is not? If the system of concern is governed by a certain causal relation, the exterior factor can only affect the input-output distribution in a certain way, and some parts of the relations in the system will remain unchanged (Pearl et al., 2009; Búhlmann, 2018). We can borrow from this philosophy; we can say that the exterior factor is affecting the distribution in a "reasonable" manner if there is always some part of the input-output relation that remains invariant under its influence. We can then work with the assumption that some part of the input-output relation is shared in common by all distributions to be considered. For the aforementioned *camel* example, camels'*essential biological feature* in the picture is not affected by the person who takes the picture. This is an example of the approach based on invariance. By finding the part of the system that stays invariant under the effect of environmental factor, one may be able to find a solution to the OOD problem. Our study explains when this is actually possible.

We do not necessarily say that this is the only way to go in all situations. For example, if environmental factor and its effects are known in advance, we may take a completely different strategy. In the animal picture recognition problem we mentioned above, if we know from the beginning that ”photographer” is the environmental factor and that they affect the ”background” part of the image, one may be able to improve the generalization ability of the trained model by allowing it to leverage the similarity relations among the photographers. In such cases, methods of distributional robustness might be particularly helpful (Ben-Tal et al., 2013; Hu et al., 2018; Najafi et al., 2019; Shafieezadeh Abadeh et al., 2015).

But in many cases, figuring out what is affecting the dataset is a difficult problem on its own, and it is often not known how similar the new photographer is to the set of photographers we have seen. This is the motivation behind the invariance-based approach to the OOD problem.

B Proof of the Invariance Results

In this section we provide the proofs for the theoretical results we presented in the section 3 of the main manuscript. Let $\mathcal{E}, X, Y$ be random variables defined with probability triple $(\Omega, \mathcal{F}, P)$. Throughout, we will follow the notation rules we introduced in the main part of the manuscript. Following the convention in probability, we use $\sigma(Z)$ to denote a sigma algebra generated by X, and write $W \in \sigma(Z)$ whenever W is measurable with respect to Z. We will generally use upper case letters to denote random variables, and use lower case letters to denote the corresponding realizations. Unless otherwise noted, we will use $\mathbb{E}$ to represent the expectation with respect to P. We will also use $\perp$ to represent independency relation, and use $A \perp B|C$ to convey that $P(A, B|C) = P(A|C)P(B|C)$.

Now, let Φ be a random variable that is mesurable with respect to $\sigma(X)$, and let us define the following elements that can be derived from Φ:

1. $\mathcal{F}_\psi = \{E \in \sigma(\mathcal{E}); E \perp \Phi\})$ is a set of random variables representable as a function of $\mathcal{E}$ that are independent from variable Φ
2. $\mathcal{E}_\psi \in \mathcal{F}_\psi$. If it exists, we define this to be a random variable that is maximal in the sense that there is no other $Z \in \mathcal{F}_\psi$ with $\sigma(\mathcal{E}_\psi) \subsetneq \sigma(Z)$.
3. $\mathcal{E}_\phi = \mathcal{E}_\psi^c$; that is, $\mathcal{E}_\phi \perp \mathcal{E}_\psi$ and $\sigma(\mathcal{E}_\phi, \mathcal{E}_\phi) = \sigma(\mathcal{E})$.

As for the existence of the last decomposition, we appeal to the result of Darmois (1953), which claims that , for any joint distribution of (X, Y) we can find a function f and a noise $N_Y \perp X$, such that $Y = f(X, N_Y)$. Therefore, for every $\epsilon \in \text{supp}(\mathcal{E})$, let us WLOG suppose a map r for which $r(\epsilon_\phi, \epsilon_\psi) = \epsilon$, and use $(\epsilon_\phi, \epsilon_\psi)$ and ϵ interchangeably. Let us also define the invariant set $\mathcal{I}$ by

$$\mathcal{I} = \{\Phi \in \sigma(X); (Y|\Phi, \mathcal{E}) = (Y|\Phi), Y \not\perp \Phi\} \tag{10}$$

and that this set is non-empty. Then the following lemma will be useful for the main result 3.1, which provides a set of conditions under which the $\Phi \in \mathcal{I}$ satisfies the OOD optimality:

Lemma B.1. *If $\Phi \in \mathcal{I}$, then $\mathcal{E}_\psi \perp (\Phi, Y)$.*

Proof. By definition, $Y \perp \mathcal{E}|\Phi$ so in particular, $Y \perp \mathcal{E}_\psi|\Phi$ and $Y \perp \mathcal{E}_\phi|\Phi$. Moreover, by definition, $\mathcal{E}_\psi \perp \Phi$. Thus $P(\mathcal{E}_\psi|Y, \Phi) = P(\mathcal{E}_\psi|\Phi) = P(\mathcal{E}_\psi)$ and $\mathcal{E}_\psi \perp (\Phi, Y)$, as desired. □

We are now ready to present the proposition 3.1. In an equation that uses multiple $\mathbb{E}$ notation, will use a notation like $\mathbb{E}_Z$ to help seeing that the inside the symbol is an integration about Z.

Proposition B.2. *Let g be a strictly convex, differentiable function and let D be the corresponding Bregman Loss function. Let $\Phi \in \sigma(\mathcal{I})$, and let w_ϕ be the measurable function such that $\Phi = w_\phi(X)$. Also, put $f^*(X) = \mathbb{E}[Y|\Phi] = g^*(w_\phi(X))$ and suppose that, for all ϵ_ϕ there exists $\tilde{\epsilon}_\psi$ such that*

$$X \perp Y|(\Phi, \epsilon_\phi, \tilde{\epsilon}_\psi) \tag{11}$$

Then

$$f^* = \arg\min_f \sup_{\epsilon \in \text{supp}(\mathcal{E})} \mathbb{E}[D(f(X), Y)|\epsilon] \tag{12}$$

Proof. We are going to leverage the fact that, if $\mathcal{G}$ is a sub sigma algebra of $\mathcal{F}$ to which Y is measurable, then Banerjee et al. (2003)

$$\arg\min_{Z \in \mathcal{G}} \mathbb{E}[D(Z, Y)] = \mathbb{E}[Y|\mathcal{G}]. \tag{13}$$

We are also going to use the fact that the Bregman loss function

$$D(a, b) = g(a) - g(b) - \langle a - b, \nabla g(b) \rangle$$

is convex about its first coordinate, a. Now, in order to show the claim, we need to show that $\sup_\epsilon \mathbb{E}[D(f(X), Y)|\epsilon] \geq \sup_\epsilon \mathbb{E}[D(f^*(X), Y)|\epsilon]$ for all measurable f. To show this, for any fixed ϵ we need to be able to find one ϵ' such that

$$\mathbb{E}[D(f(X), Y)|\epsilon'] \geq \mathbb{E}[(D(f^*(X), Y)|\epsilon]. \tag{14}$$

Suppose a fixed ϵ and suppose that $\epsilon = r(\epsilon_\phi, \epsilon_\psi)$ for the appropriate invertible map r, and let ϵ' be such that $\epsilon' = r(\epsilon_\phi, \tilde{\epsilon}_\psi)$ with $\tilde{\epsilon}_\psi$ that satisfies the condition in the claim. Now, the Bregman Loss function $D(x, y)$ is convex with respect to x. With Jensen's inequality and repeated application of Tower rule (Durrett, 2019),

$$\begin{aligned}
\mathbb{E}_{X,Y}[D(f(X), Y)|\epsilon'] &= \mathbb{E}_{\Phi,Y}[\mathbb{E}_X[D(f(X), Y)|\epsilon', \Phi, Y]] && (15)\\
&\geq \mathbb{E}_{\Phi,Y}[D(\mathbb{E}_X[f(X)|\Phi, Y, \epsilon'], Y)|\epsilon']] && (16)\\
&= \mathbb{E}_{\Phi,Y}[(D(\mathbb{E}_X[f(X)|\Phi, \epsilon'], Y)|\epsilon']] && (17)\\
&= \mathbb{E}_{\Phi,Y}[D(h(\Phi, \epsilon'), Y)|\epsilon'] && (18)
\end{aligned}$$

where we set $h(\Phi, \epsilon') = \mathbb{E}[f(X)|\Phi, \epsilon']$ and used Jensen's inequality in the second line. Here, the random variable that is integrated in the last expression is Φ and Y only, and $e_\phi, \epsilon_\psi, \tilde{\epsilon}_\psi$ are all constant. Moreover, by the lemma B.1, $(Y, \Phi)|\epsilon_\phi, \tilde{\epsilon}_\psi = (Y, \Phi)|\epsilon_\phi, \epsilon_\psi$. Therefore, using the property of $\tilde{\epsilon}_\psi$,

$$\begin{aligned}
\mathbb{E}_{\Phi,Y}[D(h(\Phi, \epsilon'), Y)|\epsilon_\phi, \tilde{\epsilon}_\psi] &= \mathbb{E}_{\Phi,Y}[D(h(\Phi, \epsilon_\phi, \tilde{\epsilon}_\psi), Y)|\epsilon_\phi, \tilde{\epsilon}_\psi] && (19)\\
&= \mathbb{E}_{\Phi,Y}[D(h(\Phi, \epsilon_\phi, \tilde{\epsilon}_\psi), Y)|\epsilon_\phi, \epsilon_\psi] && (20)\\
&\geq \mathbb{E}_{\Phi,Y}[D(\mathbb{E}_Y[Y|\Phi, \epsilon_\phi, \epsilon_\psi], Y)|e_\phi, \epsilon_\psi] && (21)\\
&= \mathbb{E}_{\Phi,Y}[D(\mathbb{E}[Y|\Phi], Y)|\epsilon_\phi, \epsilon_\psi] && (22)\\
&= \mathbb{E}_{\Phi,Y}[D(f^*(X), Y)|\epsilon_\phi, \epsilon_\psi] && (23)
\end{aligned}$$

Above, the inequality in the third line follows just from the optimality of the conditional expecation about the Bregman divergence, and the equality in the fourth line follows from the fact that $\Phi \in \mathcal{I}$.

All together we have

$$\mathbb{E}[D(f(X), Y)|\epsilon'] \geq [(D(f^*(X), Y)|\epsilon]. \tag{24}$$

□

Indeed, the condition in the claim 3.1 does not hold for just any arbitrary $\Phi \in \mathcal{I}$. In order for $g^*(\Phi)$ to be optimal in the sense of 3.1, Φ has to be special in some sense.

Proposition B.3. *Suppose $\Phi \in \mathcal{I}$ and suppose that ϵ satisfies $X \perp Y|\Phi, \epsilon$. Then for this particular ϵ,*

$$\Phi = \underset{Z \in \sigma(X)}{\arg\max}\, I(Y; Z|\epsilon).$$

Proof. Suppose that there exists $B \in \sigma(X)$ with $I(Y; B, \Phi|\epsilon) > I(Y; \Phi|\epsilon)$. Notice then that

$$I(Y; B, \Phi|\epsilon) = I(Y; \Phi|\epsilon) + I(Y; B|\Phi, \epsilon) \tag{25}$$

and it follows that $I(Y; B|\Phi, \epsilon) > 0$. Since $I(Y; X|\Phi, \epsilon) \geq I(Y; B|\Phi, \epsilon)$ this implies that $Y \not\perp X|\Phi, \epsilon$ and in particular, ϵ does not satisfy $X \perp Y|\Phi, \epsilon$. Thus, in order for $X \perp Y|\Phi, \epsilon$ to hold for ϵ, such B cannot exist. In other words, Φ must be optimal among all $Z \in \sigma(X)$ on any environment satisfying (3) . □

The necessary condition for (3) can be also more succinctly stated as follows.

Proposition B.4. *Suppose $\Phi \in \mathcal{I}$ and suppose that for some ϵ^*, $\Phi \perp Y|\epsilon^*$ Then there exists no $\tilde{\Phi} \in \mathcal{I}$ with $\Phi \in \sigma(\tilde{\Phi})$ such that $I(Y; \Phi) < I(Y; \tilde{\Phi})$.*

Proof. Let $\tilde{\Phi}$ be as stated in the assumption. Then since $g(X) \perp Y|\Phi, e^*$, it particularly follows that $\tilde{\Phi}(X) \perp Y|\epsilon^*$. Thus we have

$$\begin{aligned}
P(Y|\tilde{\Phi}) &= P(Y|\tilde{\Phi}, \epsilon^*) && (26)\\
&= P(Y|\tilde{\Phi}, \Phi, \epsilon^*) && (27)\\
&= P(Y|\Phi, \epsilon^*) && (28)\\
&= P(Y|\Phi) && (29)
\end{aligned}$$

Where the first and the last equalities follow from the invariance property, the second equality follows from the tower rule of conditional expectation, and the third equality follows from the assumption about ϵ^*. Note that both Φ and $\tilde{\Phi}$ are functions about X. Also, note that the mutual information depends only on conditional distribution. It follows that $I(Y; \tilde{\Phi}) = I(Y; \Phi)$, and the claim follows. □

Theorem B.5. *Suppose that there exists at least one Φ for which there is a corresponding ϵ_ψ for every ϵ_ϕ such that $X \perp Y|\Phi, \epsilon_\phi, \epsilon_\psi$. Suppose also that $\mathcal{I}$ is generated by one Φ_0. In this case $E[Y|\Phi^*]$ is OOD optimal if*

$$\Phi^* = \underset{\Phi \in \mathcal{I}}{\arg\max}\, I(Y; \Phi). \tag{30}$$

Proof. If $\tilde{\Phi} \in \mathcal{I}$ is such that for every ϵ_ϕ there is a corresponding ϵ_ψ, Φ is automatically OOD optimal, and it needs to be maximal by the proposition B.4. Now, if $\mathcal{I}$ is generated by one Φ_0, then Φ_0 is the unique maximal variable in the sense that it is equivalent to Φ^* up to its sigma field. Because $\tilde{\Phi} \in \sigma(\Phi_0)$, $\sigma(\Phi_0) = \sigma(\tilde{\Phi})$ necessarily by the maximality of $\tilde{\Phi}$ as well. We thus have $\sigma(\Phi_0) = \sigma(\tilde{\Phi}) = \sigma(\Phi^*)$. Thus, for Φ_0 there exists ϵ_ψ satisfying 3 for every ϵ_ϕ, and $E[Y|\Phi_0] = E[Y|\Phi^*]$ is OOD optimal. □

As one *explicit* example, that $\mathcal{I}$ is generated by one variable can happen when the underlying distribution P is UG-faithful, that is, P is faithful to the independence relations for which there exists some undirected graph G that induces a perfect map (Gasse, 2017; Lucas et al., 2007; Castillo et al., 2012). More succinctly stated, we have the following corollary.

Corollary B.6. *Suppose that there exists at least one Φ for which there is a corresponding ϵ_ψ for every ϵ_ϕ such that $X \perp Y|\Phi, \epsilon_\phi, \epsilon_\psi$, and suppose that P is UG-faithful to some undirected graph. Then MIP implies OOD optimality.*

Proof. If P is UG-faithful to some undirected graph, we know that strong union law of the conditional independence applies. That is, if $X \perp Y|\Phi_1, \mathcal{E}$ and $X \perp Y|\Phi_2, \mathcal{E}$, then $X \perp Y|\Phi_1, \Phi_2, \mathcal{E}$ as well. Let Φ^* be a MIP. If there exists another $\Phi \in \mathcal{I}$ such that $\Phi \notin \sigma(\Phi^*)$, then $\Phi_0 = [\Phi, \Phi^*]$ has larger mutual information than Φ^*. Also, by the strong union property, $\Phi_0 \in \mathcal{I}$. This contradicts the maximality of Φ^*, so $\mathcal{I} = \sigma(\Phi^*)$ necessarily, and the claim follows by the application of the theorem B. □

This corollary might be applicable to some DAG cases as well. For more detailed relation between DAG and undirected graph, see Bishop (2007). We shall also mention that, when P is UG-faithful to some undirected graph, the MIP is everything that excludes $\mathcal{E}$ and Y. As we emphasize over and over, we are considering the case in which the identity of $\mathcal{E}$ as well as the correlation amongst the covariates and variates, so even if it is known that P is faithful to *some* undirected graph, finding MIP might be better strategy in such case.

The following is another sufficient condition for the OOD optimality that can hold if we can make a slightly stronger assumption about $\mathcal{E}$.

Corollary B.7. *Suppose that $\mathcal{E}$ admits an independence decomposition $(\mathcal{E}_\phi, \mathcal{E}_\psi)$ such that $\mathcal{E}_\psi = \mathcal{E}_\phi^c$ and $\mathcal{E}_\phi \perp \Psi$. If there exists one $\tilde{e}_\psi$ for which $\Psi \perp Y|\tilde{e}_\psi$, then $\mathbb{E}[Y|\Phi]$ is OOD optimal.*

Proof. Since $\Phi \perp (\Psi, \mathcal{E}_\psi)$, we have $(\Phi \perp \Psi)|\mathcal{E}_\psi$ because

$$P(\Phi|\Psi, \mathcal{E}_\psi) = P(\Phi) = P(\Phi|\mathcal{E}_\psi) \tag{31}$$

Likewise, by the assumption we have $\mathcal{E}_\phi \perp (\Psi, \mathcal{E}_\psi)$, so we have $(\mathcal{E}_\phi \perp \Psi)|\mathcal{E}_\psi$. Now, we also have the assumption that $\Psi \perp Y|\tilde{e}_\psi$. Altogether we have $(\Psi \perp (Y, \mathcal{E}_\phi, \Phi))|\tilde{e}_\psi$. This tells us that

$$P(\Psi|Y, \mathcal{E}_\phi, \Phi, \tilde{e}_\psi) = P(\Psi|\tilde{e}_\psi) = P(\Psi|Y, \tilde{e}_\psi) \tag{32}$$

This is to say that $\Psi \perp Y|\Phi, \mathcal{E}_\phi, \tilde{\epsilon}_\psi$, and the claim follows by the application of 3.1. □

Wa also want to make a comment regarding the choice of $\mathcal{E}_\psi$.

Remark B.8. *Making $\mathcal{E}_\psi$ maximal in the assumption (3) in general cases the difficulty of satisfying (3). To see this, suppose that $\mathcal{E}_\psi$ is not maximal, and that there exists $\tilde{\mathcal{E}_\psi}$ with the corresponding complement $\tilde{\mathcal{E}_\phi}$ for which $\sigma(\tilde{\mathcal{E}_\psi}) \subsetneq \mathcal{E}_\psi$. Now, again in the way of Darmois (1953), let $\mathcal{E}'_\psi \perp \mathcal{E}_\psi$ be such that $\sigma(\mathcal{E}'_\psi, \mathcal{E}_\psi) = \sigma(\tilde{\mathcal{E}_\psi})$. This way we can let the triplet $(\tilde{\mathcal{E}_\phi}, \mathcal{E}'_\psi, \mathcal{E}_\psi)$ to represent e, and let the pair $(\tilde{\mathcal{E}_\phi}, \mathcal{E}'_\psi)$ serve as a $\mathcal{E}_\phi$. Whenever one can say that, for all ϵ_ϕ there exists ϵ_ψ satisfying (3), I can also say as well that, for all $\tilde{e}_\phi$ there exists $\tilde{e}_\psi = (\epsilon'_\psi, \epsilon_\psi)$ for which (3) holds. This is clear by construction. However, the reverse is not true in general. Even if for $\tilde{\epsilon}_\phi$ there exists $\tilde{\epsilon}_\psi$ for which (3) holds, it is not necessarily true that, for all ϵ_ϕ there exists ϵ_ψ satisfying (3). To see this, suppose that, for a given $\tilde{\epsilon}_\phi$, there is a unique $(\epsilon'^*_\psi, \epsilon^*_\psi)$ such that (3) holds. In this case, for any $\epsilon_\phi = (\tilde{\epsilon}_\phi, \epsilon'_\psi)$ with $\epsilon'_\psi \neq \epsilon'^*_\psi$, there is no ϵ_ψ for which (3) holds.*

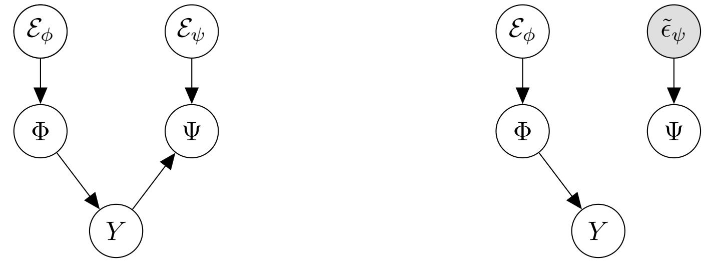


Figure 8: (Left) A rough schematic view of the condition assumed in the corollary B.7, interpreted in the language of graphical model. (Right) When there is a $\tilde{\epsilon}_\psi$ that breaks the edge from Y to Ψ, the conditional expectation $\mathbb{E}[Y|\Phi]$ will be OOD optimal.

C More details about Method

C.1 Difficulties in finding the MIP

At a first glance, the concept we presented in the last section seems to require two steps: (i) the search for the solution Φ^* of (30) and (ii) the computation of $\mathbb{E}[Y|\Phi^*]$, which is a function of Φ^*. However, finding Φ^* is difficult on its own because $I(Y;\Phi)$ requires $p(Y|\Phi)$(density) for its evaluation. Without reasonably accurate knowledge about the density of X and Y, it is hard to compute (30), let alone the condition for Φ's invariance.

Classic constrained optimization methods often use regularization terms to enforce the constraint. Let us use Q to approximate P, and let us parametrize Q by ξ and Φ by η. We can interpret our problem (30) as the optimization of

$$\begin{aligned}&\arg\min_{\xi,\eta}\ \mathbb{E}[d_{KL}[p(Y|X)\|p(Y|\Phi_\eta)]] \\ &+\mathbb{E}[d_{KL}[p(Y|\Phi_\eta)\|q_\xi(Y|\Phi_\eta)]]+\lambda I_q(Y;\mathcal{E}|\Phi_\eta)\end{aligned} \tag{33}$$

about both Φ and q. The last regularization term encourages Φ to be in $\mathcal{I}$, and the second term encourages q to be close to p. The first term encourages Φ to be closely correlated with Y. Now, $I_q(Y,\mathcal{E}|\Phi_\eta)$ can be approximated by

$$\mathbb{E}[d_{KL}(q_\xi(Y|\Phi_\eta,\mathcal{E})\|q_\xi(Y|\Phi_\eta))] \tag{34}$$

and it also follows that

$$\begin{aligned}&\mathbb{E}[d_{KL}[p(Y|X)\|q_\xi(Y|\Phi_\eta)]] \\ &=\mathbb{E}[d_{KL}[p(Y|X)\|p(Y|\Phi_\eta)]] \\ &+\mathbb{E}[d_{KL}[p(Y|\Phi_\eta)\|q_\xi(Y|\Phi_\eta)]].\end{aligned} \tag{35}$$

With the understanding that the pair of the parameters (ξ,η) represents a function, let us write $d_{KL}[p(Y|X)\|q_\xi(Y|\Phi_\eta)]$ as $L_e(\xi,\eta)$. We can think of (33) as the minimizer of the following equation about ξ,η;

$$\mathbb{E}[L_\mathcal{E}(\xi,\eta)]+\lambda\mathbb{E}[d_{KL}\left(q_\xi(Y|\Phi_\eta,\mathcal{E})\|q_\xi(Y|\Phi_\eta)\right)]. \tag{36}$$

However, we encounter another problem yet again. In general, the form of $q(y|\phi)$ as a function of ϕ depends on the choice of ϕ [3]. This problem is subtle but important. For instance, if $q_\xi(y|\phi_\eta)$ is modeled as $f(y,x,\xi,\eta)$, then $f(y,x,\xi,\eta')$ for $\eta'\neq\eta$ is generally not equal to $q_\xi(y|\phi_{\eta'})$. Rather, it will most likely equal $\tilde{q}_\xi(y|\phi_{\eta'})$ for completely other $\tilde{q}_\xi$. To see this, note that $q_\xi(y|\phi_{\eta'})$ is solution about the optimization of $\arg\min_r d_{KL}[q_\xi(y|\phi_\eta)\|r(y|\phi_{\eta'})]$ about the density r, and it is clear that the optimal r depends on η. Thus, the choice of η and the choice of ξ are not independent, and the simultaneous update of ξ and η can be nonsensical.

While aiming for different invariant feature, IRM (Arjovsky et al., 2019) too encountered a similar problem and endeavored to skirt this problem by assuming that $\mathbb{E}[Y|\Phi]$ is linear with respect to Φ, and modeled $\mathbb{E}[Y|\Phi]=\langle\mathbf{1},\Phi\rangle$ by requiring Φ to absorb all linear transformation. However, it is unlikely that $\mathbb{E}[Y|\Phi]$ takes the same form throughout the algorithm in search of a good Φ. The same things can be said to $P(Y\in A|\Phi)$ for any A, which is, in essence, $\mathbb{E}[1_A(Y)|Z]$.

[3] Recall that ϕ is a symbol representing the realization of Φ

C.2 Penalty derivation

In this section, we provide the derivation of the penalty term we omitted in the section 4.2. For the notations, please take a look at the main manuscript.

$$I(Y;\mathcal{E}|\Phi) \cong I_q(Y;\mathcal{E}|\Phi) = \mathbb{E}[d_{KL}(q_\theta(Y|X,\mathcal{E})\|q_\theta(Y|X))] \tag{37}$$

$$\cong \mathbb{E}[\log q_\theta(Y|X,\mathcal{E}) - \log q_\theta(Y|X)] \tag{38}$$

$$= \mathbb{E}[L_\mathcal{E}(\theta - \alpha\nabla_\theta \mathbb{E}[L_\mathcal{E}(\theta)]) - L_\mathcal{E}(\theta - \alpha\nabla_\theta L_\mathcal{E}(\theta))] \tag{39}$$

$$\cong \alpha(\mathbb{E}[\nabla_\theta L_\mathcal{E}(\theta)^T \nabla_\theta L_\mathcal{E}(\theta)] - \mathbb{E}[\nabla_\theta L_\mathcal{E}(\theta)]^T \mathbb{E}[\nabla_\theta L_\mathcal{E}(\theta)]) \tag{40}$$

$$= \alpha \text{ trace}(\text{Var}(\nabla_\theta L_\mathcal{E}(\theta)) \tag{41}$$

Also, in our implementation, we took advantage of the smallness of α to approximate $q(y|x;\theta)$ in the first term with $f(y|x;\theta)$ instead of $f(y|x;\theta - \alpha\nabla_\theta \mathbb{E}_e[L_e(\theta)])$. In fact, faithfully computing the first term with $f(y|x;\theta - \alpha\nabla_\theta \mathbb{E}[L_\mathcal{E}(\theta)])$ did not make much difference in the training process.

D Implementation Detail

In this section we describe the details of the experiment design along with the architectures of the models we used. In order to present a self-contained material, we first restate the experimental setting we already described in the main manuscript.

D.1 Colored MNIST

Colored MNIST is an experiment that was used in (Arjovsky et al., 2019). The goal of the task in *Colored MNIST* is to predict the label of a given digit in the presence of varying exterior factor, $\mathcal{E}$. The left panel of the figure 10 is a Bayesian Network representation of this experiment. Each member of the *Colored MNIST* dataset is constructed from an image-label pair (x, y) in MNIST, as follows.

1. Assign a binary label $\hat{y}_{obs}$ from y with the following rule: $\hat{y}_{obs} = 0$ if $y \in \{0 \sim 4\}$ and $\hat{y}_{obs} = 1$ otherwise.
2. Flip $\hat{y}_{obs}$ with a fixed probability p to produce y_{obs}.
3. Let x_{fig} be the binary image corresponding to y.
4. Put $y_{obs} = \hat{x}_{ch1}$, and construct x_{ch1} from $\hat{x}_{ch}$ by flipping $\hat{x}_{ch1}$ with probability e.
5. Construct $x_{ods} = x_{fig} \times [x_{ch0}, (1-x_{ch0}), 0]$.(that is, red if $x_{ch1} = 1$ and green if $x_{ch1} = 0$.) Indeed, x_{obs} has exactly same information as the pair (x_{fig}, x_{ch1}).

In this experiment, only (Y_{obs}, X_{obs}) are assumed observable. At training times, the machine learner will be given a set of datasets $\mathcal{D}_{train} = \{D_e; e \in R_{train}\}$ in which D_e is a set of observations gathered when $\mathcal{E} = e$. We set $|R_{train}| = 2$, and choose $|D_e| = 25000$. More particularly, for the e_1 we chose the flip-rate(p) to be 0.1, and chose $p = 0.2$ for the e_2. Each image was resized to 14×14 resolution.

For the test evaluation, we randomly sampled 10 instances of p uniformly from the range $[0, 1]$ to construct R_{test}, and approximated the OOD accuracy by computing the worst performance over all R_{test} . We used 5 seeds to produce each numerical result. For the model, we used 4 Layers MLP with 2500 units per each layer and elu activation(Clevert et al., 2015), and did not use bias term in the last sigmoid activation. We used batch normalization (BN)(Ioffe & Szegedy, 2015) for each layer, and optimized the model using Adam(Kingma & Ba, 2014) with alpha = 0.0015, beta1=0.0, beta2=0.9 over 500 iterations. In general, less number of iterations yielded better results when $|\mathcal{R}_{train}|$ was small (less overfitting).

D.2 Extended Colored MNIST

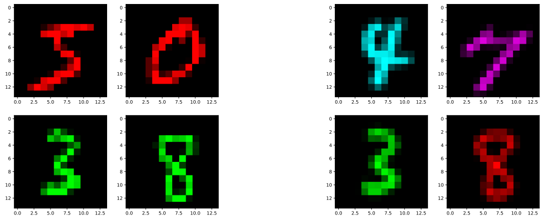
As described in the main manuscript, Extended Colored MNIST is a modified version of colored MNIST, in which the dataset was constructed using the following procedure. The right panel of the figure 10 is a Bayesian Network representation of this experiment.

1. Set x_{ch2} to 1 with probability e_{ch2}. Set it to 0 with probability $1 - e_{ch2}$.
2. Construct $\hat{y}_{obs}$ in the same way as in *Colored MNIST*. If $e_{ch2} = k$, construct y_{obs} by flipping $\hat{y}_{obs}$ with probability $p_k(k \in \{0, 1\}$.)
3. Put $y_{obs} = \hat{x}_{ch0}$, and construct x_{ch0} from $\hat{x}_{ch}$ by flipping $\hat{x}_{ch1}$ with probability e_{ch0}.
4. Construct x_{obs} as $x_{fig} \times [x_{ch0}, (1 - x_{ch0}), x_{ch2}]$. As an RGB image, this will come out as an image in which the red scale is *turned on* and the green scale is *turned off* if $x_{ch0} = 1$, and otherway around if $x_{ch0} = 0$. Blue scale is turned-on only if $x_{ch2} = 1$.

In this experiment, we set $|R_{train}| = 5$, and choose $|D_e| = 10000$, and resized each image in the dataset to 14×14 resolution. To produce $e \in R_{train}$, we selected e_{ch0} randomly from the range $[0.1, 0.2]$, and selected e_{ch2} randomly from the range $[0.3, 0.4]$.

Mean while, we set $|R_{test}| = 9$. To produce n-th member of R_{test}, we set $e_{ch0} = 0.1$ and we selected e_{ch2} randomly from the range $[0.0, 1.0]$. We chose $p_0 = 0.25$, $p_1 = 0.75$ for both R_{test} and R_{train}.

We used 5 seeds to produce each numerical result. For the model, we used 4 Layers MLP with 2500 units per each layer, and did not use bias term in the last sigmoid activation. We used batch normalization for each layer, and optimized the model using Adam with alpha = 0.0005, beta1=0.0, beta2=0.9 over 2000 iterations. The performance-values of IRM in the Table 3 of the main article are the results produced by the model that achieved the best average ***train*** accuracy among all models trained with $\lambda > 10^4$. The averages were computed over 5 seeds.



(a) Example images of *Colored MNIST*. Only two channels are used for all images, and the colors are flipped randomly by the exterior factor.

(b) Example images of *Extended Colored MNIST*. The first two channels and the third channel are perturbed by the different mechanism. See the main manuscript for the way of the construction.

Figure 9: Example of *Colored MNIST* and *Extended Colored MNIST*

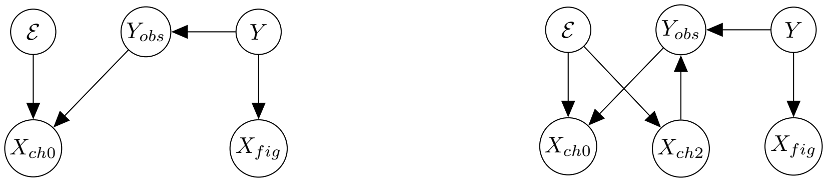
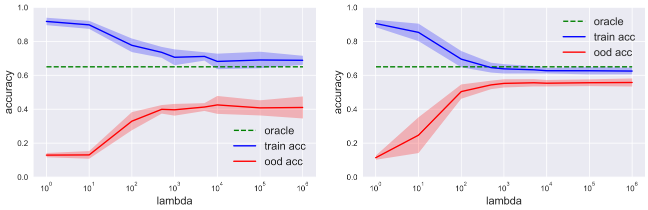


Figure 10: The graphical model of *Colored MNIST*(left) and *Extended Colored MNIST*(right)

E Additional Result

E.1 More experimental results on C-MNIST and EC-MNIST

The result of *Extended Colored Mnist* with $p_0 = 0.25$, $p_1 = 0.65$ for both R_{test} and R_{train}. Our algorithm outperforms the Invariant Risk Minimization(IRM)(Arjovsky et al., 2019) in this case as well in Figure 11.



(a) IRM

(b) IGA

Figure 11: Result on *Extended Colored MNIST* ($p_0 = 0.25, p_1 = 0.65.$)

In general, IRM does not work well with standard gradient descent when we implement MLP without Batch Normalization (Figure 12).

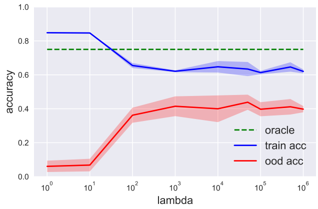
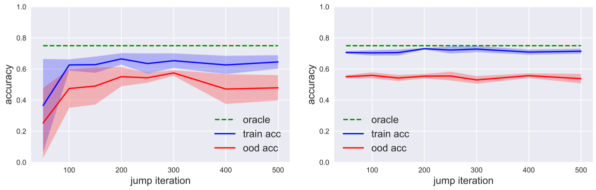


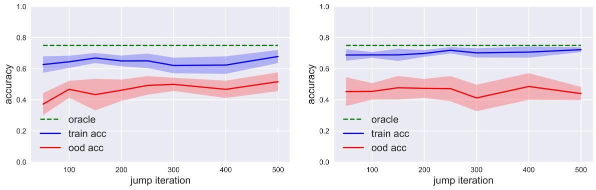
Figure 12: Result of IRM on *Colored MNIST* with MLP without BN.

We shall note that the original implementation of the IRM published in Github (`https://github.com/facebookresearch/InvariantRiskMinimization`) uses a very specific schedule for the regularization parameter λ, and it makes λ to jump to a very large value at a very specific timing that is difficult to identify from the training set. The following figures are the result of their original algorithm on *Colored MNIST* and *Extended Colored MNIST* implemented with various jump-timings of λ. For *Colored MNIST*, the original IRM works for specific choices of the jump timing($200 \sim 300$). For *Extended Colored MNIST*, the original algorithm does not work too well for any choice of the jump timings. Meanwhile, IRM works relatively well on *Colored MNIST* consistently if we apply batch normalization, and it works well even without "jumping" the λ. For the tables we present in the main manuscript, we reported the result of IRM implemented *with* batch normalization, which consistently yielded better results than the original implementation.



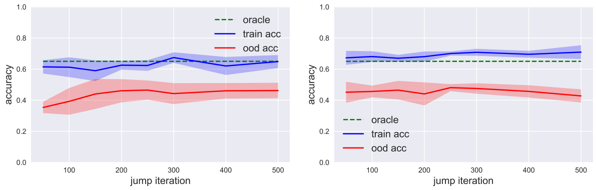
(a) MLP (b) MLP with BN

Figure 13: The plot of jump timing against the accuracies on *Colored MNIST*



(a) MLP (b) MLP with BN

Figure 14: The plot of jump timing of IRM against the accuracies on *Extended Colored MNIST*.



(a) MLP (b) MLP with BN

Figure 15: The plot of jump timing of IRM against the accuracies on *Extended Colored MNIST*.

E.2 Ablation study for two phase training with nonlinear predictor

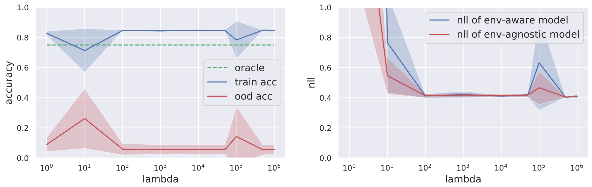
IRM is a type of two-phase training 4 that uses linear predictor $g : \Phi \to Y$. Meanwhile, IR (Chang et al., 2020) uses another variant of two-phase training in which non-linear $g : \Phi \to Y$ is considered. If $\mathcal{C}(f|\epsilon)$ is the loss of the function $f : X \to Y$ on the environment ϵ, (Chang et al., 2020) optimize the following objective:

$$\arg\min_{g,h} \max_{g*_\epsilon} \left\{ \mathbb{E}[\mathcal{C}(g \circ h|\mathcal{E})] + \lambda a(\mathbb{E}[\mathcal{C}(g \circ h|\mathcal{E})] - \mathbb{E}[\mathcal{C}(g^*_{\mathcal{E}} \circ h|\mathcal{E})]) \right\} \tag{42}$$

where $a(t)$ is a convex function that is monotonically increasing in t when $t < 0$, and strictly monotonically increasing in t when $t \geq 0$, $g^*_\epsilon = \arg\min \mathcal{C}(f|\epsilon)$ for each ϵ, and λ is the regularization parameter. We use such monotonic function because in true expectation with respect to X and Y, $C(g \circ h \| \epsilon) > C(g^*_h \circ h \| \epsilon)$ by the optimiality of g^*_h. We conducted this optimzation on both C-MNIST and EC-MNIST, and studied the relation between λ and the final accuracy as well as the value of the regularization term. Unlike in the original that searches h of the form $M \circ X$ with binary mask M, we searched h over the space of black-box function represented by MLP.

We used MLP for both g and h. For g, we used MLP with 4 layers containing 1500 nodes each and activation function elu and did not use bias term in the last sigmoid activation. For h, we used MLP with 4 layers containing 1500 nodes each and activation function elu and did not use bias term in the last sigmoid activation. As is done in both Arjovsky et al. (2019) and Chang et al. (2020), we optimized both models in parallel without propagating the loss of g_h to h. For both C-MNIST and EC-MNIST, we evaluated the model performance in the same way as in the IRM experiments.

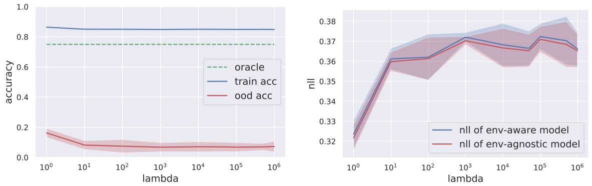
As we see in the plots below, even when the environment agnostic loss $\mathcal{C}(g \circ h|\mathcal{E})$ is very close to environment aware loss $\mathcal{C}(g^*_\epsilon \circ h|\mathcal{E})$, the the performance on the training environments does not generalize to all environments. This tendency was observed irrespective of the presence of Batch normalization. This is possibly because true g^*_ϵ is not estimated well in the training process because of the inter-dependency between g and h.



(a) accuracy

(b) The final nll of environment agnostic predictor and environment aware predictor

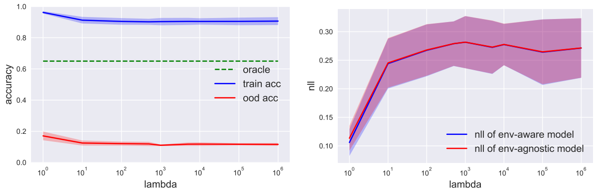
Figure 16: The two phase training results for *Colored MNIST* with MLP encoder and MLP predictor



(a) accuracy

(b) The final nll of environment agnostic predictor and environment aware predictor

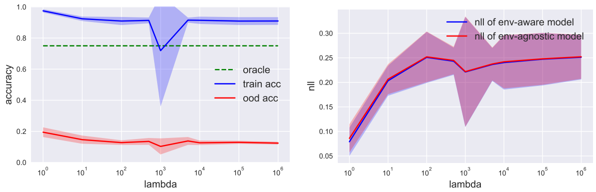
Figure 17: The two phase training results for *Colored MNIST* with MLP + BN encoder and MLP + BN predictor



(a) accuracy

(b) The final nll of environment agnostic predictor and environment aware predictor

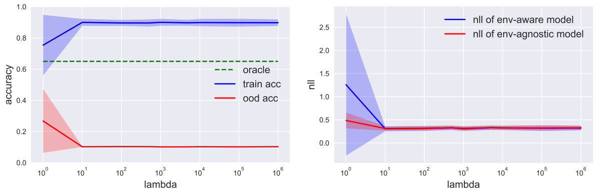
Figure 18: The two phase training results for *Extended Colored MNIST*$(p_0 = 0.25, p_1 = 0.65)$ with MLP + BN encoder and MLP + BN predictor



(a) accuracy

(b) The final nll of environment agnostic predictor and environment aware predictor

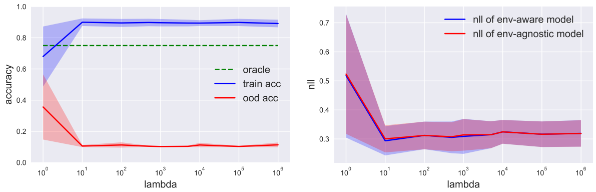
Figure 19: The two phase training results for *Extended Colored MNIST*$(p_0 = 0.25, p_1 = 0.75)$ with MLP + BN encoder and MLP + BN predictor



(a) accuracy

(b) The final nll of environment agnostic predictor and environment aware predictor

Figure 20: The two phase training results for *Extended Colored MNIST*$(p_0 = 0.25, p_1 = 0.65)$ with MLP encoder and MLP predictor



(a) accuracy

(b) The final nll of environment agnostic predictor and environment aware predictor

Figure 21: The two phase training results for *Extended Colored MNIST*$(p_0 = 0.25, p_1 = 0.75)$ with MLP encoder and MLP predictor