ZIN: When and How to Learn Invariance by Environment Inference?

Anonymous Author(s)
Affiliation
Address
email

Abstract

It is commonplace to encounter heterogeneous data, of which some aspects of the data distribution may vary but the underlying causal mechanisms remain constant. When data are divided into distinct environments according to the heterogeneity, recent invariant learning methods have proposed to learn robust and invariant models based on this environment partition. It is hence tempting to utilize the inherent heterogeneity even when environment partition is not provided. Unfortunately, in this work, we show that learning invariant features under this circumstance is fundamentally impossible without further inductive biases or additional information. Then, we propose a framework to jointly learn environment partition and invariant representation, assisted by additional auxiliary information. We derive sufficient and necessary conditions for our framework to provably identify invariant features under a fairly general setting. Experimental results on both synthetic and real world datasets validate our analysis and demonstrate an improved performance of the proposed framework. As a byproduct, our findings also raise the need of making the role of inductive biases more explicit in future works, when learning invariant models without environment partition.

1 Introduction

Machine learning algorithms with empirical risk minimization (ERM) generally assume independent and identically distributed (i.i.d.) data in training and test sets. Due to changing circumstances, selection bias, or time shifts, data distributions are often heterogeneous across different environments in practical applications. When the distributions of training and test data are indeed different, model performance can severely degrade [44, 15, 35], as ERM based algorithms may exploit the spurious correlations that are particularly useful to ERM in the training data but do not exist in the test data.

To tackle the distribution-shift or out-of-distribution generalization problem, a recent line of methods propose to utilize the causally invariant mechanisms (rather than the spurious correlations in the training data) that are supposed to be stable across different environments. For example, Peters et al. [33] propose to exploit the invariance principle to learn a linear model that merely relies on the direct causes of the target. Such a model is shown to be robust to potential interventions on any variable except the target itself. Arjovsky et al. [5] propose invariant risk minimization (IRM) to capture invariant correlations by learning representations that elicit an optimal invariant predictor across multiple training environments. Ahuja et al. [3], Jin et al. [19], Krueger et al. [22], Xie et al. [48] further develop variants of IRM by introducing game theory, regret minimization, variance penalization, etc., and Xu and Jaakkola [50], Chang et al. [8], Lin et al. [24] try to learn invariant features by coupled adversarial neural networks.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Noticeably, the aforementioned invariant learning methods require datasets to be explicitly partitioned into environments (or domains) according to the heterogeneity underlying the data. However, such an environment partition may be unavailable or hard to obtain in practice [26, 9]. Recently, a line of works to learn invariance with no environment indexes, i.e., the dataset is assembled by merging data from multiple environments. For instance, Creager et al. [9] propose environmental inference for invariant learning (EIIL), a two-stage method by firstly inferring the environments with an ERM based trained biased model and then performing invariant learning on inferred environments. Liu et al. [26] devise an interactive mechanism, called heterogeneous risk minimization (HRM), with environment inference and invariant learning on raw feature level, and Liu et al. [27] extend it to representation level with kernelized trick.

However, as we will show, learning invariant models under this circumstance (with only input-label pairs) can be *risky*, both empirically and theoretically. As our first contribution, in Section 4, we prove that learning invariant features from heterogeneous data without environment indexes is in theory impossible. Specifically, we provide a counter-example (Example 1) and a general theoretic result (Theorem 1), to show that the invariant features are unidentifiable if no environment information is provided. This impossibility result, similar to the identifiability issue in causal discovery [42, 34], independent component analysis (ICA) [18], and unsupervised learning of disentangled representations [29], motivates us to consider when and how to learn causally invariant features.

In this work, we turn to additional auxiliary variables that encode some information about the latent heterogeneity. This strategy can be treated as a case between the two existing directions: 1) it is less restrictive than requiring ideal environment partition according to distribution heterogeneity; and 2) it also obtains theoretic guarantee which is missing in the case with no environment information at all. Notably, such auxiliary information is often cheaply available for every input in practice [49, 47]. Examples include time index of the data in time series forecasting tasks [7, 30], locations (longitude and latitude) of collected satellite data in remote sensing [13, 37], and meta information associated with an instance [28]. Based on the additionally observed variables, we proceed to propose a framework to jointly learn environment partition and invariant representation in Section 5. Under a fairly general setting, we derive both sufficient and necessary conditions for our framework to identify invariant features. Extensions to other settings are discussed in Section 6 and Appendix B.

Finally, Section 7 presents experimental analysis on both synthetic and real world datasets. The proposed framework achieves an improved performance over existing methods like EIIL and HRM, and has a comparable performance to IRM with ground-truth environment partition.

2 Related Work

Invariant learning methods can be interpreted as robust to certain interventions or heterogeneity from a causal perspective [33]. On the other hand, heterogeneous data have also placed challenges and potential benefits in causal inference tasks. For example, Huang et al. [17] consider heterogeneous datasets to identify variables with changing local mechanism and further the causal directions. Here confounders are assumed as functions of the domain index or time. Tillman and Spirtes [45], Danks et al. [10], Huang et al. [16] study the multiple dataset setting with non-identical sets of variables for causal discovery, where the datasets may have different exogenous noise distributions and only part of the variables is present in each dataset.

Besides the invariant learning methods described in Section 1, another large class of methods for generalizing beyond training data is distributionally robust optimization (DRO) [6, 40, 23, 12, 11, 51]. DRO methods propose to optimize the worst-case risk over a set of distributions close to the training distribution. A notable instance of DRO methods is group DRO, which optimizes the worst-case loss over groups in the training data [38]. As with most invariant learning methods, group DRO generally requires a partition of groups obtained according to group annotation and label of each data sample.

Obtaining group annotations, however, can be costly or even impossible in practice. We may not know *a priori* the inherent spurious features associated with data. Thus, a number of works have proposed to infer group partition or directly identify the minority group by looking at features produced by

biased models, e.g., [41, 2, 9, 25]. Notice that sometimes a small number of labelled annotations are still needed to find a proper trained model, which is different from the setting considered in this work. For example, Liu et al. [25] upweights samples with high losses from the initial ERM model and relies on a small validation set of annotated data to tune parameters. Additionally, Nam et al. [31], Sanh et al. [39] try to obtain a more robust model by boosting from the wrongly specified samples, based on the observation that models with limited capacity tend to learn spurious correlations or shortcuts. Unfortunately, as we will show in Section 4, it is in theory impossible to distinguish spurious or invariant features from only the observed data consisting of input-label pairs.

3 Preliminaries

Throughout this paper, upper-cased letters such as $\boldsymbol{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ denote random variables, and lower-cased letters such as $\boldsymbol{x}$ and y denote deterministic instances. Suppose there exist invariant features $\boldsymbol{X}_v \in \mathbb{R}^{d_v}$ and spurious (non-invariant) features $\boldsymbol{X}_s \in \mathbb{R}^{d_s}$. We observe a scrambled version $\boldsymbol{X} = q(\boldsymbol{X}_v, \boldsymbol{X}_s) \in \mathbb{R}^d$ with q being an injective function. Consider the dataset $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where data are collected from multiple environments $e \in \mathcal{E}_{supp}$. We use superscript e to indicate the environment index with a variable, e.g., $\boldsymbol{X}^e$ and Y^e, when it is useful to make the index explicit. Notice that the true environment index e is not provided in training, unless otherwise stated.

In this work, we assume an underlying structural causal model (SCM) governing the data generation process [32, 34]. An SCM considers a set of variables associated with vertices of a directed acyclic graph, where directed edges represent direct causation. Each variable is obtained as a result of an assignment of a deterministic function depending on the parental variables in the graph and an exogenous random variable. As an example, we present the following data generation process that are also considered in [4, 5]:

$$Y^e = g_v(\boldsymbol{X}_v^e, \epsilon_v), \;\; \boldsymbol{X}_v^e \perp\!\!\!\perp \epsilon_v; \qquad \boldsymbol{X}^e = q(\boldsymbol{X}_v^e, \boldsymbol{X}_s^e), \tag{1}$$

where g_v denote some non-degenerate deterministic functions and ϵ_v is independent noise variables. We observe data $\{\boldsymbol{X}^e, Y^e\}_{e=1}^E$ from multiple environments, each with probability $P(\boldsymbol{X}^e, Y^e)$. Then the marginal distribution of the mixed data can be represented as $P(\boldsymbol{X}, Y) = \sum_{e=1}^E \alpha_e P(\boldsymbol{X}^e, Y^e)$ for some $\alpha_e > 0$ and $\sum_{e=1}^E \alpha_e = 1$.

We consider the model as the composition of a feature extractor Φ and a classifier (or predictor) f_ω that is parameterized by ω. To seek for out-of-distribution generalization ability, we hope that Φ merely encodes the information of invariant features. Let $\ell(\cdot, \cdot)$ denote a loss function such as cross entropy loss or squared error. Our goal is to learn a robust and invariant model that minimizes the following objective over all the considered environments $e \in \mathcal{E}_{supp}$:

$$\sup_{e \in \mathcal{E}_{supp}} \mathbb{E}_{\boldsymbol{X}^e, Y^e} \left[\ell\left(f_w(\Phi(\boldsymbol{X}^e)), Y^e\right) \right].$$

While there is no single, common definition of *invariant feature* in the literature, we treat the direct causal parents of Y as invariant features, as in [33, 5]. This is because the conditional probability of the outcome Y given its parental variables or direct causes remains unchanged with any intervention (not on Y). In contrast, if a non-parental variable is included in the conditioned variables, then the conditional probability could change after some intervention. As such, the underlying SCM governing data generation determines the desired invariance.

4 Impossibility Result

The first question we ask is whether it is possible to learn invariant models from the heterogeneous data of multiple environments with unknown environmental indexes. The following toy example immediately gives a negative answer.

Example 1. Consider that there are two environments $e \in \{1, 2\}$ with $\alpha_1 = \alpha_2 = 0.5$, and we observe binary-valued features X_1, X_2 and label Y. Suppose the joint distribution of mixed environments is given in Eqn. (2). A learning algorithm that tries to learn invariant feature based on this dataset would result in a model that (deterministically) depends on either X_1, X_2, or both to predict Y.

However, as discussed in last section, whether the learned features are invariant is determined by the underlying data generation process. We now present two possible data generation processes that generate the same distribution yet have different invariant features:

- X_1 is the invariant feature while X_2 is spurious, where $p_s^{e=1} = 0.7$ and $p_s^{e=2} = 0.9$:

$$X_1 \sim \text{Bernoulli}(0.5), \quad Y = \begin{cases} X_1, & w.p.\ 0.5, \\ 1 - X_1, & w.p.\ 0.5, \end{cases} \qquad X_2^e = \begin{cases} Y, & w.p.\ p_s^e, \\ 1 - Y, & w.p.\ 1 - p_s^e, \end{cases}$$

- X_2 is the invariant feature while X_1 is spurious where $p_s^{e=1} = 0.25$ and $p_s^{e=2} = 0.75$:

$$X_2 \sim \text{Bernoulli}(0.5), \quad Y = \begin{cases} X_2, & w.p.\ 0.8, \\ 1 - X_2, & w.p.\ 0.2, \end{cases} \qquad X_1^e = \begin{cases} Y, & w.p.\ p_s^e, \\ 1 - Y, & w.p.\ 1 - p_s^e, \end{cases}$$

$$\begin{cases} Y = 0, & w.p.\ 0.5, \\ Y = 1, & w.p.\ 0.5, \\ X_2 = X_1 = Y, & w.p.\ 0.4, \\ X_2 \neq X_1 = Y, & w.p.\ 0.1, \\ X_1 \neq X_2 = Y, & w.p.\ 0.4, \\ X_2 = X_1 \neq Y, & w.p.\ 0.1. \end{cases} \tag{2}$$

The joint distribution of mixed environments for each case is consistent with Eqn. (2). Thus, from the joint distribution $P(X_1, X_2, Y)$, a learned model only depends on either X_1, X_2 or both, and fails to generalize for at least one of the two data generation processes. Thus, the true invariant features cannot be identified. On the other hand, notice that when the partition is given, one can verify that IRM would correctly identify X_1 (resp. X_2) as the invariant feature for the first (resp. second) senario.

Moreover, such an example is not rare. Indeed, for any data distribution $P(\boldsymbol{X}, Y)$ that can be generated by some data generation process like (1) (which can be replaced by other forms of SCM), we can find a different data generation process that induces the same distribution yet has different invariant/spurious features. Similarly, it is impossible to find the minority group that is determined by spurious features, based on only the data $P(\boldsymbol{X}, Y)$. This result is formally summarized in Theorem 1.

Theorem 1. *Let $\boldsymbol{X}_v$ and $\boldsymbol{X}_s$ be respectively invariant and spurious features, with label Y. For the joint distribution $P(\boldsymbol{X}, Y)$ consisting of data from multiple environments with $P(\boldsymbol{X}_s, Y) \neq 0$, we can find $\boldsymbol{X}_v'$ and $\boldsymbol{X}_s'$ as invariant and spurious features so that: (a) $\boldsymbol{X}_v' \neq \boldsymbol{X}_v$ and $\boldsymbol{X}_v' \cap \boldsymbol{X}_s \neq \varnothing$; (b) $(\boldsymbol{X}_v', \boldsymbol{X}_s')$ together with some noise variables generate the same distribution $P(\boldsymbol{X}, Y)$.*

A proof is provided in Appendix A.1. Under this circumstance, we have to introduce additional assumptions/conditions or certain "inductive bias" to identify the invariant features. The latter may come implicitly from the proposed algorithm, model, and/or the data. We believe that this is the reason that makes existing methods achieve improved empirical performance in their considered scenarios. For example, in [26, 27], the discrepancy of spurious features among clusters are expected to be larger than that of causal features, and in [9, 31, 39, 25], the ERM model of the first stage should heavily or fully rely on the spurious feature. However, these inductive biases may not be always guaranteed. Thus, as suggested by Theorem 1, the role of inductive biases shall be discussed more explicitly when learning invariant models from heterogeneous data without environment indexes.

Alternatively, in this work, we consider that there exist additionally observed variable $\boldsymbol{Z}$ with the data, which has been often considered in the literature [17, 18, 20, 49, 47]. The variable $\boldsymbol{Z}$ could be, for example, the time index in a time series, some kind of class labels, concurrently observed variables, or human annotations on potential unmeasured variable [43]. It is worth noting that Xie et al. [49] also take advantage of auxiliary information to help improve OOD performance in a semi-supervised manner. However, they consider a different setting with access to unlabeled (out-of-domain) test data together with additional unlabeled training data. In the next section, we show how to utilize $\boldsymbol{Z}$ to learn invariant models from the mixed dataset.

5 ZIN: Learning Invariance with Additional Auxiliary Information

This section proposes our framework to learn invariant models. We derive conditions for invariance identification and show these conditions are both sufficient and necessary in a fairly general setting.

5.1 Method

We consider that there exists additional auxiliary information $\boldsymbol{Z} \in \mathbb{R}^{d_z}$ in company with the data $(\boldsymbol{X}, Y)$. In the following, we propose ZIN, *auxiliary information* $\boldsymbol{Z}$ *for environmental INference*, for invariant learning from the mixture dataset $\mathcal{D}$ without environment indexes.

We aim to learn a function $\rho(\cdot) : \mathbb{R}^{d_z} \to \mathbb{R}^K$ that softly assigns a sample to K environments. Here K is a pre-specified number (a hyper-parameter) and its choice will be empirically investigated in our experiments. Let $\rho^{(k)}(\cdot)$ denote the k-th entry of $\rho(\cdot)$. Then we have $\rho(\boldsymbol{Z}) \in [0,1]^K$ with $\sum_k \rho^{(k)}(\boldsymbol{Z}) = 1$. Denote the ERM loss as $\mathcal{R}(\omega, \Phi) = \frac{1}{n}\sum_{i=1}^n \ell\left(f_w(\Phi(\boldsymbol{x}_i)), y_i\right)$ and the loss in the k-th inferred environment as $\mathcal{R}_{\rho^{(k)}}(\omega, \Phi) = \frac{1}{n}\sum_{i=1}^n \rho^{(k)}(\boldsymbol{z}_i)\ell(f_w(\Phi(\boldsymbol{x}_i)), y_i)$.

Suppose that the environments have been given according to a fixed $\rho(\cdot)$. Recall that IRM [5] learns an invariant representation Φ, upon which there is a classifier ω that is simultaneously optimal in all environments. To measure the optimality of ω in environment k, we can fit an environment dependent classifier ω_k on the data from environment k. If ω_k achieves a smaller loss than ω, it means ω is not optimal in this environment. We can further train a set of environment dependent classifiers $\{\omega_k\}_{k=1}^K$ (one for each environment), to measure whether ω is simultaneously optimal in all environments. Thus, when $\rho(\cdot)$ is provided, our formulation to learn invariance is then given as follows:

$$\min_{\omega,\Phi} \max_{\{\omega_k\}} \mathcal{L}(\Phi, \omega, \omega_1, \ldots, \omega_K, \rho) := \mathcal{R}(\omega, \Phi) + \lambda \underbrace{\sum_{k=1}^K \left[\mathcal{R}_{\rho^{(k)}}(\omega, \Phi) - \mathcal{R}_{\rho^{(k)}}(\omega_k, \Phi)\right]}_{\text{invariance penalty}} \tag{3}$$

If Φ extracts spurious features that are unstable in the inferred environments, $\mathcal{R}_{\rho^{(k)}}(\omega, \Phi)$ will be larger than $\mathcal{R}_{\rho^{(k)}}(\omega_k, \Phi)$ and it induces a non-zero penalty.

Next, we consider how to learn the partition function $\rho(\cdot)$. A *good* partition function should generate environments among which the spurious feature exhibit instability. If Φ extracts spurious features, we expect to have a large penalty resulting from the environments. Thus, we seek for an environment partition that maximizes the invariance penalty. The overall framework is provided below:

$$\min_{\omega,\Phi} \max_{\rho,\{\omega_k\}} \mathcal{L}(\Phi, \omega, \omega_1, \ldots, \omega_K, \rho). \tag{4}$$

The idea of our method can be summarized as inferring an environment partition and then learning invariant models based on the inferred environments. Creager et al. [9] has used a similar intuition and adopted a pre-trained biased model to estimate the environments. However, their two-stage method cannot be jointly optimized, and the environment partition relies on the given model and lacks a theoretic guarantee. In contrast, we will first present sufficient conditions on $\boldsymbol{Z}$ so that the proposed framework can provably identify the invariant features. We further show that these conditions are also necessary, and that violation of these conditions will lead to failure of invariance identification. Assigning independent weights to each data sample, which is adopted in EIIL [9], is unfortunately an instance of such violations and may fail to identify invariant features.

5.2 Sufficient Conditions for Identifiability

In this section, we try to understand ZIN from a theoretic perspective. We start with a simple yet general setting: $\boldsymbol{X} = [\boldsymbol{X}_v; \boldsymbol{X}_s]$ (i.e., no scramble on the observation), $\Phi \in \{0,1\}^d$ is an element-wise feature selection mask, and f_w is general non-linear function; the next section will present theoretic results with linear feature transformations, the same setting as in [5]. We also focus on classification tasks with $\ell(\cdot,\cdot)$ being the cross-entropy loss, as there is an interesting connection to (conditional) Shannon entropy. Extension to other loss functions is left to Appendix B.

In this setting, our goal is equivalent to learning the optimal feature mask that merely selects invariant features, i.e., $\Phi_v = [\mathbf{1}^{d_v}; \mathbf{0}^{d_s}]$. For a given feature mask Φ, the objective function is equal to

$$\hat{\mathcal{L}}(\Phi) = \min_{\omega} \max_{\rho,\{\omega_k\}} \mathcal{L}(\Phi, \omega, \omega_1, \cdots, \omega_k, \rho).$$

Then ZIN can correctly identify the invariant features, or solution to Problem 4 is equivalent to Φ_v, if and only if $\hat{\mathcal{L}}(\Phi_v) < \hat{\mathcal{L}}(\Phi)$ for all $\Phi \neq \Phi_v$. This observation will be used to establish our

main theoretic result. With a little abuse of notation, we use $H(Y|\boldsymbol{X}')$ to denote expected loss of an optimal classifier over some $\boldsymbol{X}'$ and Y, and similarly $H(Y|\rho(\boldsymbol{Z}), \boldsymbol{X}')$ to denote the minimum risk $\sum_{k=1}^{K} \mathcal{R}_{\rho^{(k)}}(\omega_k, \Phi)$ for a given $\rho(\boldsymbol{Z})$. With cross-entropy loss and when $\rho(\boldsymbol{z}_i)$ gives exactly one environment, i.e., $\rho(\boldsymbol{Z})$ is one-hot, it can be verified that the optimal loss $H(\cdot|\cdot)$ coincides with conditional entropy. In the following we first state the assumptions for our identifiabitiliy result.

Assumption 1. *For a given feature mask Φ and any constant $\epsilon > 0$, there exists $f \in \mathcal{F}$ such that $\mathbb{E}[\ell(f(\Phi(\boldsymbol{X})), Y)] \leq H(Y|\Phi(\boldsymbol{X})) + \epsilon$.*

Assumption 2. *If two features are invariant w.r.t. an environment partition, then the concatenated features are also invariant. That is, for $\boldsymbol{X}_1, \boldsymbol{X}_2 \subset \boldsymbol{X}$ and $\rho(\cdot)$, if $H(Y|\boldsymbol{X}_1) - H(Y|\rho(\boldsymbol{Z}), \boldsymbol{X}_1) = 0$ and $H(Y|\boldsymbol{X}_2) - H(Y|\rho(\boldsymbol{Z}), \boldsymbol{X}_2) = 0$, we have $H(Y|\boldsymbol{X}_1, \boldsymbol{X}_2) - H(Y|\rho(\boldsymbol{Z}), \boldsymbol{X}_1, \boldsymbol{X}_2) = 0$.*

Assumption 3. *If a feature representation violates the invariance constraint, adding another feature would not make the penalty vanish. This implies that there exists a constant $\delta > 0$ so that for spurious feature $\boldsymbol{X}_1 \subset \boldsymbol{X}_s$ and any other feature $\boldsymbol{X}_2 \subset \boldsymbol{X}$, $H(Y|\boldsymbol{X}_1, \boldsymbol{X}_2) - H(Y|\rho(\boldsymbol{Z}), \boldsymbol{X}_1, \boldsymbol{X}_2) \geq \delta\left(H(Y|\boldsymbol{X}_1) - H(Y|\rho(\boldsymbol{Z}), \boldsymbol{X}_1)\right)$.*

Assumption 4. *For any distinct features $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $H(Y|\boldsymbol{X}_1, \boldsymbol{X}_2) \leq H(Y|\boldsymbol{X}_1) - \gamma$ with fixed $\gamma > 0$.*

Assumption 1 is a common assumption that requires the function space $\mathcal{F}$ be rich enough such that, given Φ, there is a $f \in \mathcal{F}$ that can fit $P(Y|\Phi(\boldsymbol{X}))$ well. Assumption 4 indicates that any feature contains some useful information w.r.t. Y, which cannot be explained by other features. Otherwise, we can simply remove such a feature (e.g., by variable selection methods [46]), as it does not affect prediction. We next present our sufficient conditions for ZIN to identify invariant features.

Condition 1 (Invariance Preserving Condition)**.** *Given invariant feature $\boldsymbol{X}_v$ and any function $\rho(\cdot)$, it holds that $H(Y|\boldsymbol{X}_v, \rho(\boldsymbol{Z})) = H(Y|\boldsymbol{X}_v)$.*

Condition 2 (Non-invariance Distinguishing Condition)**.** *For any feature $\boldsymbol{X}_s^k \in \boldsymbol{X}_s$, there exists a function $\rho(\cdot)$ and a constant $C > 0$ such that $H(Y|\boldsymbol{X}_s^k) - H(Y|\boldsymbol{X}_s^k, \rho(\boldsymbol{Z})) \geq C$.*

We remark that Condition 1 can be met if $H(Y|\boldsymbol{X}_v, \boldsymbol{Z}) = H(Y|\boldsymbol{X}_v)$ (a proof is provided in Appendix A.7). Condition 1 requires that invariant features should remain invariant w.r.t any environment partition induced by $\rho(\boldsymbol{Z})$. Otherwise, if there exists a partition where an invariant feature becomes non-invariant, then this feature would induce a positive penalty. Condition 2 implies that for each spurious feature, there exists at least one partition so that this feature is non-invariant in the split environments. If a spurious feature does not incur any invariance penalty in all possible environment partitions, we can never distinguish it from true invariant features. With these conditions, our main result follows and a proof is given in Appendix A.2.

Theorem 2 (Identifiability of Invariant Features)**.** *With Assumptions 1, 3 and 4 and Conditions 1-2, if $\epsilon < \frac{C\gamma\delta}{4\gamma + 2C\delta H(Y)}$ and $\lambda \in [\frac{H(Y)+1/2\delta C}{\delta C - 4\epsilon} - \frac{1}{2}, \frac{\gamma}{4\epsilon} - \frac{1}{2}]$, then we have $\hat{\mathcal{L}}(\Phi_v) < \hat{\mathcal{L}}(\Phi)$ for all $\Phi \neq \Phi_v$, where $H(Y)$ denotes the entropy of Y. Thus, the solution to Problem 4 identifies invariant features.*

5.3 Necessary Conditions for Identifiability

Conditions 1 and 2 may appear rather strong at first glance. In this section, we prove that they are also necessary conditions for ZIN to identify invariant features. Specifically, Proposition 1 shows that if Condition 1 is violated, then some invariant features will be excluded in the solution of Problem 4; and in Proposition 2, violation of Condition 2 renders some spurious features included in the solution.

Proposition 1. *With Assumptions 1, 3 and 4, if Condition 1 is violated, i.e., there exists $\rho(\cdot)$ so that $H(y|\boldsymbol{X}_v) - H(y|\boldsymbol{X}_v, \rho(\boldsymbol{Z})) \geq C' > 0$, then there exists a feature mask $\Phi' \neq \Phi_v$ with $\hat{\mathcal{L}}(\Phi_v) > \hat{\mathcal{L}}(\Phi')$.*

A proof is provided in Appendix A.3, which is similar to the first step of the proof of Theorem 2. A question is how Condition 1 could be violated. On the other hand, while we introduce $\mathbf{Z}$ as additional variables, it is also interesting to know whether we can obtain a valid partition based on only the training data $(\boldsymbol{X}, Y)$. Below we provide examples regarding these questions.

Corollary 1. *Condition 1 is violated for the following cases: if there is a function $\rho(\cdot)$ and an injective function $h(\cdot)$ so that (a) $\rho(\boldsymbol{Z}) = h(Y)$, (b) $\rho(\boldsymbol{Z}) = h(\boldsymbol{X}, Y)$, or (c) $\rho(\boldsymbol{Z}) = h(\mathrm{Index}(\boldsymbol{X}, Y))$.*

A proof is given in Appendix A.4. As $h(\mathrm{Index}(\boldsymbol{X}, Y))$ can be treated as learning independent weights for each sample based on $\{(\boldsymbol{x}_i, y_i)\}_i$, this case includes EIIL [9] as an instance.

Proposition 2. *With Assumptions 1- 4, if Condition 2 is violated, i.e., there exists a feature $\boldsymbol{X}_s^k \in \boldsymbol{X}_s$ such that $H(Y|\boldsymbol{X}_s^k) - H(Y|\boldsymbol{X}_s^k, \rho(\boldsymbol{Z})) = 0$ for any $\rho(\cdot)$, then there exists a feature mask $\Phi' \neq \Phi_v$ with $\hat{\mathcal{L}}(\Phi_v) > \hat{\mathcal{L}}(\Phi')$.*

Since there exists a spurious feature $\boldsymbol{X}_s^k$ that is "invariant" in all possible environment partition, adding this feature to Φ_v does not induce any penalty but increases the prediction power. Thus, such a feature mask can achieve a smaller loss than $\hat{\mathcal{L}}(\Phi_v)$. For completeness, we provide a proof in Appendix A.5. Proposition 2 again indicates in theory that $\boldsymbol{Z}$ should be sufficiently diverse and informative so that each spurious feature can be recognized.

6 Beyond Feature Selection: Linear Feature Transformations

In Section 5, both the invariant and spurious features can be directly observed and we focus on the ability of ZIN to identify the invariant features. In this section, we extend feature selection to feature learning, and show that ZIN is able to learn the invariant features given a scrambled observation in a linear form following [5, 36]. Specifically, we consider the same data generation process as [5]:

$$Y^e = \boldsymbol{X}_v^e \cdot \beta + \epsilon_v, \quad \boldsymbol{X}_v^e \perp\!\!\!\perp \epsilon_v, \quad \mathbb{E}[\epsilon_v] = 0; \qquad \boldsymbol{X}^e = \boldsymbol{W} \cdot [\boldsymbol{X}_v^e; \boldsymbol{X}_s^e], \tag{5}$$

where $\beta \in \mathbb{R}^{d_v}$ and $\boldsymbol{W} \in \mathbb{R}^{d \times (d_v + d_s)}$. We assume that there exists $\tilde{\boldsymbol{W}} \in \mathbb{R}^{d_v \times d}$ so that $\tilde{\boldsymbol{W}}(\boldsymbol{W}[\boldsymbol{x}_v; \boldsymbol{x}_s]) = \boldsymbol{x}_v$ for all $\boldsymbol{x}_v$ and $\boldsymbol{x}_s$. Both the feature extractor and predictor take a linear form, i.e., Φ takes values in $\mathbb{R}^{d \times d}$ and $\omega \in \mathbb{R}^d$. The prediction for $\boldsymbol{X}$ is $\omega \circ \Phi(\boldsymbol{X}) = (\Phi \boldsymbol{X})^T \omega$.

A major difficulty in this setting lies in how to characterize the effect of an invariant/spurious feature in a *quantitive* way: the feature extractor Φ may extract an arbitrarily small portion of spurious information. Following [5, 36], we consider a constrained form of Problem 4 for theoretic analysis:

$$\min_{\omega, \Phi} \mathcal{R}(\omega, \Phi), \quad \text{subject to} \max_{\rho, \{\omega_k\}} \sum\nolimits_{k=1}^{K} \left[\mathcal{R}_{\rho^{(k)}}(\omega, \Phi) - \mathcal{R}_{\rho^{(k)}}(\omega_k, \Phi)\right] = 0. \tag{6}$$

As in Conditions 1 and 2, the auxiliary information should also be sufficiently informative so that the inferred environments can be diverse enough but also maintain the underlying invariance. This is in accordance to existing conditions for identifiability in the linear case [5, 36]. In this paper, we utilize such a condition, *linear general position* condition, from [5]. For our analysis, we take squared error as our loss function and consider that $\rho(\cdot)$ partitions the environments in a hard manner, i.e., each data sample would be assigned to exactly one environment. We also assume that the environments are non-degenerate, i.e., each inferred environment contains some data samples; otherwise, we can simply remove such an environment. Our identifiability result for the linear case then follows.

Theorem 3. *Assume that there exists $\rho(\cdot)$ such that the generated environments, denoted as $\{\boldsymbol{X}^k\}_{k=1}^K$, lie in linear general position of degree r, i.e., $K > d - r + d/r$ for some $r \in \mathbb{N}$ and for all non-zero $\boldsymbol{x} \in \mathbb{R}^d$: $\dim\left(\mathrm{span}\left(\left\{\mathbb{E}_{\boldsymbol{X}^k}\left[\boldsymbol{X}^k \boldsymbol{X}^{k\intercal}\right]\boldsymbol{x} - \mathbb{E}_{\boldsymbol{X}^k, \epsilon_v}\left[\boldsymbol{X}^k \epsilon_v\right]\right\}_k\right)\right) > d - r$. If $\Phi \in \mathbb{R}^{d \times d}$ has rank $r > 0$, then Problem 6 results in the desired invariant predictor.*

A proof is given in Appendix A.6. We can also replace the above assumption by the non-degenerate conditions in Eqn. (9) and (10) of [36], and a similar identifiability result follows.

7 Experiments

This section empirically verifies our theoretic analysis and the effectiveness of ZIN on both synthetic and real world datasets. We compare ZIN with several existing methods: ERM, IRM, [5], GroupDRO [38], EIIL [9], HRM [26], and LfF [31]. We provide the ground-truth partition to IRM and group DRO. LfF tries to learn a robust model by directly boosting from wrongly specified samples of shallow neural networks. Notice that LfF only works for classification tasks.

Table 1: Test Mean and Worst accuracy (%) on four temporal heterogeneity synthetic datasets.

Env Partition	$p_s(t)$	0.999, 0.7				0.999, 0.8				0.999, 0.9			
	p_v	0.9		0.8		0.9		0.8		0.9		0.8	
	Test Acc	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst
No	ERM	75.37	57.31	59.65	25.81	68.72	41.97	55.90	15.07	60.61	23.39	52.85	7.57
	EIIL	38.41	16.80	64.89	49.15	50.77	46.67	68.36	56.35	61.99	53.81	70.10	59.36
	HRM	50.00	49.99	49.98	49.93	50.00	49.98	50.01	49.99	50.00	49.98	49.99	49.97
	ZIN	**87.50**	**85.36**	**77.85**	**75.39**	**86.35**	**82.91**	**76.79**	**72.77**	**83.71**	**75.89**	**73.55**	**64.69**
Yes	IRM	87.57	85.47	77.99	75.65	86.57	83.25	77.00	73.39	83.99	76.48	73.84	65.33

We implement both ρ and f in ZIN as a two-layer MLP with 32 hidden units. The number of inferred environments K is set to be 2 as default. We implement Φ as a two-layer MLP for the synthetic dataset and house price prediction, ResNet-18 [14] for CelebA, and 1D CNN for Landcover. Please see Appendix D for more details.

7.1 Synthetic Dataset

Temporal Heterogeneity. We consider temporal heterogeneity with distributional shift w.r.t. time. Let $t \in [0, 1]$ be time index and $X_v(t) \in \mathbb{R}$ the invariant feature. The data generation process is

$$X_v(t) \sim \begin{cases} \mathcal{N}(1, \sigma^2), & w.p.\ 0.5, \\ \mathcal{N}(-1, \sigma^2), & w.p.\ 0.5, \end{cases} \quad Y(t) \sim \begin{cases} \text{sign}(X_v(t)), & w.p.\ p_v, \\ -\text{sign}(X_v(t)), & w.p.\ p_v', \end{cases} \quad X_s(t) \sim \begin{cases} \mathcal{N}(Y(t), \sigma^2), & w.p.\ p_s(t), \\ \mathcal{N}(-Y(t), \sigma^2), & w.p.\ p_s'(t). \end{cases}$$

where p_v is a constant w.r.t. t, indicating a stable correlation between $Y(t)$ and $X_v(t)$, $p_v' = 1 - p_v$, and $p_s'(t) = 1 - p_s(t)$. Notice here $p_s(t)$ varies with time t. A similar setting with two-dimensional spatial variable is considered in Appendix F.1. Our goal is to learn a model that purely relies on X_v. We simulate two heterogeneous environments along time, namely, $\{[0, 0.5), [0.5, 1]\}$, and $p_s(t)$ will be set differently. We use tuple of $p_s(t)$ in the two environments to denote a simulated case. For example, $(0.999, 0.7)$ stands for $p_s(t) = 0.999, t \in [0, 0.5)$ and $p_s(t) = 0.7, t \in [0.5, 1]$. We evaluate the performance on four distinct test environments with $p_s \in \{0.999, 0.8, 0.2, 0.1\}$ and p_v being constant. We use t as the auxiliary information Z. More implementation details are given in Appendix D

Results. Table 1 reports the test accuracy. In all simulated settings, the worst accuracy of ERM is much lower than the mean accuracy, indicating that ERM tends to rely on spurious feature X_s. EIIL can improve the worst accuracy in some cases, e.g., when $p_s(t) = (0.999, 0.8)$ and $p_s(t) = (0.999, 0.9)$. However, its performance is even worse than ERM for some other settings. This may be attributed to the first stage of EIIL, where the trained biased model is not guaranteed and may learn both spurious and invariant features. The proposed method ZIN improves the worst test accuracy significantly. For instance, when $p_s(t) = (0.999, 0.7)$ and $p_v = 0.9$, ZIN outperforms ERM and EIIL by over 28% and 65%, respectively. Moreover, ZIN is very close to IRM that knows the ground-truth environment partition, showing that ZIN can infer the environments effectively. Finally, it seems that HRM does not learn a useful model for this classification task.

Ablation Study. In Appendix F.2, we conduct ablation study to verify our theoretic results in Section 5 by choosing different auxiliary variables. We also empirically study the choice of hyper-parameter K. A notable result is that K has a relatively small impact, as shown in Figure 2.

7.2 Real World Datasets

House Price Prediction. This experiment considers a real world regression dataset of house sales prices from Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques). The target variable is house price and each sample contains 17 predictive variables like the built year of the house, number of bedrooms, etc. The dataset is split according to the built year, with training dataset in period $[1900, 1950]$ and test dataset in period $(1950, 2000]$. We normalize the prices of the houses with the same built year, and our target is to predict the normalized price. We choose the built year as auxiliary information for ZIN. As there is no well-defined ground-truth partition for IRM, we manually split the training dataset equally into 5 segments with 10-year range in each segment.

Table 2: House price prediction (MSE).

Method	Env Index	Train	Test Mean	Test Worst
IRM	Yes	0.1112	0.3328	0.4913
GroupDRO	Yes	0.1213	0.6887	1.005
ERM	No	**0.1007**	0.3597	0.4968
EIIL	No	0.6841	0.9625	1.3909
EIIL(+$\boldsymbol{Z}$)	No	0.6912	0.9701	1.4201
HRM	No	0.3466	0.3521	0.5248
HRM(+$\boldsymbol{Z}$)	No	0.3190	0.3764	0.5726
ZIN	No	0.2275	**0.3118**	**0.4285**

Table 3: Accuracy (%) on CelebA task.

Method	Env Index	Train	Test Mean	Test Worst
IRM	Yes	81.30±1.53	78.44±0.48	75.03±1.29
GroupDRO	Yes	89.32±0.67	74.28±0.22	58.11±0.61
ERM	No	**90.97**±0.66	70.76±0.26	47.58±0.46
LfF	No	59.89±0.72	52.97±0.56	44.38±2.01
EIIL	No	44.78±5.85	55.26±1.17	43.12±8.98
ZIN(1)	No	90.62±0.78	70.79±0.61	47.62±0.98
ZIN(4)	No	83.57±1.40	75.20±0.71	63.47±1.41
ZIN(7)	No	83.06±1.28	**76.29**±0.60	**67.27**±1.15

Table 2 reports the mean squared error (MSE), and ZIN again outperforms other methods on test dataset. We also investigate whether this type of additional information $\boldsymbol{Z}$, together with $\boldsymbol{X}$, helps EIIL and HRM. The corresponding results are marked with $(+\boldsymbol{Z})$ in Table 2. We observe that the results of EIIL and HRM with additional information $\boldsymbol{Z}$ are worse than those without using $\boldsymbol{Z}$. This implies that such information may not be directly utilized by or even can harm the two methods.

CelebA. This task is to predict *Smiling* based on the image from CelebA [28], and by construction the target is spuriously correlated with *Gender*. We have access to the meta annotations of CeleA images and they serve as $\boldsymbol{Z}$ in our framework. In particular, we pick seven features as potential auxiliary variables: {*Young*, *Blond Hair*, *Eyeglasses*, *High Cheekbones*, *Big Nose*, *Bags Under Eyes*, *Chubby*}. Notice that the annotation of *Gender* is not provided to ZIN, EIIL, HRM and ERM, but is made available to IRM to create an oracle environment partition.

Empirical results are given in Table 3, where ZIN(#) represents ZIN using # of the seven features as $\boldsymbol{Z}$. ERM achieves the highest training accuracy, while only has 47.58% worst test accuracy. EIIL has the worst performance on both training and test data, showing that it may not learn the causal features. Finally, we observe that ZIN performs better when more auxiliary information is provided; specifically, ZIN(7) achieves the best test performance based on the mixed dataset.

Landcover. Our final task is land cover prediction that classifies the land cover type (e.g., grasslands) from satellite data [13, 37]. We take the same setup from [49]: the time series data input dimension is 46×8; the target Y is one of six land cover classes; six climate related variables are the auxiliary variables. We also consider non-African locations as training data and Africa as test data. In addition, we take location (latitude and longitude) as possible additional information. For ZIN, we average climate variables over the time dimension and use those means to predict environments. In Table 4, we see that ZIN with location achieves the best OOD performance, while ZIN with climate variables is slightly worse. This is because climate variables indeed contain some information for predicting the target, violating our condition $\boldsymbol{Z}$. EIIL also has a good performance in this task, and we conjecture that its first stage has learned the spurious features as desired. However, together with previous empirical results, it can be risky as the first stage may not be guaranteed.

Table 4: Test Accuracy (%) on Landcover task.

Method	IID Test	OOD Test
ERM [48]	75.92	58.31
ERM (+climate) [48]	76.58	54.78
LfF	49.90	48.65
EIIL	71.51	64.37
ZIN (climate)	72.35	63.45
ZIN (location)	72.16	**65.01**

8 Concluding Remarks

This paper investigates when and how to learn invariance from heterogeneous data without explicit environment indexes. We first show that learning invariant models in this case is generally impossible. Then we propose ZIN to jointly learn environment partition and invariant representation using some additional auxiliary variables. We provide theoretic guarantees in both feature selection and linear feature learning scenarios. Experimental results verify our analysis and demonstrate an improved performance of ZIN over existing methods. Our results also raise the need of future works to make the role of inductive biases more explicit, when learning invariance from heterogeneous data without environment indexes. A limitation of the present work is the lack of theoretic guarantee with general nonlinear feature extractors, which is also an open problem for IRM even with environment indexes.

References

[1] Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2020.

[2] Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron C. Courville. Systematic generalisation with group invariant predictions. In *ICLR*, 2021.

[3] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, 2020.

[4] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or invariant risk minimization? A sample complexity perspective. *International Conference on Learning Representations*, 2021.

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

[6] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[7] Joos-Hendrik Böse, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Dustin Lange, David Salinas, Sebastian Schelter, Matthias Seeger, and Yuyang Wang. Probabilistic demand forecasting at scale. *Proceedings of the VLDB Endowment*, 10(12):1694–1705, 2017.

[8] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, 2020.

[9] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, 2021.

[10] David Danks, Clark Glymour, and Robert Tillman. Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems*, 2009.

[11] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[12] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *arXiv preprint arXiv:1712.06050*, 2020.

[13] Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R. Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[16] Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery from multiple data sets with non-identical variable sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[17] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Clark Glymour, and Bernhard Schölopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21, May 2020.

[18] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[19] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Domain extrapolation via regret minimization. *arXiv preprint arXiv:2006.03908*, 2020.

[20] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014.

[22] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 2021.

[23] Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, 2018.

[24] Yong Lin, Lian Qing, and Tong Zhang. An empirical study of invariant risk minimization on deep models. 2021.

[25] Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, 2021.

[26] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, 2021.

[27] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Kernalized heterogeneous risk minimization. In *NeurIPS*, 2021.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[29] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

[30] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-Science Reviews*, 190:310–322, 2019.

[31] Jun Hyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *NeurIPS*, 2020.

[32] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

[33] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[34] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

[35] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, 2019.

[36] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *International Conference on Learning Representations*, 2021.

[37] Marc Russwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[38] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

[39] Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *ICML*, 2021.

[40] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

[41] Nimit Sharad Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *NeurIPS*, 2020.

[42] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, second edition, 2000.

[43] Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, 2020.

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[45] Robert Tillman and Peter Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

[46] Jorge Vergara and Pablo Estevez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24, 01 2014.

[47] Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B. Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12(2), 2020.

[48] Chuanlong Xie, Fei Chen, Yue Liu, and Zhenguo Li. Risk variance penalization: From distributional robustness to causality. *arXiv preprint arXiv:2006.07544*, 2020.

[49] Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-N-Out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations*, 2021.

[50] Yilun Xu and Tommi Jaakkola. Learning representations that support robust transfer of predictors. *arXiv preprint arXiv:2110.09940*, 2021.

[51] Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhi-Ming Ma. Improved OOD generalization via adversarial training and pre-training. In *International Conference on Machine Learning*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In the conclusion section.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In Appendix H.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

A.1 Proof of Theorem 1

Proof. First, we assume Y and $\boldsymbol{X}_v$ are univariate variables, i.e., $Y, \boldsymbol{X}_v \in \mathbb{R}$. Let $\eta_1 \sim \text{Uniform}(0,1)$ independent of $\boldsymbol{X}_s$ and Y, and set $\boldsymbol{X}'_v = \boldsymbol{X}_s$. Define the conditional cumulative distribution function and its inverse as:

$$F_{Y|\boldsymbol{X}_s=\boldsymbol{x}_s}(y) = P(Y \leq y \mid \boldsymbol{X}_s = \boldsymbol{x}_s),$$
$$Y' = f'_1(\boldsymbol{X}'_v, \eta_1) = F^{-1}_{Y|\boldsymbol{X}_s}(\eta_1) = \inf\{y \in \mathbb{R} : F_{Y|\boldsymbol{X}_s}(y) \geq \eta_1\} \text{ with } \boldsymbol{X}_s = \boldsymbol{X}'_v.$$

By definition, we would have

$$\begin{aligned} P(Y' \leq y \mid \boldsymbol{X}'_v) &= P(f'_1(\boldsymbol{X}'_v, \eta_1) \leq y) \\ &= P(F^{-1}_{Y|\boldsymbol{X}_s}(\eta_1) \leq y) \\ &= P(F_{Y|\boldsymbol{X}_s} \circ F^{-1}_{Y|\boldsymbol{X}_s}(\eta_1) \leq F_{Y|\boldsymbol{X}_s}(y)) \\ &= P(\eta_1 \leq F_{Y|\boldsymbol{X}_s}(y)) \\ &= P(Y \leq y \mid \boldsymbol{X}_s). \end{aligned} \tag{7}$$

Similarly, we can construct $\boldsymbol{X}'_s = f'_2(\boldsymbol{X}'_v, Y', \eta_2) = F^{-1}_{\boldsymbol{X}_v|Y,\boldsymbol{X}_s}(\eta_2)$ so that $P(\boldsymbol{X}'_s|Y', \boldsymbol{X}'_v) = P(\boldsymbol{X}_v|Y, \boldsymbol{X}_s)$ with $\eta_2 \sim \text{Uniform}(0,1)$. Thus, we have

$$\begin{aligned} P(\boldsymbol{X}'_v, \boldsymbol{X}'_s, Y) &= P(\boldsymbol{X}'_s|Y', \boldsymbol{X}'_v)P(Y'|\boldsymbol{X}'_v)P(\boldsymbol{X}'_v) \\ &= P(\boldsymbol{X}_v|Y, \boldsymbol{X}_s)P(Y|\boldsymbol{X}_s)P(\boldsymbol{X}_s) \\ &= P(\boldsymbol{X}_v, \boldsymbol{X}_s, Y). \end{aligned} \tag{8}$$

Next, we can easily find a function $q'(\cdot)$ so that $\boldsymbol{X}' = q'(\boldsymbol{X}'_v, \boldsymbol{X}'_s) = q(\boldsymbol{X}_v, \boldsymbol{X}_s)$, where we have chosen $\boldsymbol{X}'_v = \boldsymbol{X}_s$ and $\boldsymbol{X}'_s = \boldsymbol{X}_v$. Together with Equation (8), we conclude that $P(X', Y') = P(X, Y)$.

Second, we consider that $\boldsymbol{X}_v$ has a multi-dimension. We may leave Y to be a univariate variable as the label in many ML problems is scalar-valued. For this case, we can pick an entry of $\boldsymbol{X}_v$, say, $\boldsymbol{X}_v^{(j)}$ for some j. Then we set $\boldsymbol{X}'_v = (\boldsymbol{X}_v^{(-j)}, \boldsymbol{X}_s)$ where $\boldsymbol{X}_v^{(-j)}$ denotes the rest entries of $\boldsymbol{X}_v$ except the j-th. Then we can similarly use the inverse conditional cumulative distribution function and an independent uniformly distributed noise variable η_1 to construct $Y' = f'(\boldsymbol{X}'_v, \eta_1)$ so that $P(Y' \leq y \mid \boldsymbol{X}'_v) = P(Y \leq y \mid \boldsymbol{X}_v^{(-j)}, \boldsymbol{X}_s)$. Similarly, we set $\boldsymbol{X}'_s = \boldsymbol{X}_v^{(j)}$ and we can construct $\boldsymbol{X}'_s = f'_2(\boldsymbol{X}'_v, Y', \eta_2)$ so that $P(\boldsymbol{X}'_s|Y', \boldsymbol{X}'_v) = P(\boldsymbol{X}_v^{(j)} \mid Y, \boldsymbol{X}_s)$. We can then get the same conclusion following the previous proof. □

A.2 Proof of Theorem 2

Proof. Our proof proceeds by two steps. First, we show that any feature mask that selects at least one spurious feature would induce a penalty. With sufficiently large λ, the penalty will dominate the expected risk and then exceed $\hat{\mathcal{L}}(\Phi_v)$. Second, we show that any proper subset of the invariant features induces a loss larger than $\hat{\mathcal{L}}(\Phi_v)$.

Step 1 Suppose that the feature mask contains at least one spurious features. Denote the selected features as $\boldsymbol{X}_{+s}$ and the corresponding feature mask as Φ_{+s}. We aim to show that

$$\hat{\mathcal{L}}_{(\Phi_{+s}} > \hat{\mathcal{L}}(\Phi_v).$$

By Assumption 1 with a given $\epsilon > 0$, we have

$$\begin{aligned} \hat{\mathcal{L}}(\Phi_v) &\leq (1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_v) + \lambda\left(H(Y|\boldsymbol{X}_v) - H(Y|\boldsymbol{X}_v, \rho(Z))\right) \\ &= (1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_v) \\ &\leq (1+2\lambda)\epsilon + H(Y). \end{aligned} \tag{9}$$

One the other hand, we have

$$\begin{aligned}\mathcal{L}(\Phi_{+s}) &\geq -(1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_{+s}) + \lambda\left(H(Y|\boldsymbol{X}_{+s}) - H(Y|\boldsymbol{X}_{+s}, \rho(Z))\right) \\ &\geq -(1+2\lambda)\epsilon + \lambda\left(H(Y|\boldsymbol{X}_{+s}) - H(Y|\boldsymbol{X}_{+s}, \rho(\boldsymbol{Z}))\right) \\ &\geq -(1+2\lambda)\epsilon + \lambda\delta C, \end{aligned} \tag{10}$$

where the last inequality is due to Assumption 3 and Condition 2. Thus, if we choose $\epsilon < \delta C/4$ and $\lambda > \frac{H(Y)+2\epsilon}{\delta C - 4\epsilon}$, we can get

$$\mathcal{L}(\Phi_v) < \mathcal{L}(\Phi_{+s}).$$

Step 2 Denote a proper subset of invariant features as $\boldsymbol{X}_{-v} \subsetneq \boldsymbol{X}_v$, and similarly the feature mask as Φ_{-v}.

In Step 1, we have shown that

$$\hat{\mathcal{L}}(\Phi_v) \leq (1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_v).$$

Similar to Equation (10), we have

$$\hat{\mathcal{L}}(\Phi_{-v}) \geq -(1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_{-v}).$$

Then according to Assumption 4, we have

$$\begin{aligned}\hat{\mathcal{L}}(\Phi_{-v}) - \hat{\mathcal{L}}(\Phi_v) &\geq -2(1+2\lambda)\epsilon + H(y|\boldsymbol{X}_{-v}) - H(y|\boldsymbol{X}_v) \\ &\geq -2(1+2\lambda)\epsilon + \gamma.\end{aligned}$$

Thus, if $\epsilon < \frac{\gamma}{2(1+2\lambda)}$, we have

$$\hat{\mathcal{L}}(\Phi_{-v}) > \hat{\mathcal{L}}(\Phi_v).$$

In conclusion, with $\lambda \in [\frac{H(Y)+1/2\delta C}{\delta^d C - 4\epsilon} - \frac{1}{2}, \frac{\gamma}{4\epsilon} - \frac{1}{2}]$, we can get

$$\hat{\mathcal{L}}(\Phi_v) < \hat{\mathcal{L}}(\Phi), \quad \forall \Phi \neq \Phi_v.$$

Notably, there exists a feasible λ if $\epsilon < \frac{C\gamma\delta}{4\gamma + 2C\delta H(Y)}$. The proof is complete by noticing that ϵ can be chosen arbitrarily according to Assumption 1. □

A.3 Proof of Proposition 1

Proof. Consider the following feature set

$$\boldsymbol{X}_{\bar{v}} := \max_{|\boldsymbol{X}'|}\{\boldsymbol{X}' \subset \boldsymbol{X} : H(Y|\boldsymbol{X}, \rho(\boldsymbol{Z})) = H(Y|\boldsymbol{X}) \quad \forall \rho(\cdot)\},$$

and the corresponding feature mask is denoted as $\Phi_{\bar{v}}$. It corresponds the largest subset of $\boldsymbol{X}$ that satisfies the invariance constraint.

Note that $\Phi_{\bar{v}} \neq \Phi_v$. By Assumption 1 with a given ϵ, we can have

$$\begin{aligned}\hat{\mathcal{L}}(\Phi_{\bar{v}}) &\leq (1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_{\bar{v}}) + \lambda(H(Y|\boldsymbol{X}_{\bar{v}}) - H(Y|\boldsymbol{X}_v, \rho(\boldsymbol{Z}))) \\ &= (1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_{\bar{v}}) \\ &\leq (1+2\lambda)\epsilon + H(Y).\end{aligned}$$

One the other hand, we have

$$\begin{aligned}\mathcal{L}(\Phi_v) &\geq -(1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_v) + \lambda\left(H(Y|\boldsymbol{X}_v) - H(Y|\boldsymbol{X}_v, \rho(\boldsymbol{Z}))\right) \\ &\geq -(1+2\lambda)\epsilon + \lambda\left(H(Y|\boldsymbol{X}_v) - H(Y|\boldsymbol{X}_v, \rho(\boldsymbol{Z}))\right) \\ &\geq -(1+2\lambda)\epsilon + \lambda\delta C',\end{aligned}$$

which can be shown similarly to Equation (10). Thus, if we choose $\epsilon < \delta C'/4$ and $\lambda > \frac{H(Y)+2\epsilon}{\delta C' - 4\epsilon}$, we would get

$$\mathcal{L}(\Phi_{\bar{v}}) < \mathcal{L}(\Phi_v).$$

□

A.4 Proof of Corollary 1

Proof. (a) Since h is injective, $H(Y|\boldsymbol{X}_v, h(Y)) = H(Y|\boldsymbol{X}_v, Y) = 0$ for any $\boldsymbol{X}_v$. By Assumption 4, we have $H(Y|\boldsymbol{X}_v) \geq H(Y|\boldsymbol{X}) + \gamma \geq \gamma$. Then Proposition 1 indicates that we cannot identify all the invariant features. (b) The proof proceeds the same as above by noting $H(Y|\boldsymbol{X}_v, h(\boldsymbol{X}, Y)) = 0$. (c) This case can be shown similarly to (a), because $H(Y|X_v, h(\mathrm{Index}(\boldsymbol{X}, Y))) = H(Y|X_v, \boldsymbol{X}, Y) = 0$. □

A.5 Proof of Proposition 2

Proof. Denote a feature set $\boldsymbol{X}_{v(+k)}$, which contains the invariant feature set $\boldsymbol{X}_v$ as well a spurious feature $\boldsymbol{X}_s^k$ in Proposition 2.

Similar to Eqn. (9), we can show that

$$\hat{\mathcal{L}}(\Phi_{v(+k)}) \leq (1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_{v(+k)}).$$

And similar to Eqn. (10), we also have

$$\hat{\mathcal{L}}(\Phi_v) \geq -(1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_v).$$

Then it follows that

$$\begin{aligned}\hat{\mathcal{L}}(\Phi_v) - \hat{\mathcal{L}}(\Phi_{v(+k)}) \geq &- 2(1+2\lambda)\epsilon + H(Y|\boldsymbol{X}_v) - H(y|\boldsymbol{X}_{v(+k)}) \\ \geq &- 2(1+2\lambda)\epsilon + \gamma.\end{aligned}$$

If $\epsilon < \frac{\gamma}{2(1+2\lambda)}$, we have

$$\hat{\mathcal{L}}(\Phi_v) > \hat{\mathcal{L}}(\Phi_{v(+k)}).$$

Thus, we cannot identify all the invariant features from Problem 4. □

A.6 Proof of Theorem 3

Proof. Step 1: No spurious feature will be learned.

Given a partition $\{\boldsymbol{X}^k\}_{k=1}^K$ that lies in linear general position of degree r, Arjovsky et al. [5, Theorem 9] shows that Φ and ω satisfies the normal equations $\Phi\mathbb{E}_{\boldsymbol{X}^k}[\boldsymbol{X}^k\boldsymbol{X}^{k^T}]\Phi^T\omega = \Phi\mathbb{E}_{\boldsymbol{X}^k,Y^k}[\boldsymbol{X}^kY^k]$ for all k if and only if Φ elicits the desired invariant predictor $\Phi^T\omega = \tilde{\boldsymbol{W}}^T\beta$. Thus, we only need to show our solution meets the same normal equations. Let Φ' and ω' denote a solution to Problem 6. According to the constraint and the general linear position condition, we know there exists a partition $\{\boldsymbol{X}_k\}_{k=1}^K$ lies in general linear position of degree r, and we have $\mathcal{R}^k(\omega', \Phi') = \mathcal{R}^k(\omega_k, \Phi')$. Notice that ω_k minimizes the mean squared error on only the k-th environment, hence it must satisfy the normal equation $\Phi'\mathbb{E}_{\boldsymbol{X}^k}[\boldsymbol{X}^k\boldsymbol{X}^{k^T}]\Phi'^T\omega_k = \Phi'\mathbb{E}_{\boldsymbol{X}^k,Y^k}[\boldsymbol{X}^kY^k]$. As ω' achieves the same minimum mean squared error on the k-th environment, ω' must satisfy the normal equation, too. So no spurious information will in included and $\boldsymbol{X}\Phi = S\boldsymbol{X}_v$ where S is a invertible function.

Step 2: No invariant information will be discarded.

By Condition 1, we have $H(Y|X_v) - H(Y|X_v, \rho(Z)) = 0$. Then $\boldsymbol{X}\Phi = S\boldsymbol{X}_v$ will satisfy the constrain of Problem 6 for any environment generated by $\rho(Z)$. Notice that $S\boldsymbol{X}_v$ achieves the when smallest loss only using invariant feature information.

Putting these two steps together completes the proof. □

A.7 Proof of Meeting Condition 1

We show that $H(Y|\boldsymbol{X}_v, \rho(\boldsymbol{Z})) = H(Y|\boldsymbol{X}_v)$ for all $\rho(\cdot)$ if $H(Y|\boldsymbol{X}_v, \boldsymbol{Z}) = H(Y|\boldsymbol{X}_v)$ holds.

Proof. On one hand, because $\rho(\boldsymbol{Z})$ contains less information than $\boldsymbol{Z}$, so

$$H(Y|\boldsymbol{X}_v, \rho(\boldsymbol{Z})) \geq H(Y|\boldsymbol{X}_v, \boldsymbol{Z}) = H(Y|\boldsymbol{X}_v).$$

On the other hand, $\boldsymbol{X}_v$ and $\rho(\boldsymbol{Z})$ contain more information than $\boldsymbol{X}_v$, so we have

$$H(Y|\boldsymbol{X}_v, \rho(\boldsymbol{Z})) \leq H(Y|\boldsymbol{X}_v).$$

Thus, we conclude $H(Y|\boldsymbol{X}_v, \rho(\boldsymbol{Z})) = H(Y|\boldsymbol{X}_v)$. □

B Extension to Other Loss Functions

Sections 5.2 and 5.3 focus on classification task with cross-entropy loss, to establish sufficient and necessary conditions for invariance identification. Similar results can be shown for other loss functions or tasks in a straightforward manner. To see this, notice that $H(\cdot|\cdot)$ in Assumptions 1-4 and Conditions 1-2 is used to represent the optimal expected risk, which coincides with the conditional entropy with cross-entropy loss and when $\rho(\boldsymbol{z}_i)$ gives exactly one environment. For other loss functions $l(\cdot,\cdot)$ like squared error, we can use $L(\cdot|\cdot)$ to represent the optimal expected risks, and $L(\cdot)$ denotes the optimal expected risk with no predictor variables, e.g., variance for the squared error loss. Replacing $H(\cdot|\cdot)$ with $L(\cdot|\cdot)$, we can follow the same proof procedure to obtain similar sufficient and necessary conditions. This is summarized in Theorem 4 for completeness.

Theorem 4. *Consider Assumptions 1-4 and Conditions 1-2 where $H(\cdot|\cdot)$ is replaced with $L(\cdot|\cdot)$ accordingly. If $\epsilon < \frac{C\gamma\delta}{4\gamma+2C\delta L(Y)}$ and $\lambda \in [\frac{L(Y)+1/2\delta C}{\delta C-4\epsilon}-\frac{1}{2}, \frac{\gamma}{4\epsilon}-\frac{1}{2}]$, we have $\hat{\mathcal{L}}(\Phi_v) < \hat{\mathcal{L}}(\Phi), \quad \forall \Phi \neq \Phi_v$ where we assume $L(Y) < \infty$. If Condition 1 or Condition 2 is violated, then there exists a feature mask $\Phi' \neq \Phi_v$ so that $\hat{\mathcal{L}}(\Phi_v) > \hat{\mathcal{L}}(\Phi')$.*

C Discussion on Auxiliary Information for Related Methods

The proposed framework ZIN relies on additional auxiliary variables. It is interesting to ask whether this type of information is also useful to related methods such as EIIL and HRM. In many scenarios, the additional variable $\boldsymbol{Z}$ may have no or little information about the target, e.g., 1) in time series tasks, the time index is rarely used to predict the label; and 2) in the CelebA classification task, features like *Young* and *Eyeglasses* are likely to provide little information about predicting *Smiling*. Then $\boldsymbol{Z}$ may not be useful to ERM and EIIL. In Section 7.2, we also provide $\boldsymbol{Z}$ to EIIL and HRM, and the experimental result shows that it does not help or even harms the test performance.

D More Experimental Details

For the synthetic datasets and the house price prediction dataset, we use the full batch gradient in optimizing. The ERM method is trained for 4000 epochs. The IRM/ZIN method is also trained for 4000 epochs, with annealing in the first 2000 epochs. We train EIIL for 4000 epochs which is divided equally, i.e., 2000 epochs, in the two stages. For the CelebA classification task, we use mini-batch training with batch size of 128. All the methods are trained for 50 epochs, with annealing strategy in the first 25 epochs. For the landcover classification task, we use mini-batch training with batch size of 1024. All the methods are trained for 400 epochs, with anneal strategy in the first 40 epochs. We use Adam [21] with learning rate 0.001 as our optimizer.

Our experiments are run in a Linux workstation with Intel Xeon 3.20GHz CPU, 128GB RAM, and Nvidia GTX 3090 GPU. It takes about 5 GPU hours for the CelebA tasks, while the house price task and Landcover task cost less than 10 minutes for training.

E Extensions of Section 4

E.1 Connection of Impossibility Example to Existing Tasks

In the counter-example in Section 4, we consider binary features and label. This is inspired by the popular classification tasks in the literature of IRM, e.g., CMNIST [5], CifarMnist [24], ColoredObject [1] , Waterbirds and CelebA [38]. Take CMNIST for example. The label takes either 0 or 1. The invariant feature, X_1, is the semantic feature '0' or '1' of the digit shape. The spurious feature X_2 (color) is also binary: either red or green. So we will denote the binary invariant and spurious features as 0/1 in the following discussion.

Recall the the construction procedure of CMNIST in [5]:

$$X_1 \sim \text{Bernoulli}(0.5), \quad Y = \begin{cases} X_1, & w.p.\ 0.75, \\ 1-X_1, & w.p.\ 0.25, \end{cases} \quad X_2^e = \begin{cases} Y, & w.p.\ p_s^e, \\ 1-Y, & w.p.\ 1-p_s^e, \end{cases}$$

where X_1 (digit shape) is the invariant feature while X_2 (color) is spurious, and $p_s^{e=1} = 0.8$ and $p_s^{e=2} = 0.9$. Notably, in the test domain, $p^{e=3} = 0.1$ is chosen to simulate further distributional shift where the conditional probability $P(Y|X_2)$ is significantly different from the training domains.

The CMNIST construction process is analogous to the first data generation process in our counter-example in Section 4. When the two environments are mixed, we observe a joint distribution that can also be generated according to a different SCM. In the next section, we construct a dataset in accordance to the CMNIST dataset.

E.2 Supportive Experimental Results

We now construct a variant of CMNIST, analogous to the second data generation process in Example 1 in Section 4. In this dataset, color is the invariant feature and the digit shape is the spurious feature. The correlation between digit shape and label is unstable and it would change dramatically in the testing domain. We name this variant of CMNIST as MCOLOR (short for MNIST-COLOR). We use the same notations where X_1 stands for the digit shape, X_2 is the color and Y denotes label.

We generate the label in MCOLOR from the color and make the digit shape spuriously correlated with the label:

$$X_2 \sim \text{Bernoulli}(0.5), \quad Y = \begin{cases} X_2, & w.p.\ 0.85, \\ 1 - X_2, & w.p.\ 0.15, \end{cases} \qquad X_1^e = \begin{cases} Y, & w.p.\ p_s^e, \\ 1 - Y, & w.p.\ 1 - p_s^e, \end{cases}$$

where X_2 (color) is the invariant feature while X_1 (digit shape) is spurious, and $p_s^{e=1} = 0.8$ and $p_s^{e=2} = 0.7$. In the testing domain, we set $p_s^{e=3} = 0.1$ to simulate the distributional shift so that $P(Y|X_1)$ is significantly different from the training domains. If we mix the training domains of CMNIST or MCOLOR, we will observe the same distribution given by

$$\begin{cases} Y = 0, & w.p.\ 0.5, \\ Y = 1, & w.p.\ 0.5, \\ X_2 = X_1 = Y, & w.p.\ 0.6375, \\ X_2 \neq X_1 = Y, & w.p.\ 0.1125, \\ X_1 \neq X_2 = Y, & w.p.\ 0.2125, \\ X_2 = X_1 \neq Y, & w.p.\ 0.0375. \end{cases}$$

With this new dataset, we can then compare EIIL, IRM (with environment indexes), and ERM. The empirical results are reported in Table 5. For ERM (oracle), we train the model only on the invariant feature in MCOLOR, i.e., color.

Table 5: Experimental results on MCOLOR.

Method	Environment Index	Train Acc	Test Acc
ERM (oracle)	No	85.0±0.0	85.0±0.0
ERM	No	86.3±0.1	80.1±0.6
IRM	Yes	84.9±0.1	84.7±0.3
EIIL	No	74.0±0.4	17.8±0.4

In Table 5, we can see that the EIIL method performs poorly on MCOLOR. Due to the inductive bias of EIIL, it will deterministically rely on the digit shape as the invariant feature, resulting in poor performance on this MCOLOR dataset. However, since we have no prior knowledge of the data generation process, the TRUE invariant feature could be the color as in the MCOLOR dataset. LfF will also rely on either color or digit shape as the invariant feature and will fail on one of the CMNIST and MCOLOR datasets. On the other hand, if we have the environment indexes, IRM can still learn the color as the invariant feature on MCOLOR. We believe that this observation will also hold on CifarMnist, ColoredObject, Waterbird or CelebA.

Table 6: Test Mean and Worst accuracy (%) on four spatial heterogeneity synthetic datasets.

Env Partition	$p_s(t)$	0.999, 0.999, 0.7, 0.7				0.999, 0.9, 0.8, 0.7				0.999, 0.999, 0.8, 0.8			
	p_v	0.9		0.8		0.9		0.8		0.9		0.8	
	Test Acc	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst
No	ERM	76.65	59.48	60.33	27.25	76.59	59.35	60.30	27.25	69.93	44.65	56.23	16.60
	EIIL	37.81	16.89	66.46	50.35	37.03	14.01	66.60	50.28	70.18	45.23	71.00	58.72
	HRM	49.98	49.95	49.97	49.92	49.99	49.98	50.00	49.99	49.97	49.95	49.99	49.97
	ZIN	**88.66**	**87.23**	**79.16**	**78.04**	**88.28**	**86.29**	**78.92**	**77.49**	**88.00**	**85.75**	**78.80**	**77.25**
Yes	IRM	83.71	82.24	73.26	71.25	86.73	83.79	75.80	73.33	84.39	81.48	73.15	69.97

F More Experiment Results

F.1 Spatial Heterogeneity

We next consider spatial heterogeneity that is also commonly encountered in practice, e.g., environments may be divided according to locations of latitude and longitude. We simulate spatial heterogeneity in the same way the data generation process for time heterogeneity but use a two-dimensional spatial variable $\boldsymbol{r} = [r_1, r_2] \in [0, 1]^2$. We simulate four environments for training by equally splitting the space into four blocks, i.e., $\{[0, 0.5) \times [0, 0.5), [0, 0.5) \times [0.5, 1], [0.5, 1] \times [0, 0.5), [0.5, 1] \times [0.5, 1]\}$. Similarly, we denote a simulated case by tuple of $p_s(\boldsymbol{r})$ in the four elements. We also evaluate the performance on four distinct test environments with $p_s \in \{0.999, 0.8, 0.2, 0.1\}$ and p_v being constant. More details regarding the implementations are given in Appendix D.

The experimental results of spatial heterogeneity in Table 6 show similar performances to those of temporal heterogeneity. An interesting observation is that ZIN outperforms IRM with ground-truth environments in this simulation, especially when "duplicated" environments exist, e.g., when $p_s(\boldsymbol{r}) = (0.999, 0.999, 0.7, 0.7)$. We conjecture that in this case the "ground-truth" partition may not be the most effective for invariant learning due to the heavily overlapped environments. As shown in the right panel of Figure 1, ZIN automatically merges duplicated environments.

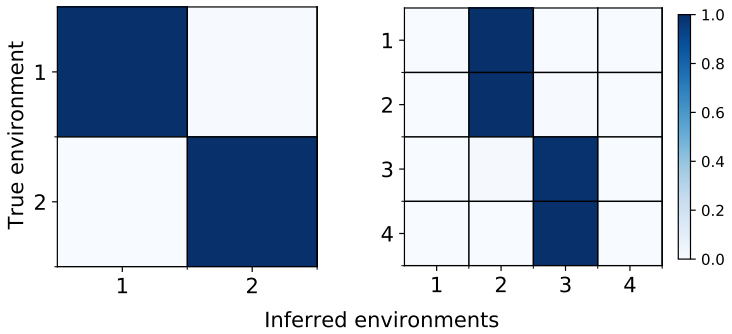

Figure 1: Visualization of inferred environments. **Left**: temporal heterogeneity setting $(0.999, 0.7)$. **Right**: spatial heterogeneity setting $(0.999, 0.999, 0.7, 0.7)$.

F.2 Ablation Study on Auxiliary Information and Number of Environments

We now verify the theoretical results in Section 5 by choosing different inputs as $\boldsymbol{Z}$. We adopt a setting of spatial heterogeneity with $p_s(\boldsymbol{r}) = (0.999, 0.999, 0.8, 0.8)$ and $p_v = 0.8$. Note that the heterogeneity in this setting is only along the second dimension r_2. The results are shown in Table 7. One can verify that Conditions 1 and 2 are satisfied when $\boldsymbol{r}$ or r_2 is chosen as $\boldsymbol{Z}$. The mean and worst test accuracy implies that ZIN based on $\boldsymbol{r}$ or r_2 can effectively remove the spurious feature. Since $p_s(\boldsymbol{r})$ does not change along r_1, choosing r_1 to infer environment partition violates Condition 2, which is reflected by the poor performance of ZIN with r_1. Using $[\boldsymbol{X}, Y]$ as input to $\rho(\cdot)$ is also a violation by Corollary 1, and the corresponding results in Table 7 confirm our analysis. Lastly, notice that $X_s \in \boldsymbol{X}$ contains some information of Y and there may exist a function $\rho(\cdot)$ so that

$H(Y|X_v) > H(Y|X_v, \rho(\boldsymbol{X}))$. This violates Condition 1 and leads to poor results when choosing $\boldsymbol{X}$ as input to $\rho(\cdot)$.

We also empirically verify the choice of hyper-parameter K using the same simulated setting. Figure 2 shows that K has a relatively small impact, especially when $K \geq 4$.

Table 7: Ablation study on choice of $\boldsymbol{Z}$.

$\boldsymbol{Z}$	Condition 1	Condition2	Test Mean	Test Worst
$\boldsymbol{r}$	✓	✓	78.80	77.25
r_1	✓	✗	56.30	16.84
r_2	✓	✓	78.79	77.21
$\boldsymbol{X}$	✗	✓	59.85	25.99
$\boldsymbol{X}, Y$	✗	✓	71.09	58.85

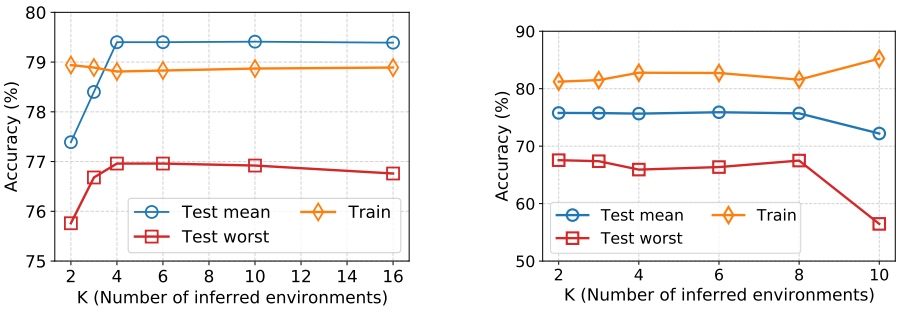

Figure 2: Ablation study on the choice of K. **Left**: synthetic dataset. **Right**: CelebA.

F.3 About the Inferred Environments on CelebA

We visualize the inferred environments of the CelebA dataset in this section. The target *Smiling* is spuriously correlated with feature *Gender*, i.e., most females are smiling while most males are not. We set K to be 2 (our framework is insensitive to K as shown in Appendix F.2). We visualize the spurious correlations in the two inferred environments during training in Fig. 3. ZIN can generate two environments where the spurious correlations differ. Then we can easily discard the spurious feature using IRM methods on the inferred environments.

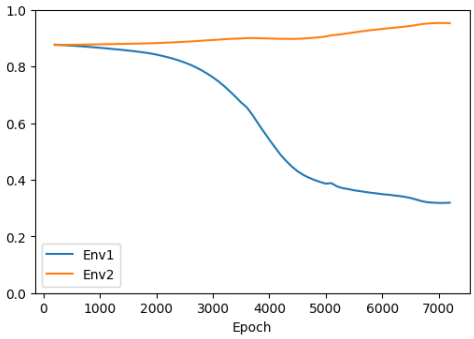

Figure 3: The correlation of the spurious feature (Gender) with the target (Smiling). The spurious correlation is calculated as the percentage of samples whose target (Smiling/Not Smiling) align with its gender (Female/Male). For example, a smiling female or a non-smiling male is counted as score 1, otherwise as score 0. The average score represents the correlation between Smiling and Gender.

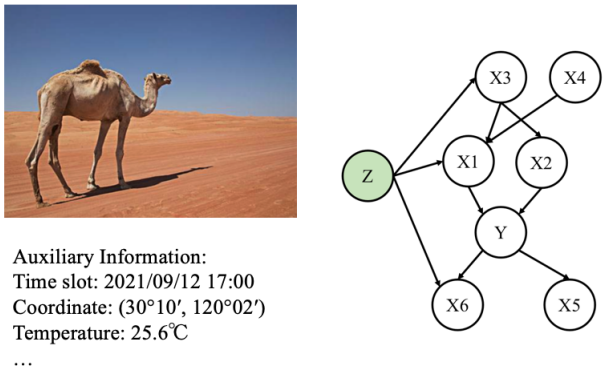

Figure 4: An illustration of a valid choice of Z satisfying Condition 1.

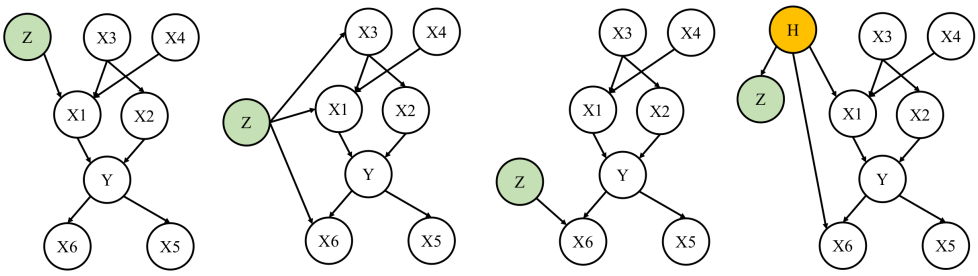

Figure 5: Examples of valid choices of Z satisfying Condition 1. Here "H" in the 4th graph denotes hidden confounders. The invariant features are X_1 and X_2, the direct causes of Y.

G How to Choose Z

Conditions 1 and 2 give the theoretical requirements of Z to successfully learn invariant features. In this section, we discuss how to find such auxiliary information in applications.

G.1 Checking Condition 1

We start with an example in Fig. 4. When we collect a sample (e.g., taking a photo), there often exists a set of meta information (e.g., the time slot, coordinates, temperature). Consider the label of the image (e.g., "camel") is generated by the causal graph on the right panel of Fig. 4. The invariant feature $\boldsymbol{X}_v$ are X_1 and X_2 in this case. The meta information is used as Z, which is colored in green in the causal graph. In this example, node Z has correlations with some nodes in the causal system. This Z satisfies Condition 1 because $H(Y|X_1, X_2) = H(Y|X_1, X_2, Z)$ or equivalently $Y \perp Z|X_1, X_2$. This example gives us some evidence of choosing Z: Z and Y should be d-separated by $\boldsymbol{X}_v$. There are other examples satisfying Condition 1, as shown in Fig. 5.

We also illustrate some failure choices of Z, illustrated in Fig. 6. Specifically, when Z is a child of Y or a direct cause of Y, Z contains additional information of Y conditional on X_1 and X_2. Then Condition 1 is violated and this kind of information cannot be used in our framework.

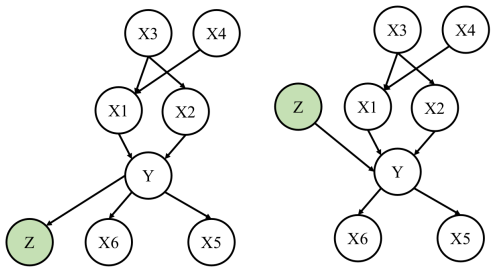

Figure 6: Invalid choices of Z that violate Condition 1. The invariant features are X_1 and X_2, the direct causes of Y.

G.2 Checking Condition 2

Though Condition 2 is needed to discard *all* spurious features, it is hard to check in practice because we generally have no prior knowledge on all the possible spurious features. Thus, we can collect as much auxiliary information as possible, which satisfies Condition 1 and also has correlations with some nodes in the causal system of interest. Then we can discard as many spurious features as possible while keeping the invariant features.

A natural question to ask is *what would happen if Condition 2 is only satisfied partially*, i.e., the condition holds only for some $\boldsymbol{X}_s^k \in \boldsymbol{X}_s$. The experimental results on CelebA in Table 3 already give us some hints. As we collect more auxiliary information, we find that the test performance increases. Even when we have limited (insufficient) auxiliary information, we can also obtain a performance gain over ERM, e.g., ZIN(4) outperforms ERM by a large margin. In this section, we present an additional theoretical result: *if Condition 2 is satisfied partially, we are able to discard spurious features partially and still preserve all invariant features.* The main result is given in the following.

Proposition 3. *Assume that Assumptions 1-4 and the following conditions are true:*

- *Condition 1 holds: given invariant feature $\boldsymbol{X}_v$ and any function $\rho(\cdot)$, it holds that $H(Y|\boldsymbol{X}_v, \rho(\boldsymbol{Z})) = H(Y|\boldsymbol{X}_v)$.*
- *Condition 2 holds **partially**: there exists a proper subset of spurious feature $\boldsymbol{X}_{\bar{s}} \subsetneq \boldsymbol{X}_s$ and for any $\boldsymbol{X}_{\bar{s}}^k \in \boldsymbol{X}_{\bar{s}}$, there exists a function $\rho(\cdot)$ and a constant $C > 0$ such that $H(Y|\boldsymbol{X}_{\bar{s}}^k) - H(Y|\boldsymbol{X}_{\bar{s}}^k, \rho(\boldsymbol{Z})) \geq C$. Further, let $\boldsymbol{X}_{sc} := \boldsymbol{X}_s \setminus \boldsymbol{X}_{\bar{s}}$, and for any function $\rho(\cdot)$, it holds that $H(Y|\boldsymbol{X}_{sc}, \rho(\boldsymbol{Z})) = H(Y|\boldsymbol{X}_{sc})$.*

Let Φ_{v+sc} denote the feature mask that selects both $\boldsymbol{X}_v$ and $\boldsymbol{X}_{sc}$. If $\epsilon < \frac{C\gamma\delta}{4\gamma+2C\delta H(Y)}$ and $\lambda \in [\frac{H(Y)+1/2\delta C}{\delta C-4\epsilon} - \frac{1}{2}, \frac{\gamma}{4\epsilon} - \frac{1}{2}]$, then we have $\hat{\mathcal{L}}(\Phi_{v+sc}) < \hat{\mathcal{L}}(\Phi)$ for all $\Phi \neq \Phi_{v+sc}$.

Proposition 3 shows that if Condition 1 is satisfied for $\boldsymbol{X}_v$ and Condition 2 is only satisfied for part of the spurious features $\boldsymbol{X}_{\bar{s}}$, then our framework can still successfully discard $\boldsymbol{X}_{\bar{s}}$, while keeping $\boldsymbol{X}_v$ and the rest of the spurious features.

Proof. The proof proceeds almost the same as that of Theorem 2, by replacing Φ_v with Φ_{v+sc}. Note that $\boldsymbol{X}_{v+sc}$ behaves as "invariant" under the inferred environments in Proposition 3. Discarding any member of $\boldsymbol{X}_{v+sc}$ will increase the ERM loss. Adding any member of $\boldsymbol{X}_{\bar{s}}$ also increases the loss due to a positive penalty (with a sufficient large weight λ). □

G.3 Take-away Message on Choosing Z

To summarize, we make the following suggestions to collect Z in practice:

- Pick Z that satisfies Condition 1 and that is correlated with nodes in the causal graph of interest. Further, the path between Z and Y should be d-separated by $\boldsymbol{X}_v$ as shown in Figures 4 and 5. Notably, Z cannot be the parent or the child of Y as shown in Figure 6.
- Collect as many such Z's as possible.

H Societal Impact

In this work, we propose to utilize the auxiliary information to aid the invariance learning without environmental indexes. This method is also helpful to fairness issues; see, e.g., the discussion about out-of-distribution generation and algorithmic fairness in [9]. For the additional information, we should also avoid using demographic, private, and sensitive information.