INCO: ENHANCE DOMAIN GENERALIZATION IN NOISY ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Invariant Risk Minimization (IRM) approach aims to address the challenge of domain generalization by training a feature representation that remains invariant across multiple environments. However, in noisy environments, IRM-related techniques such as IRMv1 and VREx may be unable to achieve the optimal IRM solution due to incorrect optimization directions. To address this issue, we introduce InCo (short for **I**nvariant **Corr**elation), a novel approach that effectively tackles the aforementioned challenges in noisy environments. Additionally, we provide a case study to analyze why previous methods may lose ground while InCo can succeed. We offer theoretical analysis from a causal perspective, demonstrating that the invariant correlation of representation with labels across environments is a necessary condition for the optimal invariant predictor in noisy environments, whereas the optimization motivations for other methods may not be. Subsequently, we empirically demonstrate the usefulness of InCo by comparing it with other domain generalization methods on various noisy datasets.

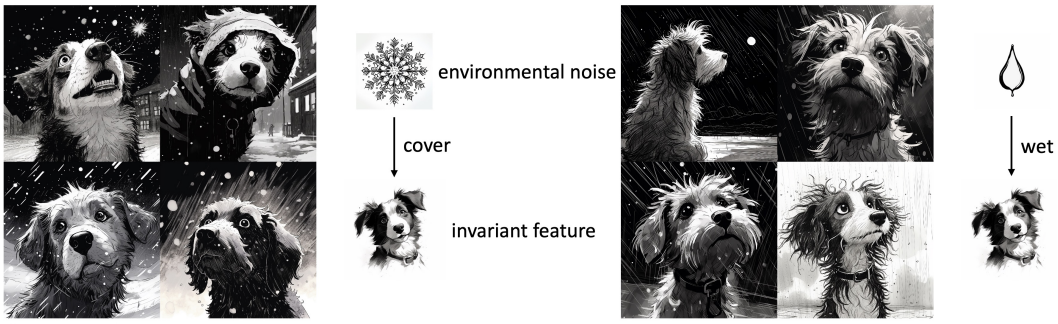
1 INTRODUCTION

Over the past decade, deep neural networks (DNNs) have made remarkable progress in a wide range of applications, such as computer vision (Simonyan & Zisserman, 2014; He et al., 2016; Krizhevsky et al., 2017) and natural language processing (Bahdanau et al., 2014; Luong et al., 2015). Typically, most deep learning models are trained using the Empirical Risk Minimization (ERM) (Vapnik, 1991) approach, which assumes that training and testing samples are independently drawn from an identical distribution (I.I.D. assumption). Nevertheless, recent studies have reported increasing instances of DNN failures (Beery et al., 2018; Geirhos et al., 2020; DeGrave et al., 2021) when this I.I.D. assumption is violated due to distributional shifts in practice.

Invariant Risk Minimization (Arjovsky et al., 2019) is a novel learning approach that addresses the challenge of domain generalization (also known as out of distribution problem) in the face of distributional shifts. The fundamental concept behind IRM is to train a feature representation that remains invariant across multiple environments (Peters et al., 2016), such that a single classifier can perform well in all of them. Although obtaining the optimal invariant feature representation is challenging, previous works employ alternative methods (Xu & Jaakkola, 2021; Shahtalebi et al., 2021; Zhang et al., 2023) to approximate it. The success of IRM approach in existing training environments can ensure its ability to generalize well in new environments with unseen distributional shifts, which is evidenced by positive empirical results (Rame et al., 2022; Chen et al., 2023).

However, in the real world, different environments (or domains) may exhibit varying levels of inherent (independent) noises, leading to various inherent losses. Even an optimal IRM model cannot mitigate these inherent losses, resulting in varying optimal losses across different environments. As shown in Fig. 1, inherent noise (such as snow or water) can impact the invariant feature (dog), such as covering the face or blurring the body, resulting in different inherent losses. Existing IRM-related methods, such as IRMv1, VREx (Krueger et al., 2021) and Fisher (Rame et al., 2022), focus on optimizing the model in different clean environments but may fail in these noisy situations.

We conduct an analysis in this study to identify the reasons why existing IRM-related methods may be ineffective in noisy environments. Upon examining the case study presented in Sec. 2.3, it has come to our attention that the optimization methods utilized for IRMv1, VREx and others may fail to converge to the optimal IRM solution due to environmental noise interference. Fortunately, our



(a) Dogs with snow (b) Dogs with water

Figure 1: A conceptual illustration of dogs in two environments: **(a)** snow, **(b)** water. As snow may cover the hair of dogs and water may wet the appearance, they can cause different environmental inherent losses. All images are generated by authors using Midjourney (`www.midjourney.com`).

proposed method (InCo) in Sec. 2.2 can effectively overcome these challenges, because independent environmental noise should have no effect on the correlation between invariant representation and label. Following the theoretical setting from Peters et al. (2016); Arjovsky et al. (2019), we also provide in Sec. 3 a formal theoretical analysis from the perspective of causality, demonstrating that the invariant correlation across environments (i.e., the optimization idea of InCo) is a necessary condition for the (optimal) invariant predictor in noisy environments, while the optimization motivations for others may not be. Furthermore, in Sec. 4, we conduct a comprehensive range of experiments to confirm the effectiveness of InCo in noisy environments.

We summarize the contributions and novelties of this work as follows:

- We propose InCo (in Sec. 2.2), which enforces the correlation constraint throughout training process, and demonstrate its benefits through theoretical analysis of causality (in Sec. 3).
- We present the motivation of InCo through a case study in Sec. 2.3, which reveals that when in noisy environments, previous IRM-related methods may fail to get the optimal IRM solution because of environmental inherent noises, whereas InCo can still converge to the optimal IRM solution.
- An extensive set of empirical results is provided to demonstrate that InCo can generalize better in noisy environments across different datasets when compared with other domain generalization methods (Sec. 4).

2 Study IRM in Noisy Environments

2.1 Preliminaries

Given that $\mathcal{X}$ and $\mathcal{Y}$ are the input and output spaces respectively, let $\mathcal{E} := \{e_1, e_2, ..., e_m\}$ be a collection of m environments in the sample space $\mathcal{X} \times \mathcal{Y}$ with different joint distributions $\mathbb{P}^e(\boldsymbol{x}^e, y)$, where $e \in \mathcal{E}$. Consider $\mathcal{E}_{tr} \subset \mathcal{E}$ to be the training environments and $\mathcal{S}^e := \{(\mathbf{x}_i^e, \mathrm{y}_i)\}_{i=1}^{n^e}$ to be the training dataset drawn from distribution $\mathbb{P}^e(\boldsymbol{x}^e, y)$ ($e \in \mathcal{E}_{tr}$) with n^e being dataset size. Given the above training datasets $\mathcal{S}^e$ ($e \in \mathcal{E}_{tr}$), the task is to learn an optimal model $f(\cdot; \boldsymbol{w}) : \mathcal{X} \to \mathcal{Y}$, such that $f(\boldsymbol{x}^e; \boldsymbol{w})$ performs well in predicting y when given $\boldsymbol{x}^e$ not only for $e \in \mathcal{E}_{tr}$ but also for $e \in \mathcal{E} \setminus \mathcal{E}_{tr}$, where $\boldsymbol{w}$ is the parameters of f.

The ERM algorithm (Vapnik, 1991) tries to solve the above problem via directly minimizing the loss throughout training environments:

$$\min_{\boldsymbol{w}:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} \mathcal{R}^e(\boldsymbol{w}), \tag{ERM}$$

where $\mathcal{R}^e(\boldsymbol{w})$, $\mathcal{R}(\boldsymbol{x}^e, \boldsymbol{w})$ are the expected loss of $f(\cdot; \boldsymbol{w})$ in e, loss of $f(\boldsymbol{x}^e; \boldsymbol{w})$, respectively.

IRM (Arjovsky et al., 2019) firstly supposes that the predictor $f(\cdot; \boldsymbol{w})$ can be made up of $g(\cdot; \Phi)$ and $h(\cdot; \boldsymbol{v})$, i.e., $f(\cdot; \boldsymbol{w}) = h(g(\cdot; \Phi); \boldsymbol{v})$, where $\boldsymbol{w} = \{\boldsymbol{v}, \Phi\}$ are the model parameters. Here,

$g(\cdot;\Phi) : \mathcal{X} \to \mathcal{H}$ extracts invariant features among $\mathcal{E}_{tr}$ through mapping $\mathcal{X}$ to the representation space $\mathcal{H}$. The classifier $h(\cdot;\boldsymbol{v}) : \mathcal{H} \to \mathcal{Y}$ is supposed to be simultaneously optimal for all training environments. The original IRM method learns $g(\cdot;\Phi)$ and $h(\cdot;\boldsymbol{v})$ through solving the following minimization problem:

$$\begin{aligned} &\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ \boldsymbol{v}:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{R}^e(\{\boldsymbol{v},\Phi\}) \\ &s.t. \quad \boldsymbol{v} \in \arg\min_{\bar{\boldsymbol{v}}:\mathcal{H}\to\mathcal{Y}} \mathcal{R}^e(\{\bar{\boldsymbol{v}},\Phi\}), \text{ for all } e \in \mathcal{E}_{tr}. \end{aligned} \tag{IRM}$$

However, IRM remains a bi-level optimization problem. Arjovsky et al. (2019) suggests, for practical reasons, to relax this strict limitation by using the method IRMv1 as a close approximation to IRM:

$$\min_{\boldsymbol{w}:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\text{tr}}} \left[\mathcal{R}^e(\boldsymbol{w}) + \lambda \left|\left|\nabla_{\boldsymbol{v}|\boldsymbol{v}=1}\mathcal{R}^e(\boldsymbol{w})\right|\right|^2\right], \tag{IRMv1}$$

where $\boldsymbol{v} = 1$ is a scalar and fixed "dumm" classifier. Furthermore, VREx (Krueger et al., 2021) adopts the following regularizer for robust optimization:

$$\min_{\boldsymbol{w}:\mathcal{X}\to\mathcal{Y}} \lambda \cdot \text{Var}(\mathcal{R}^e(\boldsymbol{w})) + \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{R}^e(\boldsymbol{w}), \tag{VREx}$$

where $\text{Var}(\mathcal{R}^e(\boldsymbol{w}))$ represents the variance of the losses $\mathcal{R}^e(\boldsymbol{w})$ in $\mathcal{E}_{tr}$. Clearly, to encourage $f(\cdot;\boldsymbol{w})$ to be simultaneously optimum, IRMv1 constrains the gradients $\nabla_{\boldsymbol{v}|\boldsymbol{v}=1}\mathcal{R}^e(\boldsymbol{w})$ to be 0 and VREx decreases the loss variance $\text{Var}(\mathcal{R}^e(\boldsymbol{w}))$ to 0.

2.2 Invariant correlation of representation with label

We now formally describe our method (InCo) to extract invariant features in noisy environments. InCo performs robust learning via stabilizing the correlation between representation and true label across environments:

$$\min_{\boldsymbol{w}:\mathcal{X}\to\mathcal{Y}} \lambda \cdot \text{Var}(\rho^e_{f,y}(\boldsymbol{w})) + \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{R}^e(\boldsymbol{w}), \tag{InCo}$$

where $\rho^e_{f,y}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}^e,y}(\tilde{f}(\boldsymbol{x}^e;\boldsymbol{w})y)$ is the correlation between $f(\boldsymbol{x}^e;\boldsymbol{w})$ and y in the environment e, $\tilde{f}(\boldsymbol{x}^e;\boldsymbol{w}) = f(\boldsymbol{x}^e;\boldsymbol{w}) - \mathbb{E}_{\boldsymbol{x}^e}(f(\boldsymbol{x}^e;\boldsymbol{w}))$, and $\text{Var}(\rho^e_{f,y}(\boldsymbol{w}))$ represents the variance of the correlation in $\mathcal{E}_{tr}$. Here $\lambda \in [0,+\infty)$ controls the balance between reducing average loss and enhancing stability of correlation, with $\lambda = 0$ recovering ERM, and $\lambda \to +\infty$ leading InCo to focus entirely on making the correlation equal. In the following, we demonstrate the power of InCo in noisy environments through the case study (Sec. 2.3) and the theoretical analysis of causality (Sec. 3), respectively.

2.3 Why is InCo necessary (a case study in two-bit environments)

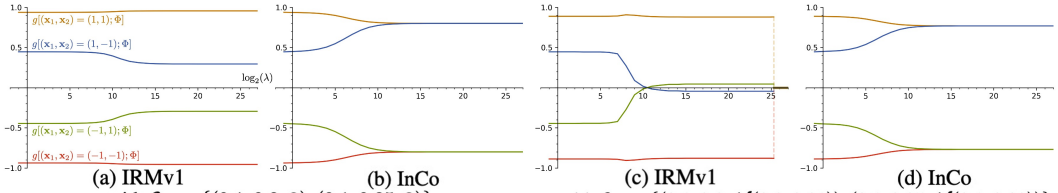
Arjovsky et al. (2019) presents the Colored-MNIST task, a synthetic challenge derived from MNIST, to demonstrate the efficacy of the IRM technique and IRMv1 in particular. Although MNIST pictures are grayscale, Colored-MNIST images are colored red or green in a manner that strongly (but spuriously) correlates with the class label. In this case, ERM successfully learns to exploit the color during training, but it fails at test time when the correlation with the color is inverted.

Kamath et al. (2021) studies an abstract version of Colored-MNIST based on two bits of input, where y is the label to be predicted, $\hat{\boldsymbol{x}}_1$ is correlated with the label of the hand-written digit ($0-4$ or $5-9$), and $\hat{\boldsymbol{x}}_2^e$ corresponds to the color (red or green).

Setting: Following Kamath et al. (2021), we initially represent each environment e with two parameters $\alpha, \beta^e \in [0,1]$. The data generation process is then defined as

$$\begin{aligned} \text{Invariant feature: } \hat{\boldsymbol{x}}_1 &\leftarrow \text{Rad}(0.5), \\ \text{True label: } y &\leftarrow \hat{\boldsymbol{x}}_1 \cdot \text{Rad}(\alpha), \\ \text{Spurious feature: } \hat{\boldsymbol{x}}_2^e &\leftarrow y \cdot \text{Rad}(\beta^e), \end{aligned} \tag{1}$$

where $\text{Rad}(\delta)$ is a random variable taking value -1 with probability δ and $+1$ with probability $1-\delta$. In addition, we also consider an environmental inherent noise η^e. That is, we can only observe the



(a) IRMv1 with $\mathcal{E}_{tr}=\{(0.1,0.2,\mathbf{0}),(0.1,0.25,\mathbf{0})\}$ (b) InCo with $\mathcal{E}_{tr}=\{(0.1,0.2,\mathbf{0}),(0.1,0.25,\mathbf{0})\}$ (c) IRMv1 with $\mathcal{E}_{tr}=\{(0.1,0.2,\mathcal{N}(0.2,0.01)),(0.1,0.25,\mathcal{N}(0.1,0.02))\}$ (d) InCo with $\mathcal{E}_{tr}=\{(0.1,0.2,\mathcal{N}(0.2,0.01)),(0.1,0.25,\mathcal{N}(0.1,0.02))\}$

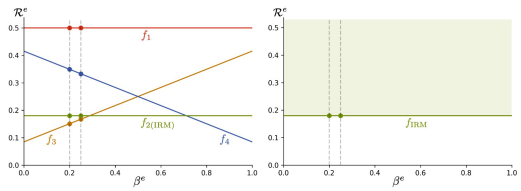
Figure 2: The output (vertical axis) of optimized $g(\boldsymbol{x}^e;\Phi)$ with four inputs $(\mathbf{x}_1,\mathbf{x}_2)=\{(1,1),(1,-1),(-1,1),(-1,-1)\}$. The horizontal axis is $\log_2(\lambda)$, with -1 representing $\lambda=0$. **(a)**, **(b)** are the results of IRMv1 and InCo for varying λ optimized with training environments $\mathcal{E}_{tr}=\{(0.1,0.2,\mathbf{0}),(0.1,0.25,\mathbf{0})\}$. **(c)**, **(d)** are the results of IRMv1 and InCo optimized with $\mathcal{E}_{tr}=\{(0.1,0.2,\mathcal{N}(0.2,0.01)),(0.1,0.25,\mathcal{N}(0.1,0.02))\}$. More results are given in App. A.

Table 1: The square losses for optimal IRM (oracle) and other optimization methods: ERM, IRMv1($\lambda=+\infty$), VREx($\lambda=+\infty$), InCo($\lambda=+\infty$). All losses in this table are computed with $\eta^e=\mathbf{0}$, **left** methods are optimized with training environments $\mathcal{E}_{tr}=\{(0.1,0.2,\mathbf{0}),(0.1,0.25,\mathbf{0})\}$, whereas **right** ones are optimized with $\mathcal{E}_{tr}=\{(0.1,0.2,\mathcal{N}(0.2,0.01)),(0.1,0.25,\mathcal{N}(0.1,0.02))\}$. The upper two rows are the results with training β^e (0.2 and 0.25), whereas the lower two rows present the results when the correlation of $\hat{\boldsymbol{x}}_2^e$ has flipped ($\beta^e=0.7,0.9$). In addition, we also provide more results of other methods in App. A. Best results are in **bold**.

$\mathcal{R}(\alpha,\beta^e,\eta^e)$	$\mathcal{E}_{tr}=\{(0.1,0.2,\mathbf{0}),(0.1,0.25,\mathbf{0})\}$					$\mathcal{E}_{tr}=\{(0.1,0.2,\mathcal{N}_{[0.2,0.01]}),(0.1,0.25,\mathcal{N}_{[0.1,0.02]})\}$				
	Oracle	ERM	IRMv1	VREx	InCo	Oracle	ERM	IRMv1	VREx	InCo
$\mathcal{R}(0.1,0.2,\mathbf{0})$	**0.18**	0.15	0.15	**0.18**	**0.18**	**0.1805**	0.15	0.50	0.50	**0.1805**
$\mathcal{R}(0.1,0.25,\mathbf{0})$	**0.18**	0.16	0.17	**0.18**	**0.18**	**0.1805**	0.16	0.50	0.50	**0.1805**
$\mathcal{R}(0.1,0.7,\mathbf{0})_{tst}$	**0.18**	0.26	0.32	**0.18**	**0.18**	**0.1805**	0.25	0.50	0.50	**0.1805**
$\mathcal{R}(0.1,0.9,\mathbf{0})_{tst}$	**0.18**	0.30	0.38	**0.18**	**0.18**	**0.1805**	0.30	0.50	0.50	**0.1805**

features interfered by environmental noise:

$$\begin{aligned}\text{Observed invariant feature: } \boldsymbol{x}_1^e &\leftarrow \hat{\boldsymbol{x}}_1+\eta^e,\\ \text{Observed spurious feature: } \boldsymbol{x}_2^e &\leftarrow \hat{\boldsymbol{x}}_2^e+\eta^e,\end{aligned} \tag{2}$$



(a) IRMv1 (b) VREx and InCo

Figure 3: The solutions for **(a)** IRMv1, **(b)** VREx and InCo when $\lambda=+\infty$. I.e., the solutions satisfy **(a)** $\nabla_{\boldsymbol{v}|\boldsymbol{v}=1}\mathcal{R}^e(\boldsymbol{w})=0$, **(b)** $\text{Var}(\mathcal{R}^e(\boldsymbol{w}))=0$ and $\text{Var}(\rho^e_{f,y}(\boldsymbol{w}))=0$ for $\mathcal{E}_{tr}=\{e_1=(0.1,0.2,\mathbf{0}),e_2=(0.1,0.25,\mathbf{0})\}$. The horizontal axis is β^e and vertical axis represents square loss for $e=(0.1,\beta^e,\mathbf{0})$. The solid circles are training losses for different solutions. Clearly, **(a)** picks f_3 and **(b)** picks f_{IRM}.

where $\eta^e\sim\mathcal{N}(\mu^e,(\sigma^e)^2)$ is an independent Gaussian noise. Then, for convenience, we denote an environment e as (α,β^e,η^e), where α represents invariant correlation between $\hat{\boldsymbol{x}}_1$ and y, β^e represents varying (non-invariant) correlation between $\hat{\boldsymbol{x}}_2^e$ and y across $\mathcal{E}$, η^e is the environmental inherent noise. We consider a linear model $(f(\boldsymbol{x}^e;\boldsymbol{w})_{\boldsymbol{v}=1}=g(\boldsymbol{x}^e;\Phi)=w_1\boldsymbol{x}_1^e+w_2\boldsymbol{x}_2^e)$ with square loss $\mathcal{R}_{sq}(\hat{y},y):=\frac{1}{2}(\hat{y}-y)^2$ in this case study. All methods are optimized in training environments $\mathcal{E}_{tr}=\{(0.1,0.2,\eta^{e_1}),(0.1,0.25,\eta^{e_2})\}$ with $\eta^{e_{1,2}}=\mathbf{0}$ (Case 1) or $\eta^{e_{1,2}}\neq\mathbf{0}$ (Case 2).

Case 1: Optimization without environmental inherent noise.

In the first case, $\mathcal{E}=\mathcal{E}_{\alpha=0.1}$ with training environments $\mathcal{E}_{tr}=\{e_1=(0.1,0.2,\mathbf{0}),e_2=(0.1,0.25,\mathbf{0})\}$, our results are similar to Kamath et al. (2021).

• **Failure of IRMv1:** Consider that IRMv1, VREx, InCo become exactly ERM when their regularization terms are $\lambda=0$. Fig. 2(a) shows the output of $g(\boldsymbol{x}^e;\Phi)$ from IRMv1 ($\lambda=0$, ERM) to IRMv1 ($\lambda=+\infty$) with four inputs. Note that IRMv1 with a specific λ is optimized by training environments $\mathcal{E}_{tr}$. We find that $g((1,-1);\Phi)$ decreases and $g((-1,1);\Phi)$ increases with growing λ; this phenomenon demonstrates the reliance on $\boldsymbol{x}_2^e$ increases when $\lambda\to+\infty$. Thus IRMv1 may find an un-invariant predictor even worse than ERM. This is also echoed by the results in Tab. 1(left): When the correlation of $\hat{\boldsymbol{x}}_2^e$ has flipped ($\beta^e=0.7,0.9$) in the test environment, the performance of the predictor from IRMv1 ($\lambda=+\infty$) may be worse than that learnt by optimal IRM and even worse than ERM.

• **Success of VREx, InCo:** Fortunately, VREx and InCo can still converge to the optimal IRM with increasing λ, as stabilizing losses (VREx) or correlations (InCo) across different training environments can effectively prevent the interference from spurious feature in this case. Fig. 2(b) demonstrates that $g(\boldsymbol{x}^e;\Phi)$ from InCo only relies on invariant feature $\boldsymbol{x}_1^e$ when $\lambda \geq 2^{11}$. Furthermore, VREx ($\lambda = +\infty$) and InCo ($\lambda = +\infty$) in Tab. 1(left) perform the same as optimal IRM in all training and test environments.

• **Why:** As shown in Fig. 3(a) (Kamath et al., 2021), there are four solutions for IRMv1 when $\lambda \to +\infty$. Unfortunately, IRMv1 picks f_3 rather than optimal IRM solution (f_2) as f_3 has the lowest training loss of those four solutions. Clearly, f_3 relies more on $\boldsymbol{x}_2^e$ and damages the performance when flipping β^e. On the other hand, Fig. 3(b) shows VREx and InCo can easily converge to the optimal IRM solution when minimizing training losses for any two training environments. The details of calculating procedure are given in App. B.1.

Case 2: Optimization with environmental inherent noise.

In the second case, we further consider training environments with environmental inherent noise, i.e., $\mathcal{E}_{tr} = \{(0.1, 0.2, \mathcal{N}(0.2, 0.01)), (0.1, 0.25, \mathcal{N}(0.1, 0.02))\}$.

• **Failure of IRMv1:** As shown in Fig. 2(c), compared with clean training environments in Fig. 2(a), noisy training environments may make IRMv1 more reliant on $\boldsymbol{x}_2^e$ when $\lambda \in [2^{10}, 2^{25.3}]$, and finally IRMv1 converges to a zero solution ($w_1 = 0, w_2 = 0$) with a non-continuous step when $\lambda > 2^{25.3}$. Thus the loss for IRMv1 ($\lambda = +\infty$) in Tab. 1(right) is 0.5 across all environments. This finding is consistent with our calculation in App. B.1, which demonstrates that IRMv1 ($\lambda = +\infty$) has only one zero solution.

• **Failure of VREx:** In noisy training environments, VREx ($\lambda = +\infty$) in Tab. 1(right) also fails to extract invariant feature, since minimizing $\mathrm{Var}(\mathcal{R}^e(\boldsymbol{w}))$ cannot help find the optimal invariant predictor when there are different environmental inherent noises. As shown in Fig. 6(a) from App. A, VREx also converges to a zero solution when $\lambda \to +\infty$.

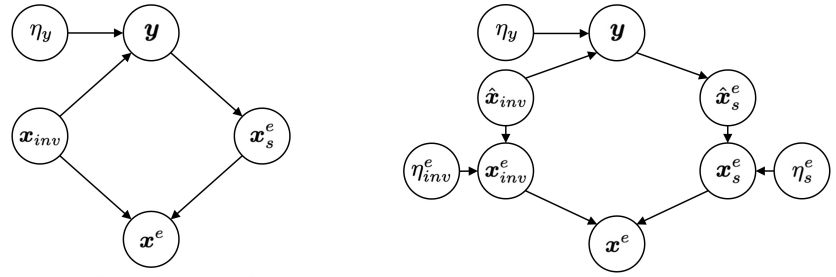
• **Success of InCo:** InCo can deal with this case as its regularization term only considers the correlation between representation and true label. In other words, it can filter out the impact of environmental noise which is independent of true label. The results in Fig. 2(d) show that InCo still converges to IRM solution in noisy training environments and Tab. 1(right) shows that InCo ($\lambda = +\infty$) has the same results with optimal IRM (oracle).

• **Why:** Due to the variability of environmental inherent losses, optimizing $||\nabla_{\boldsymbol{v}|\boldsymbol{v}=1}\mathcal{R}^e(\boldsymbol{w})|| \to 0$ or $\mathrm{Var}(\mathcal{R}^e(\boldsymbol{w})) \to 0$ may be impractical in noisy training environments. That is to say, if an optimal IRM predictor operates in noisy training environments, there may exist $||\nabla_{\boldsymbol{v}|\boldsymbol{v}=1}\mathcal{R}^e(\boldsymbol{w})|| \neq 0$ and $\mathrm{Var}(\mathcal{R}^e(\boldsymbol{w})) \neq 0$ due to different environmental inherent noises. Nevertheless, the independence between η^e and y ensures that $\mathrm{Var}(\rho^e_{f,y}(\boldsymbol{w})) = 0$ holds for the optimal IRM predictor. Details of the calculation are given in App. B.1. (We also provide formal proofs under a more general setting for the above claims in the next section.)

• **Failure of other methods:** In addition, gradient-based optimization methods for optimal IRM can also be unsuccessful in noisy environments. In this noisy case with $\mathcal{E}_{tr} = \{e_1, e_2\}$, IGA (Koyama & Yamaguchi, 2020) minimizes $||\nabla_{\boldsymbol{w}}\mathcal{R}^{e_1}(\boldsymbol{w}) - \nabla_{\boldsymbol{w}}\mathcal{R}^{e_2}(\boldsymbol{w})||_2^2$, Fish (Shi et al., 2021) increases $\nabla_{\boldsymbol{w}}\mathcal{R}^{e_1}(\boldsymbol{w}) \cdot \nabla_{\boldsymbol{w}}\mathcal{R}^{e_2}(\boldsymbol{w})$, AND-mask (Parascandolo et al., 2020) and Mansilla et al. (2021) update weights only when $\nabla_{\boldsymbol{w}}\mathcal{R}^{e_1}(\boldsymbol{w})$ and $\nabla_{\boldsymbol{w}}\mathcal{R}^{e_2}(\boldsymbol{w})$ point to the same direction, Fishr (Rame et al., 2022) reduces $||\mathrm{Var}(\nabla_{\boldsymbol{w}}\mathcal{R}(\boldsymbol{x}^{e_1}, \boldsymbol{w})) - \mathrm{Var}(\nabla_{\boldsymbol{w}}\mathcal{R}(\boldsymbol{x}^{e_2}, \boldsymbol{w}))||_2^2$. Clearly, they may be failed in noisy environments as their penalty terms are also affected by environmental inherent noises. We provide more simulation and calculation results for some of these methods in App. A and App. B.2 respectively.

3 THEORETICAL ANALYSIS FROM CAUSAL PERSPECTIVE

In this section, we present our theoretical understanding of InCo from the perspective of causality. Following the theoretical setting from Peters et al. (2016); Arjovsky et al. (2019), we formally prove that (1) $\mathrm{Var}(\rho^e_{f,y}(\boldsymbol{w})) = 0$ is a necessary condition for the optimal invariant predictor in noisy environments; (2) $||\nabla_{\boldsymbol{v}|\boldsymbol{v}=1}\mathcal{R}^e(\boldsymbol{w})|| = 0$, $\mathrm{Var}(\mathcal{R}^e(\boldsymbol{w})) = 0$ and some other minimal penalty terms may not be necessary for the optimal invariant predictor in noisy environments.



(a) DAG (Arjovsky et al., 2019) (b) DAG (ours)

Figure 4: Comparison of the DAG form **(a)** (Arjovsky et al., 2019) and **(b)** ours. Different from **(a)**, the observed invariant feature $\boldsymbol{x}_{inv}^e$ in **(b)** is affected by the environmental inherent noise η_{inv}^e, such as snow covering the face or water blurring the body in Fig. 1.

Setting: Consider several training environments $\mathcal{E}_{tr} = \{e_1, e_2, ...\}$ and $\boldsymbol{x}^e$ to be the observed input of $e \in \mathcal{E}_{tr}$. We adopt an anti-causal framework (Arjovsky et al., 2019) with data generation process as follows:

$$
\begin{aligned}
&y = \gamma^\top \hat{\boldsymbol{x}}_{inv} + \eta_y, \\
&\boldsymbol{x}_{inv}^e = \hat{\boldsymbol{x}}_{inv} + \eta_{inv}^e, \quad \boldsymbol{x}_s^e = \hat{\boldsymbol{x}}_s^e + \eta_s^e, \\
&\boldsymbol{x}^e = S\begin{pmatrix} \boldsymbol{x}_{inv}^e \\ \boldsymbol{x}_s^e \end{pmatrix},
\end{aligned}
$$

where $\gamma \in \mathbb{R}^{d_{inv}}$ and $\gamma \neq \mathbf{0}$, the hidden invariant feature $\hat{\boldsymbol{x}}_{inv}$ and the observed invariant feature $\boldsymbol{x}_{inv}^e$ take values in $\mathbb{R}^{d_{inv}}$, the hidden spurious feature $\hat{\boldsymbol{x}}_s^e$ and the observed spurious feature $\boldsymbol{x}_s^e$ take values in $\mathbb{R}^{d_s}$, and $S : \mathbb{R}^{(d_{inv}+d_s)} \to \mathbb{R}^d$ is an inherent mapping to mix features. The hidden spurious feature $\hat{\boldsymbol{x}}_s^e$ is generated by y with any *non-invariant* relationship, η_{inv}^e and η_s^e are independent Gaussian with bounded mean and variance changed by environments, η_y is an independent and invariant zero-mean Gaussian with bounded variance. As the directed acyclic graph (DAG) in Fig. 4(b) shows, the hidden invariant feature $\hat{\boldsymbol{x}}_{inv}$ generates the true label y and y generates the hidden spurious feature $\hat{\boldsymbol{x}}_s^e$. In consideration of environmental noise, we can only observe the input $\boldsymbol{x}^e$ which is a mixture of $\boldsymbol{x}_{inv}^e$ and $\boldsymbol{x}_s^e$ after mapping. (Note that the observed feature is generated by applying environmental noise to the hidden feature.) We aim to learn a classifier to predict y based on $\boldsymbol{x}^e$, i.e., $f(\boldsymbol{x}^e; \boldsymbol{w}) = h(g(\boldsymbol{x}^e; \Phi); \boldsymbol{v})$.

Drawing upon the foundational assumption from IRM (Arjovsky et al., 2019), i.e., assume that there exists a mapping $\tilde{S} : \mathbb{R}^d \to \mathbb{R}^{d_{inv}}$ such that $\tilde{S}(S(\begin{smallmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{smallmatrix})) = \mathbf{x}_1$ for all $\mathbf{x}_1 \in \mathbb{R}^{d_{inv}}$ and $\mathbf{x}_2 \in \mathbb{R}^{d_s}$, the following theorem mainly states that, in noisy environments, if there exists a representation Φ that elicits the optimal invariant predictor $f(\cdot; \boldsymbol{w})$ across all possible environments $\mathcal{E}$, then the correlation between $f(\boldsymbol{x}^e; \boldsymbol{w})$ and y remains invariant for all $e \in \mathcal{E}$.

Theorem 3.1 *Assume that there exists a mapping $\tilde{S} : \mathbb{R}^d \to \mathbb{R}^{d_{inv}}$ such that $\tilde{S}(S(\begin{smallmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{smallmatrix})) = \mathbf{x}_1$ for all $\mathbf{x}_1 \in \mathbb{R}^{d_{inv}}, \mathbf{x}_2 \in \mathbb{R}^{d_s}$. Then, if Φ elicits the desired (optimal) invariant predictor $f(\cdot; \boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$, we have $\mathrm{Var}(\rho_{f,y}^e(\boldsymbol{w})) = 0$.*

Thm. 3.1 indicates that in noisy environments, minimizing the regularization term of InCo, i.e., $\mathrm{Var}(\rho_{f,y}^e(\boldsymbol{w}))$, is a necessary condition to find the invariant features. The intuition behind Thm. 3.1 is that, the correlation between the representation and the true label can effectively prevent interference in noisy environments, whereas IRMv1 and VREx may get stuck. In the following, we would like to point out that the regularization strategies employed in IRMV1, VREx and others may not be the most effective.

Corollary 3.2 *If Φ elicits the desired invariant predictor $f(\cdot; \boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$, then there exists e satisfies*

$$
\frac{\partial \mathcal{R}^e(\boldsymbol{w})}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}} \neq 0
$$

in noisy environments.

Cor. 3.2 suggests that $||\nabla_{\boldsymbol{v}|\boldsymbol{v}=1} \mathcal{R}^e(\boldsymbol{w})|| = 0$ (IRMv1) may not be a necessary condition for the optimal invariant predictor in noisy environments, as environmental inherent losses can lead to non-

Table 2: Comparison of MLP on ColoredMNIST with varying training noises, i.e., first training environment without noise, second training environment with $\mathbf{0}$, $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(0, 1)$, respectively. We repeat each experiment with 100 times and report the best, worst and average accuracies (%) on the test environment with $\mathbf{0}$, $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(0, 1)$, respectively. Best results are in **bold**.

Test noise	Method	$\{\mathbf{0}, \mathbf{0}\}_{\text{train}}$			$\{\mathbf{0}, \mathcal{N}(0, 0.5)\}_{\text{train}}$			$\{\mathbf{0}, \mathcal{N}(0, 1)\}_{\text{train}}$		
		Best	Worst	Mean	Best	Worst	Mean	Best	Worst	Mean
$\mathbf{0}$	ERM	50.85	10.08	27.08	51.77	17.70	35.38	51.71	10.48	35.64
	IRMv1	70.12	63.31	67.46	50.65	17.36	36.92	50.42	10.13	31.19
	VREx	**70.84**	64.80	69.02	58.66	23.50	43.18	51.69	14.43	32.98
	CLOvE	69.07	41.32	64.97	34.00	10.61	15.83	50.61	10.77	31.41
	Fishr	70.48	**66.01**	**69.07**	50.18	20.50	36.98	50.87	9.87	27.01
	InCo	70.56	65.25	68.33	**69.40**	**26.69**	**53.73**	**68.11**	**18.18**	**44.16**
$\mathcal{N}(0, 0.5)$	ERM	51.44	22.63	36.11	51.71	13.18	32.98	51.63	12.28	32.94
	IRMv1	59.59	53.22	56.75	51.19	11.61	32.01	50.89	10.28	29.33
	VREx	58.73	53.52	56.61	51.35	30.44	41.97	51.54	13.69	35.09
	CLOvE	49.28	36.10	44.87	42.45	20.33	31.40	49.76	23.32	41.37
	Fishr	**62.64**	**57.10**	**60.54**	51.40	26.28	38.17	50.36	10.87	30.63
	InCo	59.38	53.02	56.32	**64.66**	**35.43**	**57.17**	**67.09**	**23.96**	**49.00**
$\mathcal{N}(0, 1)$	ERM	50.90	32.70	42.36	51.54	20.61	36.94	50.99	18.31	35.86
	IRMv1	55.31	49.94	52.91	51.08	18.95	36.16	51.19	15.34	32.72
	VREx	54.20	50.33	52.55	50.85	34.57	43.87	51.47	22.92	39.29
	CLOvE	47.39	40.19	45.05	46.12	31.23	39.65	49.83	33.78	45.29
	Fishr	**57.76**	**53.48**	**55.81**	51.05	34.63	42.17	51.15	17.33	34.90
	InCo	54.51	50.36	52.65	**60.06**	**44.56**	**55.27**	**63.51**	**40.14**	**52.26**

zero $||\nabla_{\boldsymbol{v}|\boldsymbol{v}=1}\mathcal{R}^e(\boldsymbol{w})||$. Even in clean environments without noise, $||\nabla_{\boldsymbol{v}|\boldsymbol{v}=1}\mathcal{R}^e(\boldsymbol{w})|| = 0$ may point to other predictors rather than the optimal invariant one (Case 1 in Sec. 2.3).

Corollary 3.3 *If Φ elicits the desired invariant predictor $f(\cdot; \boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$, there exists $\eta_{inv}^{e_1} \neq \eta_{inv}^{e_2}$ in noisy environments $\{e_1, e_2\}$ such that*

$$\text{Var}(\mathcal{R}^e(\boldsymbol{w})) \neq 0.$$

Cor. 3.3 shows that $\text{Var}(\mathcal{R}^e(\boldsymbol{w}))$ (REx) may also be failed to represent as an indicator for the optimal invariant predictor in noisy environments. Given different inherent losses across environments, it seems unreasonable to enforce all losses to be equal. In App. C, we further prove that the regularization terms for IGA, Fishr and IB-ERM (Ahuja et al., 2021) may also be not necessary to find the optimal invariant predictor in such noisy situations. All proofs are given in App. D.

In conclusion, in this section, we examine InCo from a causal perspective and provide theoretical analysis that minimizing $\text{Var}(\rho_{f,y}^e(\boldsymbol{w}))$ is a necessary condition to find the invariant features in noisy environments. On the other hand, IRMv1, VREx and others may be ineffective in obtaining the optimal invariant predictor due to the impact of environmental noise on their regularization terms.

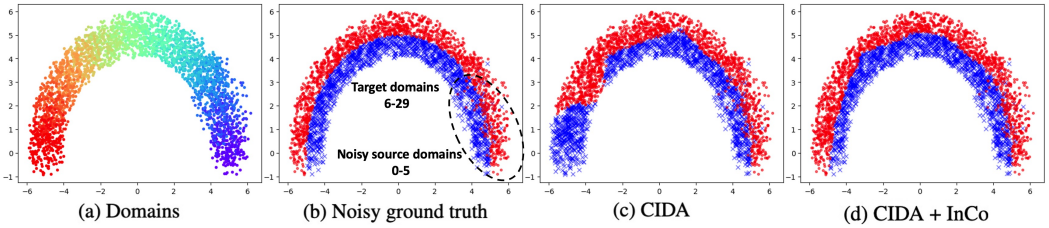
4 Experiments

In this section, we implement extensive experiments with ColoredMNIST (Arjovsky et al., 2019), Circle dataset (Wang et al., 2020a) and noisy DomainBed (Gulrajani & Lopez-Paz, 2020) framework. The first part includes comprehensive experiments on ColoredMNIST using multi-layer-perceptrons (MLP) with varying environmental noises. In the second part, we conduct further experiments to verify the effectiveness of InCo in extracting invariant features in noisy environments.

4.1 MLP with ColoredMNIST

Training setting: This proof-of-concept experiment of ColoredMNIST follows the settings from Arjovsky et al. (2019); Krueger et al. (2021). The MLP consists of two hidden layers with 256 and 256 units respectively. Each of these hidden layers is followed by a ReLU activation function. The final output layer has an output dimension of number of classes. All networks are trained with the Adam optimizer, ℓ_2 weight decay 0.001, learning rate 0.001, batchsize 25000 and epoch 500. Note that we use the exactly same hyperparameters as Arjovsky et al. (2019); Krueger et al. (2021), only replacing the IRMv1 penalty and VREx penalty with InCo penalty and other penalties.

ColoredMNIST setting: We create three MNIST environments (two training and one test) by modifying each example as follows: firstly, give the input a binary label $\tilde{y}$ depending on the digit: $\tilde{y} = 0$



(a) Domains (b) Noisy ground truth (c) CIDA (d) CIDA + InCo

Figure 5: Results on the noisy circle dataset with 30 domains.

for digits 0 to 4 and $\tilde{y} = 1$ for digits 5 to 9; secondly, define the final true label y by randomly flipping $\tilde{y}$ with a probability 0.25; the third step is to randomly choose the color id c by flipping y with probability $\mathbb{P}_c^e$, where $\mathbb{P}_c^e$ is 0.2 in the first environment, 0.1 in the second environment, and 0.9 in the test environment. Finally, if c is 1, the image is colored in red, otherwise it is colored in green.

Evaluating setting: There are three training groups in our experiments: $\{\mathbf{0}, \mathbf{0}\}$, $\{\mathbf{0}, \mathcal{N}(0, 0.5)\}$ and $\{\mathbf{0}, \mathcal{N}(0, 1)\}$. Specifically, the first training environment is clean without noise (i.e., $\mathbf{0}$ across all three groups), the second training environment differs in three groups: non-noise $\mathbf{0}$ in the first group, noise $\mathcal{N}(0, 0.5)$ in the second group and noise $\mathcal{N}(0, 1)$ in the third group. We train each network in these three training groups respectively with 100 times. Note that, following Krueger et al. (2021), we only record the test accuracy which is less than corresponding training accuracy for each experiment. We then report the best, average and worst performances (among 100 runs) in the test domain with environmental noise $\mathbf{0}$, $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(0, 1)$, respectively.

Remark: As shown in Tab. 2, there is no significant difference in the performances of IRMv1, VREx and InCo (Fishr performs relatively better and CLOvE (Wald et al., 2021) performs relatively worse) when trained in clean environments (first thick column). However, InCo is the only method to efficiently tackle noisy training environments (second and third thick columns). For example, with the training group $\{\mathbf{0}, \mathcal{N}(0, 0.5)\}$, InCo can achieve 69.4% best accuracy in the clean test environment, others can only get up to 58.66%; InCo can achieve 57.17% average accuracy in the $\mathcal{N}(0, 0.5)$ noisy test environment, while VREx, Fishr and IRMv1 only get 41.97%, 38.17% and 32.01%, respectively.

4.2 More empirical results

The **Circle Dataset** (Wang et al., 2020a) consists of 30 domains, with indices ranging from 0 to 29. The domains are depicted in Fig. 5(a) using distinct colors (in ascending order from 0 to 29, from right to left). Each domain consists of data related to a circle, and the objective is to perform binary classification. Fig. 5(b) illustrates the positive samples as red dots and negative samples as blue crosses. We utilize domains 0 to 5 as source domains (inside dashed circle), and the remaining domains as target domains. To create noisy environments during training, we apply Gaussian noises $\mathcal{N}(0, \text{index}/10)$ to source domains 0 to 5, respectively, while keeping target domains 6 to 29 clean. All other settings are same with Wang et al. (2020a). As shown in Fig. 5(c), the performance of CIDA (Wang et al., 2020a) in noisy training environments is not good enough, but it can be improved by adding the InCo penalty term as depicted in Fig. 5(d).

To further substantiate the effectiveness of InCo, we conduct an evaluation within the **DomainBed** (Gulrajani & Lopez-Paz, 2020) framework with two datasets: noisy PACS (Li et al., 2017) and noisy VLCS (Fang et al., 2013). In these datasets, we introduce environmental noise in the form of small Gaussian perturbations, denoted as $\mathcal{N}(0, i/5)$, where i represents the index of the respective environment. As shown in Tab. 3, our findings reveal that InCo consistently exhibits marked improvements in noisy environments when contrasted with other methods such as ERM, IRMv1, VREx, GroupDRO (Sagawa et al., 2019), and Fishr. Although Fishr demonstrates superior performance in the first environment of noisy PACS, the discernible accuracy difference between InCo and Fishr is minimal, amounting to merely 0.1%. In all other environments across the two datasets, InCo consistently delivers the highest level of performance. More empirical results and details are given in App. E.

5 Related work

The domain generalization problem is initially explicitly described by Blanchard et al. (2011) and then defined by Muandet et al. (2013), which takes into account the potential of the target data

Table 3: Domain generalization performances using noisy DomainBed evaluation protocol (with small environmental noises). All methods are trained with default hyper-parameter. We choose the checkpoint using the test domain validation set and report the corresponding test domain accuracy (%).

	Noisy PACS				Noisy VLCS			
	A	C	P	S	C	L	S	V
ERM	86.5	83.9	94.3	83.4	97.8	65.1	69.9	**79.2**
IRMv1	88.1	85.2	**96.4**	73.1	96.1	67.9	72.1	77.6
VREx	86.7	84.3	95.2	84.8	96.4	67.4	73.4	76.4
GroupDRO	87.8	84.7	95.2	81.3	97.1	67.9	70.8	77.4
Fishr	**89.6**	82.0	94.3	84.9	98.5	63.2	70.5	**79.2**
InCo	89.5	**85.5**	**96.4**	**86.4**	**98.9**	**69.6**	**73.7**	**79.2**

being unavailable during model training. A large body of literature seeks to address the domain generalization challenge, typically through additional regularizations of ERM (Vapnik, 1991). The regularizations from Motiian et al. (2017); Namkoong & Duchi (2016); Sagawa et al. (2019) enhance model robustness against minor distributional perturbations in the training distributions, Zhang et al. (2022b); Liu et al. (2021a); Yao et al. (2022) further improve this robustness with extra assumptions, while regularizations of Ganin et al. (2016); Sun & Saenko (2016); Li et al. (2018c;b); Dou et al. (2019); Zhao et al. (2019) promote domain invariance of learned features.

In addition, there has been a growing trend towards integrating the principle of causal invariance (Pearl, 2009; Louizos et al., 2017; Goudet et al., 2018; Ke et al., 2019; Schölkopf et al., 2021) into representation learning (Peters et al., 2016; Arjovsky et al., 2019; Creager et al., 2021; Parascandolo et al., 2020; Wald et al., 2021; Ahuja et al., 2021). In this context, the IRM (Arjovsky et al., 2019) approach has been proposed to extract features that remain consistent across various environments, following the invariance principle introduced in Peters et al. (2016). As of late, there have been several IRM-related methods developed in the community. Ahuja et al. (2020a) offers novel perspectives through the incorporation of game theory and regret minimization into invariant risk minimization. Ahuja et al. (2021) proposes to combine the information bottleneck constraint with invariance to address the case in which the invariant features capture all the information of the label. Zhou et al. (2022) studies IRM for overparameterized models. Ahuja et al. (2020b); Liu et al. (2021b) endeavor to learn invariant features when explicit environment indices are not provided. Chen et al. (2022) suggests utilizing the inherent low-dimensional structure of spurious features to recognize invariant features in logarithmic environments. Rosenfeld et al. (2020) studies IRM in the non-linear regime and finds it can fail catastrophically. Kamath et al. (2021) analyzes the success and failure cases of IRMv1 in clean environments. Zhang et al. (2022a) proposes constructing diverse initializations to stabilize domain generalization performance under the trade-off between ease of optimization and robust of domain generalization. Due to space limit, we provide more related work in App. F.

In contrast to prior research, this paper investigates IRM in noisy environments where environmental noises can corrupt invariant features. As a result, previous IRM-related approaches may not be effective in such scenarios. Nevertheless, our InCo technique can successfully handle noisy cases by utilizing the principle that the correlation of invariant representation with label is invariant across noisy environments.

6 Conclusion

In this paper, we introduced an IRM-related method named InCo, which utilizes the correlation between representation and labels to overcome the challenge of training an invariant predictor in noisy environments. In a case study with two-bit environments, we analyzed why other methods may fail while InCo can succeed in noisy environments. Through some theoretical analyses of causality, we demonstrated the necessity of invariant correlation across noisy environments for the optimal IRM solution. Moreover, we conducted extensive experiments which demonstrate the superior performance of InCo compared to other methods in such noisy scenarios.

Reproducibility Statement: We provide the detailed calculation process for Sec. 2.3 in App. B, the proofs for Sec. 3 in App. D, and the code for the experiments in the supplemental file.

REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155. PMLR, 2020a.

Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R Varshney. Empirical or invariant risk minimization? a sample complexity perspective. *arXiv preprint arXiv:2010.16412*, 2020b.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, 2019.

Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 2022.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 2021.

Yining Chen, Elan Rosenfeld, Mark Sellke, Tengyu Ma, and Andrej Risteski. Iterative feature matching: Toward provable domain generalization with logarithmic environments. *Advances in Neural Information Processing Systems*, 35:1725–1736, 2022.

Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, MA KAILI, Han Yang, Peilin Zhao, Bo Han, et al. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. In *The Eleventh International Conference on Learning Representations*, 2023.

Yo Joong Choe, Jiyeon Ham, and Kyubyong Park. An empirical study of invariant risk minimization, 2020.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, pp. 1657–1664, 2013.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. *Explainable and interpretable models in computer vision and machine learning*, pp. 39–80, 2018.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. In *ICLR*, 2023.

Hyungu Kahng, Hyungrok Do, and Judy Zhong. Domain generalization via heckman-type selection models. In *ICLR*, 2023.

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. In *ICLR*, 2023.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

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*, pp. 5815–5826. PMLR, 2021.

Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Out-of-distribution robustness via disagreement. In *ICLR*, 2023.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1446–1455, 2019a.

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, 2018b.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 624–639, 2018c.

Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *International Conference on Machine Learning*, pp. 3915–3924. PMLR, 2019b.

Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16021–16030, 2022.

Evan Z 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 *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021a.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Kernelized heterogeneous risk minimization. *arXiv preprint arXiv:2110.12425*, 2021b.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Lucas Mansilla, Rodrigo Echeveste, Diego H Milone, and Enzo Ferrante. Domain generalization via gradient surgery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6630–6638, 2021.

Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5715–5725, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pp. 10–18. PMLR, 2013.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

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)*, 78(5):947–1012, 2016.

Kha Pham, Hung Le, Man Ngo, and Truyen Tran. Improving out-of-distribution generalization with indirection representations. In *ICLR*, 2023.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022.

Alexander Robey, George J. Pappas, and Hamed Hassani. Model-based domain generalization. In *Advances in Neural Information Processing Systems*, 2021.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2020.

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. *arXiv preprint arXiv:1911.08731*, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

Jie-Jing Shao, Lan-Zhe Guo, Xiao-Wen Yang, and Yu-Feng Li. Log: Active model adaptation for label-efficient ood generalization. *Advances in Neural Information Processing Systems*, 35: 11023–11034, 2022.

Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Anoopkumar Sonar, Vincent Pacelli, and Anirudha Majumdar. Invariant policy optimization: Towards stronger generalization in reinforcement learning. In *Learning for Dynamics and Control*, pp. 21–33. PMLR, 2021.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, and Wen-Sheng Chu. Rethinking domain generalization for face anti-spoofing: Separability and alignment. *arXiv preprint arXiv:2303.13662*, 2023.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

Tan Wad, Qianru Sun, Sugiri Pranata, Karlekar Jayashree, and Hanwang Zhang. Equivariance and invariance inductive bias for learning from insufficient data. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pp. 241–258. Springer, 2022.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in neural information processing systems*, 34:2215–2227, 2021.

Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *ICML*, 2020a.

Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, pp. 159–176. Springer, 2020b.

Xinyi Wang, Michael Saxon, Jiachen Li, Hongyang Zhang, Kun Zhang, and William Yang Wang. Causal balancing for domain generalization. In *ICLR*, 2023a.

Zhe Wang, Jake Grigsby, and Yanjun Qi. PGrad: Learning principal gradients for domain generalization. In *ICLR*, 2023b.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Runpeng Yu, Hong Zhu, Kaican Li, Lanqing Hong, Rui Zhang, Nanyang Ye, Shao-Lun Huang, and Xiuqiang He. Regularization penalty optimization for addressing data quality variance in ood algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8945–8953, 2022.

Jianyu Zhang, David Lopez-Paz, and Léon Bottou. Rich feature construction for the optimization-generalization dilemma. In *International Conference on Machine Learning*, pp. 26397–26411. PMLR, 2022a.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022b.

YiFan Zhang, Xue Wang, Jian Liang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Free lunch for domain adversarial training: Environment label smoothing. *arXiv preprint arXiv:2302.00194*, 2023.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, 2019.

Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020.

Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pp. 27222–27244. PMLR, 2022.

The appendices can be summarized as follows:

- App. A: We present more case study results mentioned in Section 2.3.
- App. B: We provide the calculation details of IRMv1, VREx, InCo and other methods.
- App. C: We provide more causality analysis given the theoretical settings mentioned in Section 3.
- App. D: We provide the detailed proofs for Thm. 3.1, Cor. 3.2 and Cor. 3.3.
- App. E: We present more experiments and details.
- App. F: In this section, we provide more related work about domain generalization and IRM.

A MORE CASE STUDY RESULTS

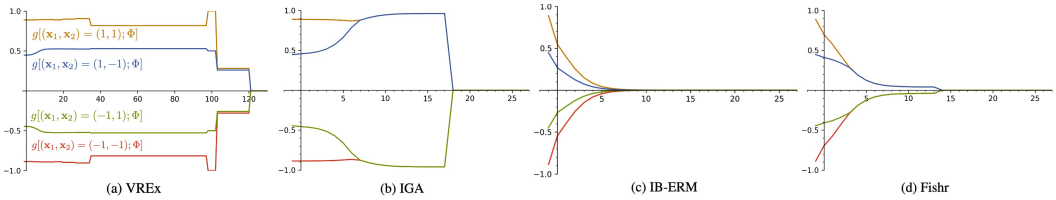
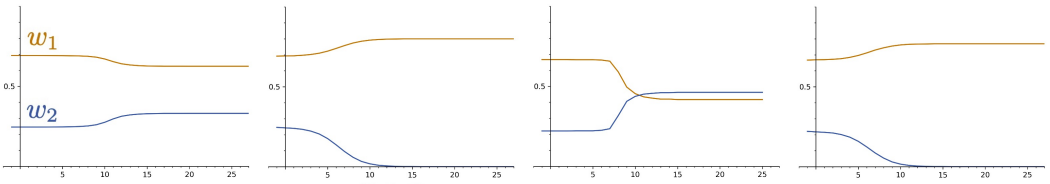


Figure 6: The output (vertical axis) of optimized $g(\boldsymbol{x}^e;\Phi)$ with four inputs $(\mathbf{x}_1,\mathbf{x}_2) = \{(1,1),(1,-1),(-1,1),(-1,-1)\}$. The horizontal axis is $\log_2(\lambda)$, with -1 representing $\lambda = 0$. **(a)**, **(b)**, **(c)**, **(d)** are the results of VREx, IGA, IB-ERM and Fishr for varying λ optimized with training environments $\mathcal{E}_{tr} = \{(0.1, 0.2, \mathcal{N}(0.2, 0.01)), (0.1, 0.25, \mathcal{N}(0.1, 0.02))\}$. Note that in **(a)** we let $\lambda = +\infty$ when $\lambda > 2^{120}$ due to numerical problems.

Table 4: The square losses for optimal IRM (oracle) and different optimization methods: IGA($\lambda = +\infty$ and 2^7), Fishr($\lambda = +\infty$ and 2^4), IB-ERM($\lambda = +\infty$). All losses in this table are computed with $\eta^e = \mathbf{0}$, and all methods are optimized with $\mathcal{E}_{tr} = \{(0.1, 0.2, \mathcal{N}(0.2, 0.01)), (0.1, 0.25, \mathcal{N}(0.1, 0.02))\}$. The upper two rows are the results with training β^e (0.2 and 0.25), whereas the lower two rows present the results when the correlation of $\hat{\boldsymbol{x}}_2^e$ has flipped ($\beta^e = 0.7, 0.9$).

$\mathcal{R}(\alpha, \beta^e, \eta^e)$	$\mathcal{E}_{tr} = \{(0.1, 0.2, \mathcal{N}_{[0.2,0.01]}), (0.1, 0.25, \mathcal{N}_{[0.1,0.02]})\}$					
	Oracle	IGA	IGA($\lambda = 2^7$)	Fishr	Fishr($\lambda = 2^4$)	IB-ERM
$\mathcal{R}(0.1, 0.2, \mathbf{0})$	**0.1805**	0.50	0.36	0.50	0.40	0.50
$\mathcal{R}(0.1, 0.25, \mathbf{0})$	**0.1805**	0.50	0.36	0.50	0.40	0.50
$\mathcal{R}(0.1, 0.7, \mathbf{0})_{tst}$	**0.1805**	0.50	0.36	0.50	0.40	0.50
$\mathcal{R}(0.1, 0.9, \mathbf{0})_{tst}$	**0.1805**	0.50	0.36	0.50	0.40	0.50

As shown in Fig. 6, we present the output of $g(\boldsymbol{x}^e;\Phi)$ which is optimized in noisy training environments $\mathcal{E}_{tr} = \{(0.1, 0.2, \mathcal{N}(0.2, 0.01)), (0.1, 0.25, \mathcal{N}(0.1, 0.02))\}$ with varying λ. **(a)** and **(c)** show that VREx and IB-ERM converge to zero solutions when $\lambda \to +\infty$. The results of IGA and Fishr are presented in **(b)** and **(d)**, respectively. Both methods converge to invariant solutions when $\lambda \geq 2^7$ for IGA and $\lambda \geq 2^4$ for Fishr, and finally they also achieve zero solutions. However, as shown in Tab. 4, these invariant solutions for IGA ($\lambda = 2^7$) and Fishr ($\lambda = 2^4$) are not optimal, as optimal loss is 0.1805 but IGA($\lambda = 2^7$) and Fishr($\lambda = 2^4$) only get 0.36 and 0.40 respectively. Note that here we choose 2^7 for IGA and 2^4 for Fishr because they are the best λ for corresponding invariant solutions. Fortunately, the results in Tab. 1 and Fig. 2(d) demonstrate the effectiveness of InCo to achieve optimal IRM solution (oracle) in this noisy case, because InCo can protect the training procedure from environmental noises. Note that all of these simulation results are consistent with our calculation in App. B. In Fig. 7, we show the change of w_1 and w_2 with respect to λ.



(a) IRMv1 with $\mathcal{E}_{tr} = \{(0.1, 0.2, \mathbf{0}), (0.1, 0.25, \mathbf{0})\}$ (b) InCo (c) IRMv1 with $\mathcal{E}_{tr} = \{(0.1, 0.2, \mathcal{N}(0.2, 0.01)), (0.1, 0.25, \mathcal{N}(0.1, 0.02))\}$ (d) InCo

Figure 7: The vertical axis is the value of w_1 and w_2 for optimized $g(\boldsymbol{x}^e; \Phi)$. The horizontal axis is $\log_2(\lambda)$, with -1 representing $\lambda = 0$. **(a)**, **(b)** are the results of IRMv1 and InCo for varying λ optimized with training environments $\mathcal{E}_{tr} = \{(0.1, 0.2, \mathbf{0}), (0.1, 0.25, \mathbf{0})\}$. **(c)**, **(d)** are the results of IRMv1 and InCo optimized with $\mathcal{E}_{tr} = \{(0.1, 0.2, \mathcal{N}(0.2, 0.01)), (0.1, 0.25, \mathcal{N}(0.1, 0.02))\}$.

Table 5: The square losses for optimal IRM (oracle) and other optimization methods: ERM, IRMv1($\lambda = +\infty$), VREx($\lambda = +\infty$), InCo($\lambda = +\infty$). All losses in this table are computed with **(left)** $\eta^e = \mathcal{N}(0.2, 0.01)$ and **(right)** $\eta^e = \mathcal{N}(0.1, 0.02)$, all methods are optimized with $\mathcal{E}_{tr} = \{(0.1, 0.2, \mathcal{N}(0.2, 0.01)), (0.1, 0.25, \mathcal{N}(0.1, 0.02))\}$. The upper two rows are the results with training β^e (0.2 and 0.25), whereas the lower two rows present the results when the correlation of $\hat{\boldsymbol{x}}_2^e$ has flipped ($\beta^e = 0.7, 0.9$).

$\mathcal{R}(\alpha, \beta^e, \eta^e)$	$\eta^e = \mathcal{N}(0.2, 0.01)$					$\eta^e = \mathcal{N}(0.1, 0.02)$				
	Oracle	ERM	IRMv1	VREx	InCo	Oracle	ERM	IRMv1	VREx	InCo
$\mathcal{R}(0.1, 0.2, \eta^e)$	**0.1953**	0.17	0.50	0.50	**0.1953**	**0.1894**	0.16	0.50	0.50	**0.1894**
$\mathcal{R}(0.1, 0.25, \eta^e)$	**0.1953**	0.18	0.50	0.50	**0.1953**	**0.1894**	0.17	0.50	0.50	**0.1894**
$\mathcal{R}(0.1, 0.7, \eta^e)_{tst}$	**0.1953**	0.27	0.50	0.50	**0.1953**	**0.1894**	0.27	0.50	0.50	**0.1894**
$\mathcal{R}(0.1, 0.9, \eta^e)_{tst}$	**0.1953**	0.32	0.50	0.50	**0.1953**	**0.1894**	0.31	0.50	0.50	**0.1894**

B CALCULATION DETAILS

B.1 CALCULATION FOR IRMV1, VREX AND INCO

Following Kamath et al. (2021) and Léon Bottou, we provide the calculation details of IRMv1, VREx and InCo solutions as follows.

Suppose $\mathcal{E}_{tr}$ consists of two environments $e_1 = (\alpha, \beta^{e_1}, \eta^{e_1})$ and $e_2 = (\alpha, \beta^{e_2}, \eta^{e_2})$. From the definition of IRMv1, VREx, InCo, for any $f(\boldsymbol{x}^e) = 1 \cdot g(\boldsymbol{x}^e; \Phi) = w_1 \boldsymbol{x}_1^e + w_2 \boldsymbol{x}_2^e$ with square loss, we have that:

when optimizing IRMv1 till $\nabla_{\boldsymbol{v}|\boldsymbol{v}=1} \mathcal{R}^e(\boldsymbol{w}) = 0$, we get

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}^{e_1}, y} \left(w_1 \boldsymbol{x}_1^{e_1} + w_2 \boldsymbol{x}_2^{e_1} - y\right) \left(w_1 \boldsymbol{x}_1^{e_1} + w_2 \boldsymbol{x}_2^{e_1}\right) = 0, \\ &\mathbb{E}_{\boldsymbol{x}^{e_2}, y} \left(w_1 \boldsymbol{x}_1^{e_2} + w_2 \boldsymbol{x}_2^{e_2} - y\right) \left(w_1 \boldsymbol{x}_1^{e_2} + w_2 \boldsymbol{x}_2^{e_2}\right) = 0; \end{aligned} \tag{3}$$

when optimizing VREx till $\mathrm{Var}(\mathcal{R}^e(\boldsymbol{w})) = 0$, we get

$$\mathbb{E}_{\boldsymbol{x}^{e_1}, y} \left(w_1 \boldsymbol{x}_1^{e_1} + w_2 \boldsymbol{x}_2^{e_1} - y\right)^2 = \mathbb{E}_{\boldsymbol{x}^{e_2}, y} \left(w_1 \boldsymbol{x}_1^{e_2} + w_2 \boldsymbol{x}_2^{e_2} - y\right)^2; \tag{4}$$

when optimizing InCo with $\mathrm{Var}[\rho_{f,y}^e(\boldsymbol{w})] = 0$, we get

$$\mathbb{E}_{\boldsymbol{x}^{e_1}, y}(w_1 \boldsymbol{x}_1^{e_1} y + w_2 \boldsymbol{x}_2^{e_1} y) = \mathbb{E}_{\boldsymbol{x}^{e_2}, y}(w_1 \boldsymbol{x}_1^{e_2} y + w_2 \boldsymbol{x}_2^{e_2} y). \tag{5}$$

Case 1: For both $\eta^{e_1} = \mathbf{0}$ and $\eta^{e_2} = \mathbf{0}$, we have (i) $\mathbb{E}[(\boldsymbol{x}_1^e)^2] = \mathbb{E}[(\boldsymbol{x}_2^e)^2] = 1$, (ii) $\mathbb{E}(\boldsymbol{x}_1^{e_i} y) = a$, $\mathbb{E}(\boldsymbol{x}_2^{e_i} y) = b_i$, (iii) $\mathbb{E}(\boldsymbol{x}_1^{e_i} \boldsymbol{x}_2^{e_i}) = a b_i$, where $a := 1 - 2\alpha$ and $b_i := 1 - 2\beta^{e_i}$ for $i \in \{1, 2\}$.

Then, according to equation 3, the solutions for IRMv1 ($\lambda = +\infty$) are

$$\begin{aligned} &(1) \quad w_1 = 0, w_2 = 0; \\ &(2) \quad w_1 = a, w_2 = 0; \\ &(3) \quad w_1 = \frac{1}{2a}, w_2 = \sqrt{\frac{1}{2} - \frac{1}{4a^2}}, \quad s.t. \quad a^2 > \frac{1}{2}; \\ &(4) \quad w_1 = \frac{1}{2a}, w_2 = -\sqrt{\frac{1}{2} - \frac{1}{4a^2}}, \quad s.t. \quad a^2 > \frac{1}{2}, \ w_2 \neq 0. \end{aligned}$$

According to (4), the solutions for VREx ($\lambda = +\infty$) are:

$$\begin{aligned} &(1) \quad w_1 = \frac{1}{a}, w_2 \in \mathbb{R}; \\ &(2) \quad w_1 \in \mathbb{R}, w_2 = 0. \end{aligned}$$

According to (5), the solution for InCo ($\lambda = +\infty$) is

$$w_1 \in \mathbb{R}, w_2 = 0.$$

Case 2: η^{e_1} and η^{e_2} are independent but not identically distributed, i.e., $\eta^{e_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\eta^{e_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$, we have (i) $\mathbb{E}[(\boldsymbol{x}_1^{e_i})^2] = \mathbb{E}[(\boldsymbol{x}_2^{e_i})^2] = 1 + \mu_i^2 + \sigma_i^2$, (ii) $\mathbb{E}(\boldsymbol{x}_1^{e_i} y) = a$, $\mathbb{E}(\boldsymbol{x}_2^{e_i} y) = b_i$, (iii) $\mathbb{E}(\boldsymbol{x}_1^{e_i} \boldsymbol{x}_2^{e_i}) = ab_i + \mu_i^2 + \sigma_i^2$, where $a := 1 - 2\alpha$ and $b_i = 1 - 2\beta^{e_i}$ for $i \in \{1, 2\}$.

According to (3), we can calculate the solution for IRMv1 ($\lambda = +\infty$) is

$$w_1 = 0, w_2 = 0.$$

According to (4), we can calculate the solution for VREx ($\lambda = +\infty$) is

$$w_1 = 0, w_2 = 0.$$

According to (5), the solution for InCo ($\lambda = +\infty$) is

$$w_1 \in \mathbb{R}, w_2 = 0.$$

These calculation results are also consistent with the simulations in Sec. 2.3 and App. A.

B.2 More calculation results

When optimizing IGA with $||\nabla_{\boldsymbol{w}} \mathcal{R}^{e_1}(\boldsymbol{w}) - \nabla_{\boldsymbol{w}} \mathcal{R}^{e_2}(\boldsymbol{w})||_2^2 \to 0$, we get

$$\begin{aligned} &\left(\mathbb{E}_{\boldsymbol{x}^{e_1}, y}((\boldsymbol{x}_1^{e_1})^2 w_1 + w_2 \boldsymbol{x}_1^{e_1} \boldsymbol{x}_2^{e_1} - \boldsymbol{x}_1^{e_1} y) - \mathbb{E}_{\boldsymbol{x}^{e_2}, y}((\boldsymbol{x}_1^{e_2})^2 w_1 + w_2 \boldsymbol{x}_1^{e_2} \boldsymbol{x}_2^{e_2} - \boldsymbol{x}_1^{e_2} y)\right)^2 +, \\ &\left(\mathbb{E}_{\boldsymbol{x}^{e_1}, y}((\boldsymbol{x}_2^{e_1})^2 w_2 + w_1 \boldsymbol{x}_1^{e_1} \boldsymbol{x}_2^{e_1} - \boldsymbol{x}_2^{e_1} y) - \mathbb{E}_{\boldsymbol{x}^{e_2}, y}((\boldsymbol{x}_2^{e_2})^2 w_2 + w_1 \boldsymbol{x}_1^{e_2} \boldsymbol{x}_2^{e_2} - \boldsymbol{x}_2^{e_2} y)\right)^2 \to 0; \end{aligned} \tag{6}$$

when optimizing Fishr with $||\mathrm{Var}(\nabla_{\boldsymbol{w}} \mathcal{R}(\boldsymbol{x}^{e_1}, \boldsymbol{w})) - \mathrm{Var}(\nabla_{\boldsymbol{w}} \mathcal{R}(\boldsymbol{x}^{e_2}, \boldsymbol{w}))||_2^2 \to 0$, we have

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}^{e_1}, y}((\boldsymbol{x}_1^{e_1})^2 w_1 + w_2 \boldsymbol{x}_1^{e_1} \boldsymbol{x}_2^{e_1} - \boldsymbol{x}_1^{e_1} y - \mathbb{E}_{\boldsymbol{x}^{e_1}, y}((\boldsymbol{x}_1^{e_1})^2 w_1 + w_2 \boldsymbol{x}_1^{e_1} \boldsymbol{x}_2^{e_1} - \boldsymbol{x}_1^{e_1} y))^2 \\ &- \mathbb{E}_{\boldsymbol{x}^{e_2}, y}((\boldsymbol{x}_1^{e_2})^2 w_1 + w_2 \boldsymbol{x}_1^{e_2} \boldsymbol{x}_2^{e_2} - \boldsymbol{x}_1^{e_2} y - \mathbb{E}_{\boldsymbol{x}^{e_2}, y}((\boldsymbol{x}_1^{e_2})^2 w_1 + w_2 \boldsymbol{x}_1^{e_2} \boldsymbol{x}_2^{e_2} - \boldsymbol{x}_1^{e_2} y))^2 \to 0, \\ &\mathbb{E}_{\boldsymbol{x}^{e_1}, y}((\boldsymbol{x}_2^{e_1})^2 w_2 + w_1 \boldsymbol{x}_1^{e_1} \boldsymbol{x}_2^{e_1} - \boldsymbol{x}_2^{e_1} y - \mathbb{E}_{\boldsymbol{x}^{e_1}, y}((\boldsymbol{x}_2^{e_1})^2 w_2 + w_1 \boldsymbol{x}_1^{e_1} \boldsymbol{x}_2^{e_1} - \boldsymbol{x}_2^{e_1} y))^2 \\ &- \mathbb{E}_{\boldsymbol{x}^{e_2}, y}((\boldsymbol{x}_2^{e_2})^2 w_2 + w_1 \boldsymbol{x}_1^{e_2} \boldsymbol{x}_2^{e_2} - \boldsymbol{x}_2^{e_2} y - \mathbb{E}_{\boldsymbol{x}^{e_2}, y}((\boldsymbol{x}_2^{e_2})^2 w_2 + w_1 \boldsymbol{x}_1^{e_2} \boldsymbol{x}_2^{e_2} - \boldsymbol{x}_2^{e_2} y))^2 \to 0; \end{aligned} \tag{7}$$

when optimizing IB-ERM till $\mathrm{Var}(g(\boldsymbol{x}^{e_1}; \Phi)|y = 1) + \mathrm{Var}(g(\boldsymbol{x}^{e_2}; \Phi)|y = 1) + \mathrm{Var}(g(\boldsymbol{x}^{e_1}; \Phi)|y = -1) + \mathrm{Var}(g(\boldsymbol{x}^{e_2}; \Phi)|y = -1) = 0$, we can get

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}^{e_1}}(w_1 \boldsymbol{x}_1^{e_1} + w_2 \boldsymbol{x}_2^{e_1} - \mathbb{E}_{\boldsymbol{x}^{e_1}}(w_1 \boldsymbol{x}_1^{e_1} + w_2 \boldsymbol{x}_2^{e_1})|y = 1)^2 = 0, \\ &\mathbb{E}_{\boldsymbol{x}^{e_2}}(w_1 \boldsymbol{x}_1^{e_2} + w_2 \boldsymbol{x}_2^{e_2} - \mathbb{E}_{\boldsymbol{x}^{e_2}}(w_1 \boldsymbol{x}_1^{e_2} + w_2 \boldsymbol{x}_2^{e_2})|y = 1)^2 = 0, \\ &\mathbb{E}_{\boldsymbol{x}^{e_1}}(w_1 \boldsymbol{x}_1^{e_1} + w_2 \boldsymbol{x}_2^{e_1} - \mathbb{E}_{\boldsymbol{x}^{e_1}}(w_1 \boldsymbol{x}_1^{e_1} + w_2 \boldsymbol{x}_2^{e_1})|y = -1)^2 = 0, \\ &\mathbb{E}_{\boldsymbol{x}^{e_2}}(w_1 \boldsymbol{x}_1^{e_2} + w_2 \boldsymbol{x}_2^{e_2} - \mathbb{E}_{\boldsymbol{x}^{e_2}}(w_1 \boldsymbol{x}_1^{e_2} + w_2 \boldsymbol{x}_2^{e_2})|y = -1)^2 = 0. \end{aligned} \tag{8}$$

Given **case 2**, the solutions for IGA ($\lambda = +\infty$), Fishr ($\lambda = +\infty$) and IB-ERM ($\lambda = +\infty$) are

$$w_1 = 0, w_2 = 0.$$

These calculation results are also consistent with the simulations in App. A.

C MORE CAUSALITY ANALYSES

Given the theoretical setting in Sec. 3, we have the following corollaries.

Setting: Consider several training environments $\mathcal{E}_{tr} = \{e_1, e_2, ...\}$ and $\boldsymbol{x}^e$ to be the observed input of $e \in \mathcal{E}_{tr}$. We adopt an anti-causal framework (Arjovsky et al., 2019) with data generation process as follows:

$$\begin{aligned} &y = \gamma^\top \hat{\boldsymbol{x}}_{inv} + \eta_y, \\ &\boldsymbol{x}^e_{inv} = \hat{\boldsymbol{x}}_{inv} + \eta^e_{inv}, \quad \boldsymbol{x}^e_s = \hat{\boldsymbol{x}}^e_s + \eta^e_s, \\ &\boldsymbol{x}^e = S\begin{pmatrix} \boldsymbol{x}^e_{inv} \\ \boldsymbol{x}^e_s \end{pmatrix}, \end{aligned}$$

where $\gamma \in \mathbb{R}^{d_{inv}}$ and $\gamma \neq \mathbf{0}$, the hidden invariant feature $\hat{\boldsymbol{x}}_{inv}$ and the observed invariant feature $\boldsymbol{x}^e_{inv}$ take values in $\mathbb{R}^{d_{inv}}$, the hidden spurious feature $\hat{\boldsymbol{x}}^e_s$ and the observed spurious feature $\boldsymbol{x}^e_s$ take values in $\mathbb{R}^{d_s}$, and $S : \mathbb{R}^{(d_{inv}+d_s)} \to \mathbb{R}^d$ is an inherent mapping to mix features. The hidden spurious feature $\hat{\boldsymbol{x}}^e_s$ is generated by y with any *non-invariant* relationship, η^e_{inv} and η^e_s are independent Gaussian with bounded mean and variance changed by environments, η_y is an independent and invariant zero-mean Gaussian with bounded variance. As the directed acyclic graph (DAG) in Fig. 4(b) shows, the hidden invariant feature $\hat{\boldsymbol{x}}_{inv}$ generates the true label y and y generates the hidden spurious feature $\hat{\boldsymbol{x}}^e_s$. In consideration of environmental noise, we can only observe the input $\boldsymbol{x}^e$ which is a mixture of $\boldsymbol{x}^e_{inv}$ and $\boldsymbol{x}^e_s$ after mapping. (Note that the observed feature is generated by applying environmental noise to the hidden feature.) We follow the assumption from IRM Arjovsky et al. (2019), i.e., assume that there exists a mapping $\tilde{S} : \mathbb{R}^d \to \mathbb{R}^{d_{inv}}$ such that $\tilde{S}(S(\begin{smallmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{smallmatrix})) = \mathbf{x}_1$ for all $\mathbf{x}_1 \in \mathbb{R}^{d_{inv}}, \mathbf{x}_2 \in \mathbb{R}^{d_s}$. and aim to learn a classifier to predict y based on $\boldsymbol{x}^e$, i.e., $f(\boldsymbol{x}^e; \boldsymbol{w}) = h(g(\boldsymbol{x}^e; \Phi); \boldsymbol{v})$.

Corollary C.1 *If Φ elicits the desired invariant predictor $f(\cdot; \boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$, there exist noisy environments $\{e_1, e_2\}$ such that*

$$\nabla_{\boldsymbol{w}} \mathcal{R}^{e_1}(\boldsymbol{w}) \neq \nabla_{\boldsymbol{w}} \mathcal{R}^{e_2}(\boldsymbol{w}).$$

Proof C.1 *If Φ elicits the desired invariant predictor $f(\cdot; \boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$ in noisy environments $\{e_1, e_2\}$, given square loss and the fixed "dummy" classifier $\boldsymbol{v} = 1$, we have*

$$\begin{aligned} \frac{\partial \mathcal{R}^e(\boldsymbol{w})}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}} &= \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}^e, y}[(f(\boldsymbol{x}^e; \boldsymbol{w}) - y)^2]}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}} \\ &= \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}^e, y}[(\boldsymbol{v}|_{\boldsymbol{v}=1}(\gamma^\top \hat{\boldsymbol{x}}_{inv} + \gamma^\top \eta^e_{inv}) - \gamma^\top \hat{\boldsymbol{x}}_{inv} - \eta_y)^2]}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}} \\ &= \mathbb{E}_{\boldsymbol{x}^e, y}\Big((\gamma^\top \hat{\boldsymbol{x}}_{inv} + \gamma^\top \eta^e_{inv})(\gamma^\top \eta^e_{inv} - \eta_y)\Big) \\ &= \mathbb{E}_{\boldsymbol{x}^e, y}\Big(\gamma^\top \eta^e_{inv} \gamma^\top \hat{\boldsymbol{x}}_{inv} + (\gamma^\top \eta^e_{inv})^2 - \gamma^\top \hat{\boldsymbol{x}}_{inv} \eta_y - \gamma^\top \eta^e_{inv} \eta_y\Big), \end{aligned} \tag{9}$$

where $e \in \{e_1, e_2\}$.

Obviously, when $\gamma \neq \mathbf{0}$, there exists $\eta^{e_1}_{inv} \neq \eta^{e_2}_{inv}$ such that $\frac{\partial \mathcal{R}^{e_1}(\boldsymbol{w})}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}} \neq \frac{\partial \mathcal{R}^{e_2}(\boldsymbol{w})}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}}$. □

Cor. C.1 shows that $||\nabla_{\boldsymbol{w}} \mathcal{R}^{e_1}(\boldsymbol{w}) - \nabla_{\boldsymbol{w}} \mathcal{R}^{e_2}(\boldsymbol{w})||_2^2 \to 0$ (IGA) may also be failed to find the optimal invariant predictor in noisy environments. Given different inherent losses, it seems unreasonable to enforce all gradients to be equal across environments.

Corollary C.2 *If Φ elicits the desired invariant predictor $f(\cdot; \boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$, there exist noisy environments $\{e_1, e_2\}$ such that*

$$\mathrm{Var}(\nabla_{\boldsymbol{w}} \mathcal{R}(\boldsymbol{x}^{e_1}, \boldsymbol{w})) \neq \mathrm{Var}(\nabla_{\boldsymbol{w}} \mathcal{R}(\boldsymbol{x}^{e_2}, \boldsymbol{w})).$$

Proof C.2 *If Φ elicits the desired invariant predictor $f(\cdot;\boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$ in noisy environments $\{e_1, e_2\}$, given square loss and the fixed "dummy" classifier $\boldsymbol{v} = 1$, we have*

$$\begin{aligned}\mathrm{Var}\left(\frac{\partial \mathcal{R}(\boldsymbol{x}^e, \boldsymbol{w})}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}}\right) =& \mathbb{E}_{\boldsymbol{x}^e,y}\Big(\left[\gamma^\top \eta_{inv}^e \left(\gamma^\top \hat{\boldsymbol{x}}_{inv} + \gamma^\top \eta_{inv}^e - \gamma_y\right)\right]^2 + \left(\gamma^\top \hat{\boldsymbol{x}}_{inv}\eta_y\right)^2 \\ & - 2\gamma^\top \hat{\boldsymbol{x}}_{inv}\eta_y \gamma^\top \eta_{inv}^e \left(\gamma^\top \hat{\boldsymbol{x}}_{inv} + \gamma^\top \eta_{inv}^e - \gamma_y\right)\Big) \\ & - \left[\mathbb{E}_{\boldsymbol{x}^e,y}\Big(\gamma^\top \eta_{inv}^e \gamma^\top \hat{\boldsymbol{x}}_{inv} + (\gamma^\top \eta_{inv}^e)^2 - \gamma^\top \hat{\boldsymbol{x}}_{inv}\eta_y - \gamma^\top \eta_{inv}^e \eta_y\Big)\right]^2,\end{aligned}$$

where $e \in \{e_1, e_2\}$.

Clearly, when $\gamma \neq \mathbf{0}$, there exists $\eta_{inv}^{e_1} \neq \eta_{inv}^{e_2}$ such that $\mathrm{Var}\left(\frac{\partial \mathcal{R}(\boldsymbol{x}^{e_1}, \boldsymbol{w})}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}}\right) \neq \mathrm{Var}\left(\frac{\partial \mathcal{R}(\boldsymbol{x}^{e_2}, \boldsymbol{w})}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}}\right)$. □

Cor. C.2 implies that looking for the optimal invariant predictor in noisy environments via $||\mathrm{Var}(\nabla_{\boldsymbol{w}} \mathcal{R}(\boldsymbol{x}^{e_1}, \boldsymbol{w})) - \mathrm{Var}(\nabla_{\boldsymbol{w}} \mathcal{R}(\boldsymbol{x}^{e_2}, \boldsymbol{w}))||_2^2 \to 0$ (Fishr) may not always be successful, for the reason that environmental inherent noises can affect the variance of gradients.

Corollary C.3 *Given $y \in \{-1, 1\}$ and the fixed "dummy" classifier $\boldsymbol{v} = 1$, if Φ elicits the desired invariant predictor $f(\cdot;\boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$, there exists e in noisy environments such that*

$$\mathrm{Var}(g(\boldsymbol{x}^e; \Phi)|y) \neq 0.$$

Proof C.3 *Given $y \in \{-1, 1\}$ and the fixed "dummy" classifier $\boldsymbol{v} = 1$, if Φ elicits the desired invariant predictor $f(\cdot;\boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$, we have*

$$\mathrm{Var}(g(\boldsymbol{x}^e; \Phi)|y) = \mathrm{Var}\big((\gamma^\top \hat{\boldsymbol{x}}_{inv}|y) + \gamma^\top \eta_{inv}^e\big).$$

Obviously, we can find a η_{inv}^e in noisy environments such that $\mathrm{Var}(g(\boldsymbol{x}^e; \Phi)|y) \neq 0$. □

Cor. C.3 suggests that the IB penalty (IB-ERM) may also be unsuccessful to find the optimal invariant predictor in noisy environments.

D PROOFS

Here, we provide the proofs for Thm. 3.1, Cor. 3.2 and Cor. 3.3, respectively.

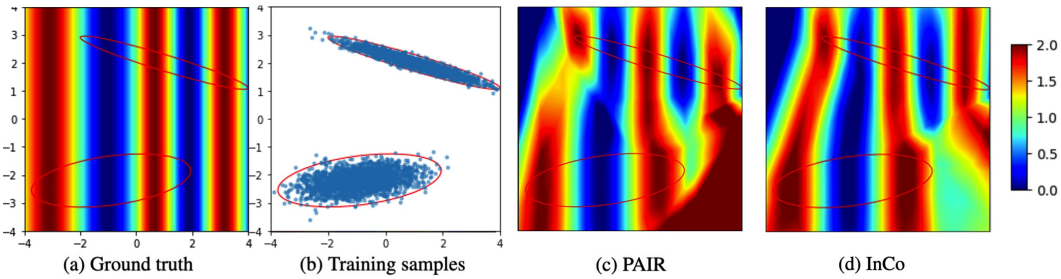
Proof 3.1 *Assume that there exists a mapping $\tilde{S} : \mathbb{R}^d \to \mathbb{R}^{d_{inv}}$ such that $\tilde{S}(S(\begin{smallmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{smallmatrix})) = \mathbf{x}_1$ for all $\mathbf{x}_1 \in \mathbb{R}^{d_{inv}}, \mathbf{x}_2 \in \mathbb{R}^{d_s}$. Then, if Φ elicits the desired (optimal) invariant predictor $f(\cdot;\boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$, we have*

$$\begin{aligned}\rho_{f,y}^e(\boldsymbol{w}) &= \mathbb{E}_{\boldsymbol{x}^e,y}[f(\boldsymbol{x}^e;\boldsymbol{w})y - \mathbb{E}_{\boldsymbol{x}^e}(f(\boldsymbol{x}^e;\boldsymbol{w}))y] \\ &= \mathbb{E}_{\boldsymbol{x}^e,y}[\gamma^\top \tilde{S}(S(\begin{smallmatrix}\boldsymbol{x}_{inv}^e\\ \boldsymbol{x}_s^e\end{smallmatrix}))y] - \mathbb{E}[\gamma^\top \tilde{S}(S(\begin{smallmatrix}\boldsymbol{x}_{inv}^e\\ \boldsymbol{x}_s^e\end{smallmatrix}))]\mathbb{E}[y] \\ &= \mathbb{E}_{\boldsymbol{x}^e,y}[\gamma^\top \boldsymbol{x}_{inv}^e y] - \mathbb{E}[\gamma^\top \boldsymbol{x}_{inv}^e]\mathbb{E}[y] \\ &= \mathbb{E}[(\gamma^\top \hat{\boldsymbol{x}}_{inv})^2] - [\mathbb{E}(\gamma^\top \hat{\boldsymbol{x}}_{inv})]^2 \\ &= \mathrm{Var}(\gamma^\top \hat{\boldsymbol{x}}_{inv}),\end{aligned} \tag{10}$$

for all $e \in \mathcal{E}$. As $\mathrm{Var}(\gamma^\top \hat{\boldsymbol{x}}_{inv})$ remains constant in all environments, we have $\mathrm{Var}(\rho_{f,y}^e(\boldsymbol{w})) = 0$. Hence, proved. □

Proof 3.2 *If Φ elicits the desired invariant predictor $f(\cdot;\boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$, consider square loss and the fixed "dummy" classifier $\boldsymbol{v} = 1$, then*

$$\begin{aligned}\frac{\partial \mathcal{R}^e(\boldsymbol{w})}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}} &= \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}^e,y}[(f(\boldsymbol{x}^e;\boldsymbol{w}) - y)^2]}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}} \\ &= \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}^e,y}[(\boldsymbol{v}|_{\boldsymbol{v}=1}(\gamma^\top \hat{\boldsymbol{x}}_{inv} + \gamma^\top \eta_{inv}^e) - \gamma^\top \hat{\boldsymbol{x}}_{inv} - \eta_y)^2]}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}} \\ &= \mathbb{E}_{\boldsymbol{x}^e,y}\Big((\gamma^\top \hat{\boldsymbol{x}}_{inv} + \gamma^\top \eta_{inv}^e)(\gamma^\top \eta_{inv}^e - \eta_y)\Big) \\ &= \mathbb{E}_{\boldsymbol{x}^e,y}\Big(\gamma^\top \eta_{inv}^e \gamma^\top \hat{\boldsymbol{x}}_{inv} + (\gamma^\top \eta_{inv}^e)^2 - \gamma^\top \hat{\boldsymbol{x}}_{inv}\eta_y - \gamma^\top \eta_{inv}^e \eta_y\Big).\end{aligned} \tag{11}$$



(a) Ground truth (b) Training samples (c) PAIR (d) InCo

Figure 8: Results of causal invariance (Chen et al., 2023) in noisy environments. We run each method with 5 times and report the average losses: **(c)** PAIR 0.8164; **(d)** InCo 0.6568.

Obviously, when $\gamma \neq \mathbf{0}$*, there exists* η^e_{inv} *in noisy environments such that* $\frac{\partial \mathcal{R}^e(\boldsymbol{w})}{\partial \boldsymbol{v}|_{\boldsymbol{v}=1}} \neq \mathbf{0}$. □

Proof 3.3 *If* Φ *elicits the desired invariant predictor* $f(\cdot;\boldsymbol{w}) = \gamma^\top \tilde{S}(\cdot)$*, consider square loss, then*

$$\begin{aligned} \mathcal{R}^e(\boldsymbol{w}) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}^e,y}\big[(\gamma^\top \hat{\boldsymbol{x}}_{inv} + \gamma^\top \eta^e_{inv} - \gamma^\top \hat{\boldsymbol{x}}_{inv} - \eta_y)^2\big] \\ &= \frac{1}{2}\mathbb{E}\big[(\gamma^\top \eta^e_{inv} - \eta_y)^2\big], \end{aligned} \tag{12}$$

where $e \in \{e_1, e_2\}$.

Clearly, when $\gamma \neq \mathbf{0}$*, there exists* $\eta^{e_1}_{inv} \neq \eta^{e_2}_{inv}$ *such that* $\mathcal{R}^{e_1}(\boldsymbol{w}) \neq \mathcal{R}^{e_2}(\boldsymbol{w})$. □

E More experiments and details

Experimental details: All experiments are implemented on NVIDIA A100 and AMD EPYC 7452 32-Core Processor.

For the experiment with ColoredMNIST, we use the exactly same setting as `https://github.com/capybaralet/REx_code_release/blob/master/InvariantRiskMinimization/colored_mnist/main.py`, only replacing the IRMv1 penalty and VREx penalty with InCo penalty and other penalties.

For the experiment with Circle Dataset, we use the exactly same setting as `https://github.com/hehaodele/CIDA/blob/master/toy-circle/main-half-circle.ipynb`, only applying noises to source domains and adding InCo penalty term.

Causal Invariance experiment: We then describe the definition of Causal Invariance specified by Peters et al. (2016); Arjovsky et al. (2019); Kamath et al. (2021); Chen et al. (2023) as in Def. E.1.

Definition E.1 *(Causal Invariance) Given a predictor* $f(\cdot;\boldsymbol{w}) = h(g(\cdot;\Phi);\boldsymbol{v})$*, the representation produced by the featurizer* Φ *is invariant over* $\mathcal{E}$ *if and only if for all* $e_1, e_2 \in \mathcal{E}$*, it holds that*

$$\mathbb{E}_{\boldsymbol{x}^{e_1},y}(y|g(\boldsymbol{x}^{e_1};\Phi) = \boldsymbol{z}) = \mathbb{E}_{\boldsymbol{x}^{e_2},y}(y|g(\boldsymbol{x}^{e_2};\Phi) = \boldsymbol{z}) \tag{13}$$

for all $\boldsymbol{z} \in \{g(\boldsymbol{x}^{e_1};\Phi)|e_1\} \cap \{g(\boldsymbol{x}^{e_2};\Phi)|e_2\}$.

As Chen et al. (2023), a regression example is designed with $\boldsymbol{x} : \mathbb{R}^2 \to y : \mathbb{R}$. The input $\boldsymbol{x}$ is with two dimensions, i.e., $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, where $\boldsymbol{x}_1$ represents horizontal axis and $\boldsymbol{x}_2$ represents vertical axis in Fig. 8. $\boldsymbol{x}_1$ is designed to be the invariant feature and $\boldsymbol{x}_2$ is designed to be the spurious feature. Consider environmental inherent noises, we assume $y = sin(1.5 * \boldsymbol{x}_1) + 1$ for domains $\boldsymbol{x}_1 < 0$ and $y = sin(2.5 * \boldsymbol{x}_1) + 1$ for domains $\boldsymbol{x}_1 \geq 0$. All other settings are same with `https://github.com/LFhase/PAIR/blob/main/Extrapolation/pair_extrapolation.ipynb`.

We evaluate InCo with Causal Invariance experiment from PAIR (Chen et al., 2023). As shown in Fig. 8, $y = sin(1.5 \cdot \boldsymbol{x}_1) + 1$ for $\boldsymbol{x}_1 < 0$ and $y = sin(2.5 \cdot \boldsymbol{x}_1) + 1$ for $\boldsymbol{x}_1 \geq 0$. y is solely determined by $\boldsymbol{x}_1$ (horizontal axis), while $\boldsymbol{x}_2$ (vertical axis) does not influence the values of y. Different colors represent different values of y. Note that we assume environmental noises influence domains $\boldsymbol{x}_1 < 0$ with $sin(1.5 \cdot \boldsymbol{x}_1)$ and domains $\boldsymbol{x}_1 \geq 0$ with $sin(2.5 \cdot \boldsymbol{x}_1)$. We sample two training areas as denoted

Table 6: Comparison of MLP on ColoredMNIST with varying training noises, i.e., first training environment without noise, second training environment with Poisson noise (with coefficient 0.1) and Uniform noise ($[-0.1, 0.1]$), respectively. We repeat each experiment with 20 times and report the best, worst and average accuracies (%) on the test environment with Poisson noise and Uniform noise, respectively.

Test noise	Method	$\{\mathbf{0}, \text{Poisson}\}_{\text{train}}$			$\{\mathbf{0}, \text{Uniform}\}_{\text{train}}$		
		Best	Worst	Mean	Best	Worst	Mean
Poisson	ERM	49.55	10.13	26.58	49.64	9.73	20.49
	IRMv1	48.79	9.41	26.19	50.04	9.95	31.69
	VREx	55.49	40.60	46.17	56.62	38.80	44.77
	CLOvE	50.11	10.65	29.48	49.87	9.52	32.66
	Fishr	53.72	41.25	45.82	54.59	40.76	44.21
	InCo	**60.95**	**44.18**	**53.31**	**59.13**	**47.73**	**53.01**
Uniform	ERM	50.33	9.77	26.12	49.66	9.47	22.10
	IRMv1	49.91	9.46	26.23	49.48	9.69	29.17
	VREx	57.21	40.12	46.36	58.69	38.93	45.41
	CLOvE	51.80	10.33	30.18	50.12	9.70	32.47
	Fishr	53.98	40.93	46.79	54.66	41.27	46.10
	InCo	**62.41**	**44.61**	**53.97**	**60.88**	**48.11**	**53.58**

by the ellipsoids colored in red (Fig. 8(b)). With 5 repeats, InCo achieves the lower average loss (0.6568) than PAIR (0.8164).

Experiments with other noises: As shown in Tab. 6, InCo also gets a better performance in Poisson noisy and Uniform noisy environments.

F More related work

Domain Generalization

Domain generalization can also be improved by model averaging (Cha et al., 2021; Arpit et al., 2022), training a model guided by meta learning (Robey et al., 2021; Li et al., 2018a; 2019a;b; Balaji et al., 2018), sample selection (Kahng et al., 2023), balanced mini-batch sampling (Wang et al., 2023a), and indirection representations (Pham et al., 2023). Additionally, by training the model on a variety of produced novel domains, data augmentation-based approaches can also increase the generalization ability, e.g. using domain synthesis to create new domains (Zhou et al., 2020). Some works also utilized the robust gradient direction to perturb data and obtained a new dataset to train the model (Shankar et al., 2018; Wang et al., 2020b; 2023b). Volpi et al. (2018) and Carlucci et al. (2019) construct a new dataset by solving the jigsaw puzzle. Lee et al. (2023) improves domain generalization through finding a diverse set of hypotheses and choosing the best one. Kaur et al. (2023) develops the technique of causally adaptive constraint minimization to improve domain generalization. Huang et al. (2023) proposes HOOD method that can leverage the content and style from each image instance to identify benign and malign (out of distribution) data. Xu et al. (2021) develops a novel Fourier-based data augmentation strategy, which linearly interpolates between the amplitude spectrums of two images, to improve domain generalization.

IRM

In addition, IRM is also widely studied. Choe et al. (2020) take an empirical study of IRMv1 across various environments by examining the performance of IRMv1 in different frameworks including text classification models and then Sonar et al. (2021) extends the IRM to the reinforcement learning task. Mitrovic et al. (2020) proposed a self-supervised setup method to learn the optimal representation by augmenting the data to build the second domain. Sun et al. (2023) studies the generalization issue of face anti-spoofing models through IRM. Shao et al. (2022) shows that active model adaptation could achieve both good performance and robustness based on the IRM principle. Wad et al. (2022) proposes a class-wise IRM method that tackles the challenge of missing environmental annotation. Lin et al. (2022) introduces Bayesian inference into IRM to its performance on DNNs. Yu et al. (2022) proposes a Lipschitz regularized IRM-related method to alleviate the influence of low quality data at both the sample level and the domain level. Lu et al. (2021) studies IRM and obtains generalization guarantees in the nonlinear setting.