When Does Group Invariant Learning Survive Spurious Correlations?

Yimeng Chen[1,2], **Ruibin Xiong**[3], **Zhiming Ma**[1,2], **Yanyan Lan**[4*]
[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]Baidu Inc.
[4]Institute for AI Industry Research, Tsinghua University
chenyimeng14@mails.ucas.ac.cn,xiongruibin@baidu.com,
mazm@amt.ac.cn,lanyanyan@tsinghua.edu.cn

Abstract

By inferring latent groups in the training data, recent works introduce invariant learning to the case where environment annotations are unavailable. Typically, learning group invariance under a majority/minority split is empirically shown to be effective in improving out-of-distribution generalization on many datasets. However, theoretical guarantee for these methods on learning invariant mechanisms is lacking. In this paper, we reveal the insufficiency of existing group invariant learning methods in preventing classifiers from depending on spurious correlations in the training set. Specifically, we propose two criteria on judging such sufficiency. Theoretically and empirically, we show that existing methods can violate both criteria and thus fail in generalizing to spurious correlation shifts. Motivated by this, we design a new group invariant learning method, which constructs groups with statistical independence tests, and reweights samples by group label proportion to meet the criteria. Experiments on both synthetic and real data demonstrate that the new method significantly outperforms existing group invariant learning methods in generalizing to spurious correlation shifts.

1 Introduction

In many real-world applications, machine learning models inevitably encounter data that are rarely presented in the training environment, i.e. being *out-of-distribution* (OOD). For example, data collected under new weather [38], locations [6], or light conditions [9] in vision tasks. However, machine learning models often fail in generalizing to OOD data, which blocks their deployment to critical applications [12; 40]. The dependence on spurious correlations that are prone to change across environments has been recognized as a major cause of such failure [4; 12; 39]. For example, it has been shown that models trained on MNLI [41] usually classify sentence pairs with high word overlap as the label 'entailment', regardless of their semantics [27]. On a new dataset where such relation no longer holds, the performance drops over 25% [27; 7].

A notable line of research on improving the robustness of models to distribution shifts is learning features with invariant conditioned label distribution across training environments [30; 4; 15], which has been termed *invariant learning* (IL). These methods are based on the assumption that the causal mechanism keeps invariant across environments [30], while the spurious correlation varies. By penalizing the variance of model prediction across environments, models are then encouraged to capture the causal mechanism instead of spurious correlations.

*Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Recently, invariant learning has been introduced to the scenario where environment labels are unknown [35; 8], which we term the *group invariant learning* (group-IL). These methods utilize prior knowledge of spurious correlations to split the training data into groups. For example, Teney et al. [35] cluster training samples with their predefined spurious features. A more generic method, EIIL [8], splits training data into the majority/minority sets on which the spurious feature conditioned label distribution varies maximally. Similar to a priori environments, these groups are supposed to encode variations of spurious information, while holding the causal mechanism.

Though some performance improvements have been gained on several datasets, much uncertainty still exists on the effectiveness of group-IL methods. In particular, *when these methods can effectively address spurious correlations* remain a question. Though some theoretical analysis on the success and failure cases of IL with known environments has been proposed [2; 33; 23; 3], they are not sufficient for group-IL. First, inferred groups may not meet the assumptions on environments in existing theoretical analysis, thus their conclusions cannot generalize to group-IL. For example, in each environment, the spurious feature is assumed to have a Gaussian type distribution in [33]. However, the inferred group may not satisfy that condition, for example each group may only contain one unique value of spurious feature. Second, as environments are known and defined with causal structures, exiting analysis on the effect of environment on IL focus mostly on their number [33; 3] but less on their property or validity. However, the later are important for group-IL. Therefore, we need specific theoretical analysis under the setting of group-IL.

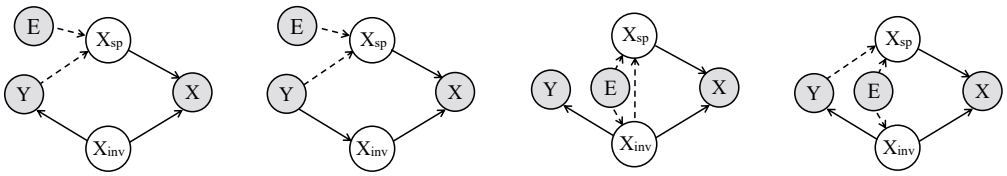
In this study, we discuss necessary conditions for group-IL to survive spurious correlations. For this purpose, we first formalize the setting of group-IL and clarify the necessary assumptions required for group-IL, which underlie our theoretical analysis. We then propose two criteria for group-IL, namely *falsity exposure criterion* and *label balance criterion*. They are respectively for judging whether spurious correlations are sufficiently exposed through their variation across groups and whether group invariance can reach spurious-free conditions. Based on that, we discuss the success and failure cases of existing methods. In some synthetic benchmarks (e.g. colored-MNIST in [4; 8]), the majority/minority groups meet the two criteria. However, in a case when the spurious feature is multivariate, the majority/minority split violates both criteria according to our theoretical analysis and observations on real datasets. As a result, existing group-IL methods are insufficient for solving spurious correlations.

To fix these problems, we propose a new group-IL method guided by the two criteria. Specifically, this method contains the following two steps to meet the two criteria. First, groups are defined by stratifying the prediction of a reference predictor which encodes spurious correlations. The strata is constructed with statistical tests, such that the spurious prediction is independent of the label on each group. Second, the label proportion of each inferred group is balanced by attaching weights to each instance within the group. Models are then trained with invariant learning objectives on the defined groups. We term this method Spurious-Correlation-Strata Invariant Learning with Label-balancing, abbreviated as SCILL. We further show that SCILL is provably sufficient in reaching spurious-free with ideal reference models.

To demonstrate the effectiveness of our proposed strategy, we conduct experiments on both synthetic and real data benchmarks on spurious correlations shifts in image classification and natural language inference (NLI). Specifically, we adopt two different invariant learning objectives, IRM (IRMv1) [4] and REx (V-REx) [15], to show the consistency of SCILL. To show the availability of SCILL, we also experiment with PGI [1] and cMMD [17; 1], which are feature invariance targets used with EIIL [8] in [1]. The experimental results show that SCILL with all the four invariance objectives consistently outperforms the existing state-of-the-art method EIIL in generalizing to spurious correlation shifts. Ablation study further shows the effectiveness of each component in SCILL.

Our main contributions can be summarized as follows.

- We propose two criteria for group-IL, and theoretically show the insufficiency of existing methods.
- Guided by the two criteria, we propose a new practical group-IL method which is provably sufficient in solving spurious correlations.
- Extensive experiments on both synthetic and real-data benchmarks show that SCILL significantly outperforms existing methods on image classification and NLI tasks.



(a) anti-causal [4] (b) anti-causal [33] (c) confounded [2; 25] (d) hybrid [39; 23]

Figure 1: The causal graph depicting different assumptions on the data generating process in existing IL works (some are simplified). Shading indicates the variable is observed. Dotted arrow indicates possible causal relation. The spurious feature is anti-causal in (a) and (b), confounded with the invariant feature in (c), and both anti-causal and confounded in (d).

2 Related works

Combating spurious correlations. A typical kind of distribution-shift is caused by the shift of spurious correlations [40], which are correlations between meaningless features (e.g. hospital tokens on a lung scan) and labels (e.g. has pneumonia or not) in the training set. The existence of spurious correlations in popular benchmarks have been revealed by many works [13; 31; 27; 34]. For example, predictive models with only incomplete semantic inputs or syntactic statistics can achieve high accuracy in NLI benchmarks [13; 31]. Geirhos et al. [11] point out that deep neural networks are prone to take easy-to-fit spurious correlations, i.e. *shortcut* strategies, in solving problems. As a result, resolving model's dependence on spurious correlation is important for their robustness to distribution shifts.

Invariant learning (IL). Many works on domain generalization focus on capturing invariances across training environments [12]. Recently a new kind of strategy which has made significant impacts is to learn features that permit an invariant predictor across environments [30; 4; 15; 39], termed invariant learning in this paper. Such strategy is grounded upon the theory of causality [29], where Structural Equation Models [30] or causal graphs [39] are used to describe assumptions on the data generation process. The feature-conditioned label distribution invariance is then induced by the invariance of causal mechanisms in different environments.

Group invariant learning. In recent works, IL is extended to the scenario without a priori environment labels, but with knowledge on spurious correlations in the training data [35; 8]. Such knowledge is proven to be necessary in this setting [18]. They are utilized to split the training data into groups, which are supposed to encode variations of spurious information so that they can be avoided by learning the invariance. For example, Teney et al. [35] cluster samples according to their question types. Liu et al. [20] construct groups with varying spurious correlations, based on the spurious features uncovered with feature selection. A more generic method proposed in EIIL [8] assume the access to a reference predictor which encodes spurious correlations, and exploit the outputs of the reference predictor to split training data into two groups, namely the majority and the minority. This strategy is shown effective, sometimes even outperform an oracle method using true environment labels in proving the OOD generalization [8] and also systematic generalization [1].

3 Formalization of group invariant learning

In this section, we introduce the problem setting and formalization of the scheme of group-IL.

Consider the task of learning a classifier $f : \mathcal{X} \to \mathcal{Y}$, which maps a value $x \in \mathcal{X}$ of the input variable X to a value $y \in \mathcal{Y}$ of the target variable Y. For example, map an image of horse on grass to the label 'horse'. We denote x_{inv} as the essential features of an instance x of the input variable X which define its class label (e.g. the shape of the horse), while x_{sp} as features of x should not inform the label of x (e.g. the grass background). X_{inv}, X_{sp} denote the corresponding random variables. The target is then to learn a *spurious-free* predictor whose predictions only depend on the feature x_{inv}, thus is supposed to have invariant performance on any dataset.

We first introduce the setting of IL. IL methods suppose that the training data $\mathcal{D}$ are collected under multiple environments $\mathcal{E}$, i.e. $\mathcal{D} = \{D_e\}_{e\in\mathcal{E}}$. $D_e = \{x_i^e, y_i^e\}_{i=1}^{n^e}$ contains data *i.i.d.* sampled from the probability distribution $\mathbb{P}^e(\mathcal{X}\times\mathcal{Y})$. $\mathbb{P}^e(Y|X_{inv})$ is invariant among different e, while $\mathbb{P}^e(Y|X_{sp})$ varies. Such an idea is based on the invariance of causal mechanisms across environments [30; 4], with assumptions on the causal structure of the data generating process. Figure 1 demonstrates four kinds of different assumptions in existing works on invariant learning. In these causal structures, environment is treated as a random variable E take values in $\mathcal{E}_{all}$, which satisfies $\mathcal{E} \subset \mathcal{E}_{all}$. $\mathbb{P}^e(\cdot) := \mathbb{P}(\cdot|E=e)$. X_{inv} and X_{sp} are latent feature variables generating the observation X, i.e. $X = r(X_{inv}, X_{sp})$. Here r is generally assumed as a bijective function so that the latent features can be recovered from the observations [33; 2; 39]. Align with the formalization in the former paragraph where X_{sp}, X_{inv} are assumed recognizable from X, in this paper we adopt the same assumption. In all the four kinds of causal graphs, $\mathbb{P}^e(Y|X_{inv}) := \mathbb{P}(Y|X_{inv}, E=e)$ keeps invariant under different $e \in \mathcal{E}_{all}$, while $\mathbb{P}^e(Y|X_{sp})$, $\mathbb{P}^e(X_{sp})$, $\mathbb{P}^e(X_{inv})$, and $\mathbb{P}^e(X_{inv}|Y)$ can vary across different $e \in \mathcal{E}_{all}$.

Suppose the predictor f can be decomposed into $f = c \circ \Phi$, where $\Phi : \mathcal{X} \to \mathcal{H}$ denotes a feature encoder which maps the input into a representation space $\mathcal{H}$, $c : \mathcal{H} \to \mathcal{Y}$ is a classifier. The target of invariant learning is then to search for a Φ which satisfies the following constraint:

$$\mathbb{P}(Y|\Phi(X), E=e) = \mathbb{P}(Y|\Phi(X), E=e'), \forall e, e' \in \mathcal{E}. \tag{EIC}$$

It is termed as *Environment Invariance Constraint* (EIC). Note that the EIC stated in [8] is a weaker form of this EIC. The constraint is incorporated into the training target via a penalty term. In a generic form, the learning target of invariant learning methods can be written as follows:

$$\min_f \sum_{e\in\mathcal{E}} \lambda_e \mathcal{R}^e(f) + \lambda \cdot \textit{penalty}(\{S_e(f)\}_{e\in\mathcal{E}}) \tag{1}$$

where $\mathcal{R}^e(f)$ stands for the expected loss of f on the environment e, weighted by a scalar λ_e. $S_e(f)$ stands for some statistics of f on e, and the *penalty* is on the variation of $S_e(f)$ to measure the deviation degree of EIC. In IRM [4], the penalty is the summation of $S_e(f) = \|\nabla_w \mathcal{R}^e(w \circ f)\|^2$, where w is a constant scalar multiplier of 1.0 for each output dimension. In V-REx [15], $S_e(f) = \mathcal{R}^e(f)$, and the penalty is the variance of $S_e(f)$ on different environments. In CLOvE [39], the penalty is defined as the summation of calibration errors of the model on each environment.

Group invariant learning methods release the dependency of invariant learning on predefined environments by splitting the training data into groups. Suppose $\mathcal{D}$ is sampled from the distribution $\mathbb{P}(\mathcal{X}\times\mathcal{Y})$. Intuitively, those groups are expected to hold the invariant mechanism $\mathbb{P}(Y|X_{inv})$, while informing the variation of X_{sp}. As a result, it is meaningless to divide groups according to X_{inv}. In group-IL methods [35; 20; 8], group inference algorithms are designed to utilize knowledge on X_{sp} or the correlation between X_{sp} and Y. Formally, denote the inferred groups as $\mathcal{G} := \{g_1, g_2, \ldots, g_m\}$. Define $G : \mathcal{X}\times\mathcal{Y} \to \mathcal{I}$ as the function which maps a sample to its group identity, i.e. $G(x,y) = i$ if and only if $(x,y) \in g_i$. $\mathcal{G}$ is then the set of events $\{G=i\}, i \in \mathcal{I}$. We have G is $\sigma(X_{sp}, Y)$-measurable, i.e. it is a function of X_{sp}, as in [35; 20], or both X_{sp} and Y [8].

In the following sections, we ground our analysis on group-IL with the causal structures in Figure 1 (a) and (b), while without additional assumptions on the causal models. Our choice of causal structures is based on the following two observations. First, in the causal graph (d), X_{sp} is a backdoor variable [29] between X_{inv} and Y and confounded with X_{inv} by an unobserved variable. As a result, whether the invariant mechanism holds on each group is indeterminate without additional assumptions on the mechanisms between X_{sp} and X_{inv}. Second, as shown by Ahuja et al. [2], invariance itself can not deal with spurious feature for causal structure (c). Additional knowledge or penalty, e.g. information bottleneck, is needed together with group-IL. Therefore, we investigate group-IL under the causal structures depicted by graphs (a) and (b).

4 Two group criteria

With the above formulation, we are ready to theoretically analyze the ability of existing group-IL methods in learning spurious-free predictors. For this purpose, in this section we derive two necessary conditions for group-IL in surviving spurious correlations, i.e. falsity exposure, and label balance. Both conditions can be used as criteria to judge the sufficiency of group-IL methods. We then show that existing methods can violate the two criteria, thus become insufficient for learning a spurious-free predictor.

4.1 Falsity exposure criterion

As groups are supposed to expose variance of spurious features so that they can be avoided by invariant learning, a natural idea is to take into account the sufficiency of such exposure on inferred groups. Ideally, if groups are split according to X_{sp}, any variance of $\mathbb{P}(Y|X_{sp})$ is then fully exposed. However, such split is only practical when the value of X_{sp} is accessible and sparse. On the contrary, we consider the condition when groups provide insufficient exposure. Intuitively, if some spurious correlation keeps invariant across groups, its variation is then not exposed, thus group invariant predictor may still depend on such correlation. Formally, this can be written as the following criterion.

Criterion 4.1 (Falsity Exposure)**.** For any $\sigma(X_{sp})$-measurable function h that satisfies $\forall g, g' \in \mathcal{G}$, $\mathbb{P}(Y|h(X_{sp}), g) = \mathbb{P}(Y|h(X_{sp}), g')$, it must satisfies $\mathbb{P}(Y|h(X_{sp})) = \mathbb{P}(Y)$.

Intuitively, if the falsity exposure criterion is not satisfied, $h(X_{sp})$ will be an invariant feature across environments, with predictive ability on Y. The predictor depending on both X_{inv} and $h(X_{sp})$ can satisfy EIC but fails to be free of spurious features. The following theorem formalizes the significance of the falsity exposure criterion.

Theorem 4.2. *Suppose the falsity exposure criterion is violated, i.e.* $\exists h$ *which satisfies* $\mathbb{P}(Y|h(X_{sp}), g) = \mathbb{P}(Y|h(X_{sp}), g') \neq \mathbb{P}(Y), \forall g, g' \in \mathcal{G}$. *Then the optimal solution of group-IL is* $f(X) = \mathbb{P}[Y|X_{inv}, h(X_{sp})]$, *which fails to generalize when* $\mathbb{P}(Y|X_{sp})$ *shifts.*

4.2 Label balance criterion

Even if we have sufficient falsity exposure, would the constraint in group-IL, i.e. EIC, guarantee the model to be free of spurious correlations? We study this problem by analyzing the relation between EIC and the constraint for a predictor to be spurious-free. In our assumed causal structures, $X_{inv} \perp\!\!\!\perp X_{sp}|Y$. Thus, a spurious-free predictor, which only depends on X_{inv}, satisfies $f(X) \perp\!\!\!\perp X_{sp}|Y$. In fact, it can be proved that this is a sufficient condition for a predictor $f(X)$ to be invariant to the intervention [29] on X_{sp} (See the appendix). As a result, we term $f(X) \perp\!\!\!\perp X_{sp}|Y$ as the *spurious-free constraint*, i.e.

$$\mathbb{P}(f(X)|X_{sp} = b, Y) = \mathbb{P}(f(X)|X_{sp} = b', Y), \forall b, b' \in \mathcal{B}, \tag{SFC}$$

where $\mathcal{B}$ is the image set of X_{sp}. The following is a necessary condition for EIC to induce the above constraint, which we term as the *label balance* criterion. It states that the label proportion in different groups should be the same.

Criterion 4.3 (Label Balance)**.** For any $g, g' \in \mathcal{G}$ and $y, y' \in \mathcal{Y}$ with non-zero $\mathbb{P}(Y = y|g), \mathbb{P}(Y = y'|g), \mathbb{P}(Y = y|g')$ and $\mathbb{P}(Y = y'|g')$, the following equation holds.

$$\mathbb{P}(Y = y|g)/\mathbb{P}(Y = y'|g) = \mathbb{P}(Y = y|g')/\mathbb{P}(Y = y'|g') \tag{2}$$

Formally, the following theorem shows the significance of this criterion on the effectiveness of group-IL.

Theorem 4.4. *With a set of groups* $\mathcal{G}$ *inferred by* (X_{sp}, Y), *i.e.* $\mathcal{G} \subset \sigma(X_{sp}, Y)$, *if the label balance criterion is violated, functions satisfying EIC can not satisfy SFC.*

4.3 Analysis of existing group invariant learning methods

Now we analyze whether the groups in existing group-IL methods meet the two criteria. Directly, randomly grouped clusters of X_{sp} as in [35] do not guarantee to meet either criteria, and clusters of $\mathbb{P}(Y|X_{sp})$ [20] do not meet the label balance criterion. In the following discussions, we focus on the majority/minority groups inferred by the EI algorithm in EIIL [8]. We find that in some synthetic datasets where the spurious feature only has two distinct values, the majority/minority groups satisfy the two criteria. However, observation on real datasets and theoretical results on the case when the spurious feature is multivariate show that they can violate both criteria.

On some synthetic datasets constructed in existing works [1; 8], the majority/minority groups satisfy these two criteria. For example, on both colored-MNIST [4] and coloured-MNIST [1], Y has a uniform distribution, and the spurious correlation has the same ratio for any spurious features, e.g.

$\mathbb{P}(Y=0|\text{color}=\textit{green}) = \mathbb{P}(Y=1|\text{color}=\textit{red})$ on colored-MNIST. It can be proved that in this case, the majority/minority groups satisfy both criteria (See the appendix).

However, it no longer holds in general cases. Empirically, we observe that on MNLI, the label distributions of the two groups inferred by EI are significantly different. Specifically, the ratio of the counts of label 0 and label 1 in the two groups are 2.87 and 0.17 respectively. The following proposition theoretically provides a case where the majority/minority split satisfies the label balance but breaks the falsity exposure, and invariant learning objectives fail to find the spurious-free classifier.

Proposition 4.5. *Suppose we have $(X, Y) \sim \mathbb{P}(X, Y)$. Y takes value in $\{0, 1\}$. X is formed with spurious feature variable $X_{sp} = (B_0, B_1)$, and invariant feature variable S, i.e. $X = r(B_0, B_1, S)$, for some bijective function r. B_0 and B_1 are both binary variables, which take values in $\{b_0^0, b_0^1\}$ and $\{b_1^0, b_1^1\}$ respectively. B_0, B_1 and S are conditionally independent given Y. Suppose $\mathbb{P}(Y = j|B_i = b_i^j) = p_i, \forall i, j \in \{0, 1\}$, and $p_0 > p_1$. Then we have 1) the majority/minority groups e_{mar}, e_{min} violate the falsity exposure criterion. 2) the optimal classifier under invariant learning objectives depends on B_1.*

In this case, we suppose the spurious feature can be decomposed into two variables that are conditionally independent with each other given the label. Such case can realize when the dataset contains multiple kinds of spurious features. For example, in the image classification task, the background pattern and the color of an object can be independent but both correlate with the label spuriously.

5 SCILL: a new method

According to the above discussion in Section 4, existing group-IL methods may fail to meet the two criteria, leading to insufficient training loss for solving spurious correlations. Faced with this challenge, we propose a new group-IL method, to satisfy the two criteria. For the generality, we only assume the access to a reference model, as in EIIL, instead of spurious features. Our new method includes two steps. In the group inference step, we define groups as spurious correlation strata constructed with the reference model, for the falsity exposure criterion. While in the training objective step, we reweight each sample with group label proportion to meet the label balance criterion. To highlight the two parts, our method is named as Spurious-Correlation-strata Invariant Learning with Label-balance, abbreviated as SCILL. Theoretically, we prove that this method is provably sufficient with ideal reference models, i.e. its optimal solution is spurious-free.

5.1 Group inference with statistical split

We first introduce the group inference step in SCILL. As in EIIL, we assume a reference classifier f_r is available, which is expected to predict only depending on X_{sp}. f_r can be a model trained with empirical risk minimization (ERM) [8] or a model designed to capture spurious correlations [21].

We propose to construct groups $\mathcal{G}$ by stratify the outputs of f_r, with the target that $Y \perp\!\!\!\perp f_r(X)|g$, $\forall g \in \mathcal{G}$. The motivation for introducing spurious correlation strata comes from the deduction of the falsity exposure criterion, i.e. when groups meet the requirements that $X_{sp} \perp\!\!\!\perp Y|g$, the falsity exposure criterion is satisfied. That is because with the above condition, we have for any function h, $h(X_{sp}) \perp\!\!\!\perp Y|g$, as a result $\mathbb{P}(Y|h(X_{sp}), g) = \mathbb{P}(Y|g)$. If $\mathbb{P}(Y|h(X_{sp}), g) = \mathbb{P}(Y|h(X_{sp}), g')$, $\forall g, g' \in \mathcal{G}$, we have $\mathbb{P}(Y|g) = \mathbb{P}(Y|g') = \mathbb{P}(Y)$. Thus $\mathbb{P}(Y|h(X_{sp}), g) = \mathbb{P}(Y)$. As a result, the falsity exposure criterion is satisfied. As different output values of the reference model f_r inform the difference in $\mathbb{P}(Y|X_{sp})$, $Y \perp\!\!\!\perp f_r(X)|g$ approximates $\mathbb{P}(Y|X_{sp}) \perp\!\!\!\perp Y|g$, which is equivalent to $X_{sp} \perp\!\!\!\perp Y|g$.

We propose to construct such groups through the *statistical-split* algorithm, inspired by the algorithm proposed in [14] for propensity score estimation. Specifically, we split current groups into subgroups according to the hypothesis-test statistics. For example, for the binary classification case, a sample set B is first divided into two subsets L_0, L_1 according to the labels of samples. Then the two-sample t-statistics t_B of $\log[f_r(x)_0/f_r(x)_1]$ are computed on the two sets. If t_B exceeds a fixed threshold thr, B is then split into two subsets according to the median of $f_r(x)_0$ on B. In this way, we enhance $f_r(X) \perp\!\!\!\perp Y$ in each group. Note that thr is a hyperparameter of this algorithm. Empirical robustness analysis on thr is conducted in our experiments. More details about the algorithm and the robustness study are provided in the appendix.

5.2 Training with reweighted loss

Now we introduce the second step. To guarantee the label balance criterion, we reweight each sample with group label proportion in the training loss. Correspondingly, the objective of invariant learning becomes the following form:

$$\mathcal{L}(f) := \sum_{g \in \mathcal{G}} \tilde{\mathcal{R}}^g(f) + \lambda \cdot \textit{penalty}(\{S_g(f)\}_{g \in \mathcal{G}}) \tag{3}$$

where $\tilde{\mathcal{R}}^g(f) = \mathbb{E}[w^g(Y)\mathcal{L}^g(f(X), Y)], \omega^g(y) := \mathbb{P}(Y = y)/\mathbb{P}(Y = y|g)$ for nonzero $\mathbb{P}(Y = y|g)$. The weight function is defined to balance the label distribution between groups, i.e. $\mathbb{P}(Y|g) = \mathbb{P}(Y|g'), \forall g, g'$, for achieving the label balance criterion.

With the above two steps, SCILL is then able to meet the two group criteria. Now we further investigate the theoretical capability of SCILL in solving spurious correlations. The following theorem shows that with a purely spurious reference model, SCILL can find spurious-free predictors.

Theorem 5.1. *If $\mathcal{G}$ satisfies $f_r^*(X) \perp\!\!\!\perp Y|g, \forall g \in \mathcal{G}$, where $f_r^* : \mathcal{X} \to \mathcal{Y}$ is spurious-only, i.e. $\sigma(X_{sp})$-measurable, and minimizes the prediction loss $\mathcal{L}_{ce}^r = \mathbb{E}[\sum_y \mathbb{P}(Y = y|X) \log f_r(X)_y]$, the optimal model minimizing the objective (3) satisfies SFC.*

6 Experiments

In this section, we first conduct experiments to show that SCILL outperforms existing group-IL methods in generalizing to spurious correlation shifts. Then we empirically analyze whether the experimental improvements are consistent with the theoretical findings. Code is availiable at `https://github.com/Beastlyprime/group-invariant-learning`.

6.1 Experimental settings

Now we describe our experimental settings, including datasets, models, and some training details. More details are provided in the appendix.

Datasets We conduct experiments on both synthetic and real-world datasets. The synthetic dataset, Patched-Colored MNIST (PC-MNIST), is constructed as a realization of the conditions in the Proposition 4.5 to verify the proposed criteria. It is derived from MNIST, by assigning two conditionally independent spurious features given label, namely the color and patch bias to each image. The design of the patch bias is inspired by [5]. MNLI-HANS is a benchmark widely used in many previous works on combating spurious correlations, such as [7; 36]. In our experiments, we follow the practice to utilize MNLI [41] as the training data and HANS [27] as the test data.

Baselines and configurations In our experiments, we compare SCILL with two baselines, i.e. ERM and EIIL [8]. ERM represents the method with the traditional empirical risk minimization (ERM) approach. EIIL is a state-of-the-art group-IL method, where groups are inferred by searching an assignment to make the reference model maximally violates the invariant learning principle. We experiment with four different invariance penalties: IRM [4], REx (V-REx) [15], cMMD [17; 1] and PGI [1]. Note that cMMD and PGI target to learn group invariant predictions conditioning on the label, different from EIC. See Appendix for more details of the four penalties.

The training configurations are presented as follows. For PC-MNIST, we adopt the classifier proposed in [4] for Colored MNIST, which is a MLP with two hidden layers of 390 neurons. The reference model is a MLP with the same structure trained with ERM on the training set, following the setting in EIIL on Colored MNIST. While for MNLI, we use a BERT-based classifier with the standard setup for sentence pair classification [10]. The reference model is the same as the biased classifier propose in [36], which is trained on top of some hand-crafted syntactic features. For each task, all implementations of SCILL and EIIL adopt the same model configurations and pretrained reference models. Since models are tested with OOD data, it is important to specify the model selection strategy, as has been revealed by Gulrajani and Lopez-Paz [12] for the case of domain generalization. In our experiments, we report results with 3 different model selection strategies, including ID, Oracle, and TEV. ID refers to the strategy based on model performance on the in-distribution validation set as

Table 1: Classification accuracy on PC-MNIST under three model selection strategies ID, Oracle, TEV. Val columns contain the accuracy values computed on the in-distribution validation set, and Test columns contain those on the test set. As the label noise rate is set to 0.25 on PC-MNIST, the optimum predictor depending on invariant features achieves an accuracy around 75% on both sets.

Method	Penalty	ID		Oracle		TEV	
		Val	Test	Val	Test	Val	Test
ERM	-	90.22 ± 0.56	50.64 ± 0.56	89.95 ± 0.45	54.53 ± 0.60	-	-
EIIL	IRM	90.21 ± 0.48	50.63 ± 0.45	78.01 ± 0.45	63.63 ± 0.71	69.81 ± 0.27	50.99 ± 0.58
	REx	90.24 ± 0.45	51.21 ± 0.64	79.10 ± 0.43	64.04 ± 0.80	70.05 ± 0.23	51.01 ± 0.68
	cMMD	90.24 ± 0.43	51.36 ± 0.61	77.27 ± 0.28	65.09 ± 0.63	70.15 ± 0.25	52.70 ± 1.40
	PGI	90.19 ± 0.46	51.07 ± 0.54	80.03 ± 1.41	64.27 ± 0.26	70.37 ± 0.14	50.64 ± 0.38
SCILL	IRM	79.65 ± 0.76	62.49 ± 0.55	71.54 ± 0.35	67.46 ± 0.19	71.54 ± 0.35	67.46 ± 0.19
	REx	80.23 ± 0.83	62.13 ± 0.99	72.59 ± 1.44	**67.60** ± 0.24	70.77 ± 0.50	67.33 ± 0.30
	cMMD	83.13 ± 0.93	59.76 ± 0.92	73.12 ± 0.47	67.49 ± 0.52	72.38 ± 0.51	**67.81** ± 0.34
	PGI	80.67 ± 1.75	**62.52** ± 0.32	71.73 ± 1.43	67.26 ± 0.14	71.35 ± 0.24	67.36 ± 0.33
SCILL_{uw}	IRM	90.27 ± 0.39	50.95 ± 0.47	90.07 ± 0.34	53.51 ± 1.38	90.28 ± 0.39	50.85 ± 0.47
maj./min.	IRM	90.18 ± 0.26	50.67 ± 0.15	80.10 ± 0.21	63.85 ± 0.58	90.18 ± 0.26	50.67 ± 0.15
SCILL_{gt}		82.55 ± 0.28	61.12 ± 1.17	74.46 ± 0.25	70.19 ± 0.39	72.30 ± 0.40	70.91 ± 0.06
SCILL_{ub}		84.37 ± 0.53	58.78 ± 0.41	79.27 ± 2.95	59.44 ± 0.44	66.07 ± 0.73	56.20 ± 0.57
opt	-	75	75	75	75	75	75

Table 2: Classification accuracy on HANS.

Method	Penalty	ID		Oracle		TEV	
		Val	Test	Val	Test	Val	Test
ERM	-	84.12 ± 0.15	64.88 ± 3.00	84.12 ± 0.15	64.88 ± 3.00	-	-
EIIL	IRM	84.01 ± 0.08	65.35 ± 0.93	83.82 ± 0.17	66.42 ± 0.98	84.01 ± 0.08	65.35 ± 0.93
	REx	84.10 ± 0.13	65.16 ± 0.19	83.91 ± 0.20	66.87 ± 2.92	84.00 ± 0.48	66.43 ± 1.00
	cMMD	83.56 ± 0.03	63.22 ± 1.76	83.22 ± 0.13	64.25 ± 1.63	83.38 ± 0.20	62.72 ± 2.03
	PGI	84.17 ± 0.08	65.57 ± 2.25	83.78 ± 0.03	66.02 ± 0.93	83.94 ± 0.64	65.57 ± 2.25
SCILL	IRM	82.75 ± 0.17	69.11 ± 1.76	82.56 ± 0.33	68.72 ± 1.24	82.67 ± 0.14	69.82 ± 1.29
	REx	82.68 ± 0.28	**69.73** ± 1.63	82.59 ± 0.22	**71.20** ± 1.81	82.56 ± 0.33	69.75 ± 1.53
	cMMD	82.74 ± 0.26	69.15 ± 1.39	82.39 ± 0.45	70.77 ± 1.40	82.61 ± 0.04	**70.92** ± 0.79
	PGI	82.79 ± 0.30	68.57 ± 0.54	81.69 ± 0.28	70.99 ± 0.48	82.79 ± 0.30	68.57 ± 0.54
EIIL_{lb}	IRM	83.39 ± 0.06	63.90 ± 1.16	83.39 ± 0.06	63.90 ± 1.16	83.16 ± 0.22	61.33 ± 0.33
SCILL_{uw}	IRM	84.15 ± 0.11	64.30 ± 0.67	83.77 ± 0.15	65.93 ± 0.12	83.84 ± 0.02	65.63 ± 1.46

used in [36]. Oracle refers to the selection based on data from the test data distribution, as used in [8; 12]. While TEV is a new strategy adapted from the training-domain validation method in [12] to the inferred groups, which alleviates the dependence on the test data as ID. Details can be found in the appendix.

6.2 Experimental results

Now we demonstrate our experimental results, including performance comparison and detailed analysis. More empirical results including the robustness analysis on the hyperparameter in SCILL can be found in the appendix.

6.2.1 Performance comparison

Table 1 and 2 show the experimental results on PC-MNIST and MNLI-HANS, respectively. The main observation is that all implementations of SCILL consistently outperform the counterpart of EIIL

across all model selection strategies, in terms of the performance on OOD data. Comparing different model selection strategies, Oracle performs the best for both EIIL and SCILL. However, SCILL with the TEV strategy has the ability to outperform EIIL with Oracle, demonstrating the superiority of our new objective. Additionally, SCILL also gains improvements against some debiasing methods utilizing the same reference model (See the appendix).

6.2.2 Ablation study and verification of the two criteria

Our main theoretical results in Section 4, i.e. Theorem 4.2 and 4.4, reveal that the two group criteria are necessary conditions for group-IL to survive spurious correlations. Now we discuss the empirical verification of the significance of the two criteria.

Falsity exposure criterion. To show the significance of the falsity exposure criterion, we compare the performance of methods under the case when the label balance criterion is satisfied. On PC-MNIST, both SCILL and EIIL groups satisfy the label balance criterion[2], while EIIL groups provably violate the falsity exposure, according to Proposition 4.5. The significant improvement of SCILL over EIIL on Table 1 then shows the importance of the falsity exposure criterion. To exclude the effect of the noise in the reference model in the group inference, we further implement SCILL with the ground-truth spurious predictor, obtaining SCILL_{gt} in Table 1. The groups then satisfy the falsity exposure criterion. We construct the ground truth majority/minority split and experiment with IL methods, obtaining results in the row maj./min. in Table 1. The significant performance drop of maj./min. compared with SCILL_{gt} verifies the importance of falsity exposure for group-IL. On MNLI, the label in EIIL groups is unbalanced (See the appendix). Therefore we attach the instance reweight step as in SCILL to EIIL, obtaining EIIL_{lb} which satisfies the label balance criterion. As shown in Table 2, EIIL_{lb} fails to achieve improved performance, which verifies the necessity of falsity exposure.

Label balance criterion. To verify the necessity of the label balance criterion, we investigate the cases when the falsity exposure is satisfied. As the SCILL_{gt} on PC-MNIST satisfies the falsity exposure, we construct such cases by disturbing the label balancing weights in SCILL. We multiply the estimated label proportion of class 0 by 0.5 to get the unbalanced SCILL_{ub}. As shown in Table 1, under Oracle selection, the test accuracy of SCILL_{ub} drops approximately 15% compared with SCILL_{gt}, thus verifying the impact of label balance. More results can be found in the appendix.

We further show the importance of the instance reweight step in SCILL, which is designed following the label balance criterion. For this, we remove the instance reweight step in SCILL, obtaining SCILL_{uw}. The experimental results in Table 1 and 2 show that SCILL_{uw} performs worse than SCILL, demonstrating the importance of the instance reweight step in SCILL.

7 Discussions

The main limitation of the paper is our assumptions on the causal structure. In fact, our conclusions can be generalized to more complex structures, e.g. those in [37] (See the appendix). The most central assumption is the conditional independence between X_{inv} and X_{sp} given Y, which is adopted in many existing works on solving spurious correlations [7; 37; 42]. It would be an important direction to find causal structures on which the assumption is not satisfied while group-IL can still be effective. An extended discussion is provided in the appendix.

As this paper focuses on analyzing group invariant learning, comparing SCILL with other algorithms besides group-IL is beyond the scope of this paper. Notably, the form of the objective of SCILL appears to be similar to two recent methods in solving spurious correlations [26; 32], though the penalties are different. However, they are only applied for the case when spurious features can be explicitly defined [32], and are also discrete as assumed in [26].

Besides the setting of group invariant learning, the two criteria may also bring benefits to the study of domain generalization. Furthermore, our empirical results show SCILL can achieve good performance with other kinds of invariances besides that in invariant learning, e.g. the invariance of the distribution of model outputs conditioned on the label. Discussing the effect of the two group criteria on other kinds of group invariance is a potential research direction.

[2]The ratio of label 0 and label 1 on group 0 and group 1 in EIIL is 1:1.03 and 1:1.04, respectively.

8 Conclusion

This paper is concerned with when group invariant learning (group-IL) can survive spurious correlations. We first formulate the setting of group-IL and necessary assumptions. Then we theoretically analyze the necessary conditions for group-IL in learning spurious-free predictors, and obtain two group criteria, i.e. falsity exposure and label balance. Considering the limitations of previous group-IL methods, we propose a new method SCILL to satisfy the two criteria. Furthermore, we theoretically prove that SCILL has the ability to learn a spurious-free predictor. Finally, we conduct extensive experiments on both synthetic and real data to evaluate the proposed new method. Experimental results show that SCILL significantly outperforms existing SOTA group-IL methods, owing to its ability to satisfy the two criteria. The empirical studies validate our theoretical findings.

9 Acknowledgement

This work was supported by National Key R&D Program of China No. 2021YFF1201600, Vanke Special Fund for Public Health and Health Discipline Development, Tsinghua University (NO.20221080053), and Beijing Academy of Artificial Intelligence (BAAI). The authors would like to thank Keyue Qiu and Yuan Li for providing useful feedback on the draft.

References

[1] Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021.

[2] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization, 2021.

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

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

[5] Jun-Hyun Bae, Inchul Choi, and Minho Lee. BLOOD: Bi-level learning framework for out-of-distribution generalization, 2022. URL `https://openreview.net/forum?id=Cm08egNmrl3`.

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

[7] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4060–4073, 2019.

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

[9] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[11] 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.

[12] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[13] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, 2018.

[14] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[15] 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*, pages 5815–5826. PMLR, 2021.

[16] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[17] 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)*, pages 624–639, 2018.

[18] Yong Lin, Shengyu Zhu, and Peng Cui. Zin: When and how to learn invariance by environment inference? *arXiv preprint arXiv:2203.05818*, 2022.

[19] 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*, pages 6781–6792. PMLR, 2021.

[20] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6804–6814. PMLR, 18–24 Jul 2021.

[21] Tianyu Liu, Zheng Xin, Baobao Chang, and Zhifang Sui. Hyponli: Exploring the artificial patterns of hypothesis-only bias in natural language inference. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6852–6860, 2020.

[22] Richard Lowry. Concepts and applications of inferential statistics, 2014.

[23] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022.

[24] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, 2020.

[25] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

[26] Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

[27] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, 2019.

[28] Junhyun Nam, Hyuntak Cha, Sung-Soo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33, 2020.

[29] Judea Pearl. *Causality*. Cambridge university press, 2009.

[30] 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.

[31] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, 2018.

[32] Aahlad Puli, Lily H Zhang, Eric K Oermann, and Rajesh Ranganath. Predictive modeling in the presence of nuisance-induced spurious correlations. *arXiv preprint arXiv:2107.00520*, 2021.

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

[34] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022.

[35] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1417–1427, 2021.

[36] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. *arXiv preprint arXiv:2005.00315*, 2020.

[37] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34, 2021.

[38] Georg Volk, Stefan Müller, Alexander Von Bernuth, Dennis Hospach, and Oliver Bringmann. Towards robust cnn-based object detection through augmentation with synthetic rain variations. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 285–292. IEEE, 2019.

[39] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

[40] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Taylan Cemgil, et al. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

[41] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.

[42] Ruibin Xiong, Yimeng Chen, Liang Pang, Xueqi Cheng, Zhi-Ming Ma, and Yanyan Lan. Uncertainty calibration for ensemble-based debiasing methods. *Advances in Neural Information Processing Systems*, 34, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [No]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] In the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A The statistical split algorithm

This section introduces details of the statistical split algorithm, which is for the binary classification case. For the multi-class case, the two-sample t-test here should be substituted by one-way ANOVA [22] or Kruskal-Wallis Test [16].

Algorithm 1 Statistical split

Input: $\mathcal{S} = \{f_r(x) | x \in \text{the training set}\}$, $threshold$ for the t-statistic.
Initialize $queue = [S]$, $G = []$
repeat
  Pop the head item B in $queue$
  Divide B into L_0, L_1 according to the label, i.e. $L_0 := \{f_r(x) \in B | \text{the label of } x \text{ is } 0\}$ and $L_1 := B \setminus L_0$
  Compute the two-sample t-statistic T_B of $\log(f_r(x)/(1 - f_r(x)))$ on L_0, L_1 and the corresponding p value
  if $T_B > threshold$ **then**
    Split B using the median value m of $\{f_r(x)_0 | x \in B\}$, i.e. $B' := \{f_r(x) | f_r(x)_0 < m, x \in B\}$, $B'' := B \setminus B'$
    Append B', B'' to the end of the $queue$
  else
    Append B to G
  end if
until $queue$ is empty
Output: G

In our main experiments on PC-MNIST and MNLI, we set the threshold for T_B to 10. It can be seen in Table 3 that the number of groups increases as T_B decreases. The two-sample t-statistic is computed with the function `scipy.stats.ttest_ind` in the Python package `scipy`.

We also experimented with the case when the condition for split the block is set as $p < p_{thr}$, where p is the p-value of the two sample test, p_{thr} is a threshold for the p-value. We set $p_{thr} = 0.01$. Results are shown in Table 3.

B Experimental Details

B.1 Model Selection

It has been argued that model selection is at the heart of domain generalization [12]. In our experiments, methods are also tested with out-of-distribution data, thus it is important to specify the model selection criteria as well. Existing works adopt either training set validation [36] or oracle validation using test data [8] to perform model selection.

In-distribution validation (ID) Hyper-parameters are selected using the in-distribution validation set, i.e. the validation set randomly split from the training set.

Test-distribution validation (Oracle) Hyper-parameters are selected using the test validation set, i.e. the validation set randomly split from the test set.

However, both approaches are suggested as non-optimal, by discussions in several literature [12]. Specifically, in-distribution validation sets can fall short in distinguishing the reference models. Oracle validation supposes the access of test distribution, which sometimes contradicts the setting of debiasing. As a result, we also test with TEV, by adapting the widely used strategy in domain generalization, i.e. Training Environments Validation (TEV) [12] to the inferred reweighted groups. A major advantage of TEV is that it supposes no access to the test data.

Training environments validation (TEV) We split the training set into training and validation subsets. In the validation step, samples in the validation set are allocated to the inferred groups in the training set. Specifically, we denote the average outputs of f_r on each group as its center. Each

Table 3: Robustness study on PC-MNIST. It shows the performance of SCILL-IRM with threshold $5, 10, 20$ on t-statistics in the statistical split algorithm and when it is substituted with a threshold on the p-value. Top 2 values are in bold. Results in Table 1 are all under threshold 10.

Method	#G	ID		Oracle		TEV	
		Val	Test	Val	Test	Val	Test
ERM	-	90.22 $\pm$ 0.56	50.64 $\pm$ 0.56	89.95 $\pm$ 0.45	54.53 $\pm$ 0.60	-	-
SCILL-thr-20	6	83.15 $\pm$ 0.47	60.14 $\pm$ 1.12	73.37 $\pm$ 0.65	**67.95** $\pm$ 0.66	72.59 $\pm$ 0.33	**67.79** $\pm$ 0.57
SCILL-thr-15	7	82.84 $\pm$ 0.61	59.79 $\pm$ 1.00	73.07 $\pm$ 0.69	**68.17** $\pm$ 0.56	72.31 $\pm$ 0.32	**67.87** $\pm$ 0.37
SCILL-thr-10	9	79.65 $\pm$ 0.76	**62.49** $\pm$ 0.55	71.54 $\pm$ 0.35	67.46 $\pm$ 0.19	71.54 $\pm$ 0.35	67.46 $\pm$ 0.19
SCILL-thr-5	15	76.91 $\pm$ 0.60	55.50 $\pm$ 1.78	66.29 $\pm$ 13.1	58.81 $\pm$ 2.35	60.29 $\pm$ 9.97	61.89 $\pm$ 3.96
SCILL-p-0.01	23	79.75 $\pm$ 0.32	**62.47** $\pm$ 0.93	69.63 $\pm$ 0.54	66.78 $\pm$ 0.64	72.63 $\pm$ 1.21	67.46 $\pm$ 0.46

sample in the validation set is allocated to the group with the nearest center, measured by the L_2 distance. The weight of the sample is then set to the same as the training samples of the same label in the group. Hyper-parameters are selected using the reweighted validation set.

B.2 Dataset Details

Patched-Colored MNIST (PC-MNIST) is a synthetic binary classification dataset. It is derived from MNIST, by assigning color and patch to each image as the spurious features. The design of the patch feature is inspired by [5]. Firstly, the handwriting with original digit label 0-4 are labeled 0, and those with 5-9 are labeled 1. Label noise is then added by flipping the label y with probability p_{noise}. After that, the color label is assigned by flipping the label y with probability p_{color}, i.e. $\mathbb{P}(Y = 0|color = 0) = 1 - p_{color}$. Similarly, the patch label is assigned by flipping the label y with probability p_{patch}. We attach a 3×3 black patch on the left top corner to the sample with patch label 1, otherwise on the right bottom corner. In our experiments, the training dataset has $p_{color} = 0.1$, $p_{patch} = 0.3$, and in the test set p_{color} and p_{patch} are both set as 0.5, i.e. uncorrelated with label y. The p_{noise} is set to 0.25 following that on Colored-MNIST [4]. The accuracy on test set is regarded as the performance of the model in solving model's dependence on spurious correlations.

MNLI-HANS is a benchmark widely used in many previous debiasing works, such as [7; 36]. In our experiments, we follow the practice to utilize MNLI [41] as the training data and HANS (Heuristic Analysis for NLI Systems) [27] as the test data. In our experiments, we consider the syntactic spurious correlations, e.g. the lexical overlap between premise and hypothesis sentences is strongly correlated with the entailment label [27]. While for HANS, the specific syntactic correlations are eliminated with manually constructed samples. Therefore, the accuracy on HANS is regarded as the performance of a concerned model in generalizing to the spurious correlation shift.

The statistics of the two datasets are shown as follows.

PC-MNIST. The training set contains 50000 instances from MNLI. In-distribution validation set, oracle set, and test set all contain 5000 instances. All four sets are generated by the same algorithm, only vary in the p_{color} and p_{patch} parameters. The training and validation set have $p_{color} = 0.1$, $p_{patch} = 0.3$. In the oracle and test set p_{color} and p_{patch} are both set as 0.5, i.e. uncorrelated with label y.

MNLI-HANS. MNLI contains approximately 393 thousand training samples. HANS contains 30000 samples. We use the MNLI-matched development as the in-distribution validation set, which contains approximately 10000 samples. The oracle set contains 1000 instances randomly selected from HANS.

B.3 Experimental Settings and Hyper-parameter Tuning

PC-MNIST. The classifier on PC-MNIST is a MLP with two hidden layers of 390 neurons. The reference model has the same structure but was trained with ERM for 100 epochs on the training set. We train each model with 800 epochs. Following Arjovsky et al. [4], the penalty is applied after

Table 4: Classification accuracy on HANS. Results of methods marked with dagger are cited from [36].

Method	Penalty	ID		Oracle		TEV	
		Val	Test	Val	Test	Val	Test
ERM	-	84.12 ± 0.15	64.88 ± 3.00	84.12 ± 0.15	64.88 ± 3.00	-	-
PoE†	-	82.8 ± 0.2	69.2 ± 2.6	-	-	-	-
ConfReg†	-	84.3 ± 0.1	69.1 ± 1.2	-	-	-	-
EIIL	IRM	84.01 ± 0.08	65.35 ± 0.93	83.82 ± 0.17	66.42 ± 0.98	84.01 ± 0.08	65.35 ± 0.93
	REx	84.10 ± 0.13	65.16 ± 0.19	83.91 ± 0.20	66.87 ± 2.92	84.00 ± 0.48	66.43 ± 1.00
	cMMD	83.56 ± 0.03	63.22 ± 1.76	83.22 ± 0.13	64.25 ± 1.63	83.38 ± 0.20	62.72 ± 2.03
	PGI	84.17 ± 0.08	65.57 ± 2.25	83.78 ± 0.03	66.02 ± 0.93	83.94 ± 0.64	65.57 ± 2.25
SCILL	IRM	82.75 ± 0.17	69.11 ± 1.76	82.56 ± 0.33	68.72 ± 1.24	82.67 ± 0.14	69.82 ± 1.29
	REx	82.68 ± 0.28	**69.73** ± 1.63	82.59 ± 0.22	**71.20** ± 1.81	82.56 ± 0.33	69.75 ± 1.53
	cMMD	82.74 ± 0.26	69.15 ± 1.39	82.39 ± 0.45	70.77 ± 1.40	82.61 ± 0.04	**70.92** ± 0.79
	PGI	82.79 ± 0.30	68.57 ± 0.54	81.69 ± 0.28	70.99 ± 0.48	82.79 ± 0.30	68.57 ± 0.54

training for several annealing epochs. Models are tested every 60 epochs to get their accuracy on 3 validation sets.

We conduct grid search on hyper-parameters. The learning rate is searched over $\{1e-4, 5e-4, 1e-3, 5e-3\}$ for all the method. For each invariant learning based method, the penalty weight λ is searched over the range of $\{0.1, 1, 10, 100\}$. The number of annealing epochs is searched over $\{100, 300, 500, 700\}$.

MNLI-HANS. On MNLI, the reference model is the bias-only classifier proposed in [36] which is trained on top of some hand-crafted syntactic features, including (1) whether all words in the hypothesis exist in the premise; (2) whether the hypothesis is a continuous sub-sequence of the premise; (3) the fraction of premise words that shared with hypotheses; (4) the mean, min, max of cosine similarities between word vectors in the premise and the hypothesis.

We follow the default setting in [36] to fine-tune the `bert-base-uncased` model 3 epochs, with the learning rate set to 5×10^{-5}. We follow [1] to set a rate which linearly ramp up the penalty weight according to batch counts. Grid search is also conducted. For each group-IL method, the penalty weight λ is searched over the range of $\{1e-2, 1e-3, 1e-4\}$. The rate to to linearly ramp up λ is searched over $\{0.2, 0.4, 0.6\}$.

B.4 Additional Comparisons

We additionally cite the results on MNLI-HANS of other state-of-the-art methods solving spurious correlations reported in [36]. These methods use the same reference model adopted in our experiments, but adjust the training objective directly based on its outputs. For example, PoE [7] reweights the sample importance via the product-of-expert method. From the table, it shows that SCILL-REx outperforms methods in out-of-distribution accuracy with ID selection strategy. When SCILL is selected with TEV, SCILL-IRM and SCILL-cMMD also show improved performance. However, these baseline methods do not admit the TEV selection strategy, as no group is defined in their algorithms.

B.5 Empirical verification for the two criteria

We conduct experiments on PC-MNIST to verify the significance of the two criteria for group-IL.

To verify the significance of the falsity exposure criterion, we compare the performance of methods under the case when label balance criterion is satisfied. On PC-MNIST, to exclude the effect of the noise in reference model in the group inference, we implement SCILL with the ground-truth spurious predictor, obtaining SCILL_{gt} in Table 6. The groups then satisfy the falsity exposure criterion. We construct the ground truth majority/minority split, which violates the falsity exposure, and experiment

Table 5: Ablation study on PC-MNIST.

Method	Penalty	ID		Oracle		TEV	
		Val	Test	Val	Test	Val	Test
ERM	-	90.22 ± 0.56	50.64 ± 0.56	89.95 ± 0.45	54.53 ± 0.60	-	-
SCILL	IRM	79.65 ± 0.76	62.49 ± 0.55	71.54 ± 0.35	67.46 ± 0.19	71.54 ± 0.35	67.46 ± 0.19
	REx	80.23 ± 0.83	62.13 ± 0.99	72.59 ± 1.44	**67.60** ± 0.24	70.77 ± 0.50	67.33 ± 0.30
	cMMD	83.13 ± 0.93	59.76 ± 0.92	73.12 ± 0.47	67.49 ± 0.52	72.38 ± 0.51	**67.81** ± 0.34
	PGI	80.67 ± 1.75	**62.52** ± 0.32	71.73 ± 1.43	67.26 ± 0.14	71.35 ± 0.24	67.36 ± 0.33
SCILL_{uw}	IRM	90.27 ± 0.39	50.95 ± 0.47	90.07 ± 0.34	53.51 ± 1.38	90.28 ± 0.39	50.85 ± 0.47
	REx	90.25 ± 0.30	51.50 ± 1.08	81.27 ± 0.13	61.63 ± 0.64	90.25 ± 0.30	51.50 ± 1.08
	cMMD	90.31 ± 0.38	51.70 ± 1.02	89.89 ± 0.28	54.50 ± 1.41	90.23 ± 0.32	52.96 ± 0.86
	PGI	90.22 ± 0.47	51.00 ± 0.52	70.05 ± 1.01	66.82 ± 1.01	90.18 ± 0.52	51.44 ± 0.58
opt	-	75	75	75	75	75	75

Table 6: Results on PCMNIST with ground-truth group splits.

Method	Penalty	ID		Oracle		TEV	
		Val	Test	Val	Test	Val	Test
ERM	-	90.22 ± 0.56	50.64 ± 0.56	89.95 ± 0.45	54.53 ± 0.60	-	-
Maj./Min.	IRM	90.18 ± 0.26	50.67 ± 0.15	80.10 ± 0.21	63.85 ± 0.58	90.18 ± 0.26	50.67 ± 0.15
	REx	90.18 ± 0.27	50.74 ± 0.15	78.95 ± 2.49	64.00 ± 1.47	90.18 ± 0.27	50.74 ± 0.15
SCILL_{gt}	IRM	82.55 ± 0.28	61.12 ± 1.17	74.46 ± 0.25	70.19 ± 0.39	72.30 ± 0.40	70.91 ± 0.06
	REx	82.22 ± 0.73	60.16 ± 0.21	73.76 ± 0.25	70.63 ± 0.36	72.21 ± 0.31	71.04 ± 0.04

with IL methods, obtaining results in the row maj./min.. The significant performance drop of maj./min. compared with SCILL_{gt} verifies the importance of falsity exposure for group-IL.

To verify the necessity of label balance criterion, we investigate the cases when the falsity exposure is satisfied. As SCILL_{gt} on PC-MNIST satisfies the falsity exposure, we construct such cases by disturbing the label balancing weights in SCILL. We multiply the estimated label proportion of class 0 by different values $1/p_{err}$ to obtain different degrees of imbalance. As shown in Table 7, label imbalance causes significant performance drop of SCILL-IRM, which verifies the impact of label balance.

We further show the importance of the instance reweight step in SCILL, which is designed following the label balance criterion. For this, we remove the instance reweight step in SCILL, obtaining SCILL_{uw}. The experimental results in Table 5 show that SCILL_{uw} performs worse than SCILL, demonstrating the importance of the instance reweight step in SCILL.

B.6 Robustness analysis

As shown in Section 5.1 in the main paper, the statistical-split algorithm contains a hyper-parameter thr. So We study the robustness of SCILL w.r.t. thr by experiments on PC-MNIST with thr set as $5, 10, 15, 20$ in SCILL-IRM. From the results shown in Table 3, the models are robust with different $thr = 10, 15, 20$, though the model with $thr = 5$ is worse than others.

B.7 Label proportion of EI groups on MNLI

Figure 2 shows the label proportion of the two groups inferred by the EI algorithm in EIIL. It can be seen that $\mathbb{P}(Y = 0|g_0)/\mathbb{P}(Y = 1|g_0) \neq \mathbb{P}(Y = 0|g_1)/\mathbb{P}(Y = 1|g_1)$. As a result, the label balance criterion is violated.

Table 7: Results on PCMNIST with ground-truth group splits and varying label proportion deviation.

Method	p_{err}	ID		Oracle		TEV	
		Val	Test	Val	Test	Val	Test
SCILL$_{\text{gt}}$ -IRM	1	82.55 $\pm$ 0.28	61.12 $\pm$ 1.17	74.46 $\pm$ 0.25	70.19 $\pm$ 0.39	72.30 $\pm$ 0.40	70.91 $\pm$ 0.06
	1.2	84.70 $\pm$ 0.09	59.40 $\pm$ 0.42	79.19 $\pm$ 0.12	65.44 $\pm$ 0.83	77.53 $\pm$ 0.01	63.43 $\pm$ 0.22
	1.5	84.61 $\pm$ 0.36	59.57 $\pm$ 0.23	80.44 $\pm$ 0.79	61.27 $\pm$ 0.10	73.17 $\pm$ 0.08	59.26 $\pm$ 0.21
	2	84.37 $\pm$ 0.53	58.78 $\pm$ 0.41	79.27 $\pm$ 2.95	59.44 $\pm$ 0.44	66.07 $\pm$ 0.73	56.20 $\pm$ 0.57

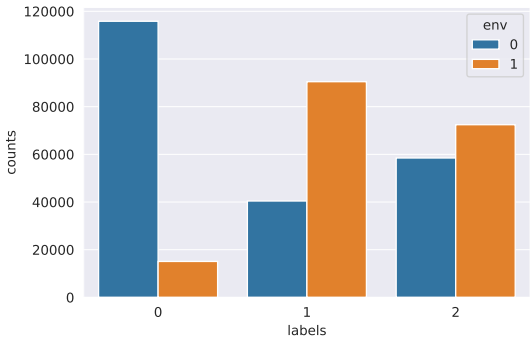


Figure 2: The label proportion varies between the two groups (denoted as 0 and 1 in the figure) inferred by EI on MNLI. The horizontal axis shows 3 labels on MNLI. The vertical axis shows the counts of the instance with the corresponding label in two groups.

B.8 Penalties

We experiment with 4 kinds of invariant learning penalties: IRM [4], REx [15], cMMD [17; 1], and PGI [1].

We follow the notations in the main paper. The penalty of IRM is defined as

$$\textit{penalty}_{\text{IRM}} := \|\nabla_w \mathcal{R}^g(w \circ f)\|^2$$

where $\mathcal{R}^g$ denotes the expected risk on group g, w is a constant scalar multiplier of 1.0 for each output dimension.

With the same notations, the penalty in V-REx writes as follows.

$$\textit{penalty}_{\text{REx}} := \text{Var}(\{\mathcal{R}^g(f)\}_{g \in \mathcal{G}})$$

where $\text{Var}(\cdot)$ denotes the variance.

Different from IRM and V-REx which enhance the invariance of feature conditioned label distribution, cMMD and PGI are two penalties to enhance the invariance of the label conditioned feature distribution across groups, i.e.

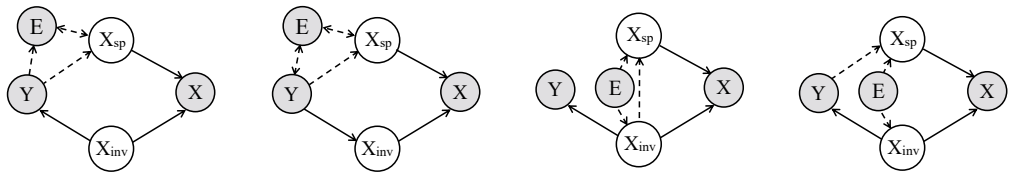
$$\mathbb{P}(f(X)|Y, g) = \mathbb{P}(f(X)|Y, g'), \forall g, g' \in \mathcal{G}.$$

The two penalties are shown to improve model's out-of-distribution generalization performance when used with EIIL in [1]. To show the availability of SCILL, we also experiments the two penalties with SCIIL.

In cMMD, the penalty is defined as the summation of the estimated MMD distances between each pair of conditional distributions, i.e.

$$\begin{aligned}\textit{penalty}_{\text{cMMD}} &:= \sum_{g,g' \in \mathcal{G}} \sum_{y} \widehat{\text{MMD}}(f(g_y), f(g'_y)) \\ &= \sum_{g,g' \in \mathcal{G}} \sum_{y} \sum_{x \in g_y, x' \in g'_y} K(f(x), f(x)) + K(f(x'), f(x')) + 2K(f(x), f(x'))\end{aligned}$$

where $g_y := g \cap \{Y = y\}$, K is a kernel function, which in our implementation is a mixture of 3 Gaussians with bandwidths [1, 5, 10], following [1]. We set $f(x)$ as the logarithm of the model's output probability, as advised in [37].



(a) [4; 37] (b) [33; 26; 37] (c) [2; 25] (d) [39; 23]

Figure 3: The causal graph depicting different assumptions on the data generating process in existing works (some are simplified). Shading indicates the variable is observed. Dotted arrow indicates possible causal relation. The spurious feature is anti-causal or correlates with Y through E in (a) and (b), confounded with the invariant feature in (c), and both anti-causal and confounded in (d).

With the same notations, in PGI, the penalty is defined as

$$\begin{aligned} \textit{penalty}_{\text{PGI}} &:= \sum_i d\left(\hat{\mathbb{E}}_{x\sim\mathbb{P}^g, y=i}\left[f(x)\right], \hat{\mathbb{E}}_{x'\sim\mathbb{P}^{g'}, y'=i}\left[f(x')\right] \right) \\ &= \sum_{g,g'\in\mathcal{G}} \sum_y \operatorname*{mean}_{x\in g_y}\left[f(x)_y\right] \log \frac{\text{mean}_{x'\in g'_y}\left[f(x')_y\right]}{\text{mean}_{x\in g_y}\left[f(x)_y\right]}. \end{aligned}$$

Here $f(x)$ is the probability estimation of the predictor, which follows [1]. $f(\cdot)_y$ denotes the component of $f(\cdot)$ on the dimension corresponding to class y.

C Extended Discussions

Assumptions in this paper. In Section 3, we stated our assumptions on the data generating process as depicted by the causal graphs (a), (b) in Figure 1 of the main paper. In fact, our conclusions can be further extended to causal structures shown in Figure 3 (a), (b). Compared with Figure 1, Figure 3 (a) further includes the case when E is a child of both X_{sp} and Y, depicting the case that X_{sp} and Y are subject to different *selection* mechanisms in different domains, as introduced in [37]. Figure 3 (b) further includes 1) E is a child of both X_{sp} and Y; 2) E is a confounder of X_{sp} and Y. In all these cases, we have $X_{inv} \perp\!\!\!\perp X_{sp}|Y$. It is the only condition required in our proofs for theorems and statements in this paper, except for SFC, which needs Y to be a backdoor variable between X_{sp} and X.

The conditional independence condition $X_{inv} \perp\!\!\!\perp X_{sp}|Y$ is an essential assumption in many related works on solving spurious correlations [37; 7; 42; 32]. For example, it is required in the proof of conditions of Theorem 4.2 in [37]. The nuisance-varying family defined in [32] satisfies that $p(x|x_b, y)$ keeps invariant, which is equivalent to $X_{inv} \perp\!\!\!\perp X_{sp}|Y$. It would be an important direction to find causal structures on which the assumption is not satisfied while group invariant learning can still be effective. For example, for the causal structure in Figure 3 (c), group invariant learning may still handy when combined with an additional information bottleneck penalty [2].

The algorithm SCIIL. In this paper, the algorithm SCIIL is proposed as a possible but not necessarily optimal solution to meet the two criteria for group-IL. As this paper focuses on analyzing group invariant learning, comparing SCILL with other algorithms besides group-IL is beyond the scope of this paper. However, it can be observed that SCILL has some advantages compared with existing methods on solving spurious correlations.

Notably, the form of the objective of SCILL appears to be similar to those in two recent methods [26; 32]. They both contain a risk term reweighted by estimations of spurious correlations and an feature invariance penalty. However, they are only applied for the case when spurious features can be explicitly defined [32], and are also discrete as assumed in [26]. Also, their feature invariance penalty is different from that in IL. Specifically, Makar et al. [26] divide samples into groups according to their spurious feature, and define the pairwise MMD distance of the distributions of embeddings on these groups as the penalty. It is equivalent to SCILL+cMMD when the spurious feature takes binary values. [32] suppose the access of X_{sp} and use a parameterized penalty term which approximates the

mutual information $I[(f(X), Y)|X_{sp}]$. Instead, SCILL only assume the access of a reference model, which fits for more general cases when X_{sp} is high dimensional or not predefined.

Compared with some other methods that exploiting a reference model [7; 24; 36; 28; 19; 42], the first term in SCIIL resembles their targets where samples are reweighted according to the outputs of the reference model. However, the IL penalty in SCIIL serves as an additional regularization. Results in Section B.4 empirically show SCILL outperforms methods in [7; 36] with the same reference model in out-of-distribution accuracy.

D Proofs

This section contains the following proofs: D.2 proof for the statement in Section 3 on the causal graph; D.3 proof for Theorem 4.2; D.4 proof for the statement in Section 4.2 that SFC is sufficient for $f(X)$ to be invariant to the intervention on spurious features; D.5 proof for Theorem 4.4; D.6 proof for the statement in Section 4.3; D.7 proof for Proposition 4.5; and D.8 proof for Theorem 5.1.

Notations. In the following contents, we denote that $(X, Y) \sim \mathbb{P}(\mathcal{X} \times \mathcal{Y})$. The image set of X_{sp}, X_{inv} is respectively denoted as $\mathcal{B}, \mathcal{S}$. $X = r(X_{sp}, X_{inv})$, where r is a bijective function. We denote x_{sp}, x_{inv} as the corresponding values of X_{sp}, X_{inv} for a given value $x \in \mathcal{X}$, i.e. $x = r(x_{sp}, x_{inv})$. $\mathcal{G}$ denotes a set of sets in $\mathcal{X} \times \mathcal{Y}$ which satisfies $\cup_{g \in \mathcal{G}} g = \mathcal{X} \times \mathcal{Y}$. $\mathcal{G}^{\mathcal{Y}} := \{g \cap \{Y = y\}\}_{g \in \mathcal{G}, y \in \mathcal{Y}}$. As $f(x) = f(r(x_{sp}, x_{inv}))$, for convenience we abbreviate $f(x) = f(x_{sp}, x_{inv})$.

D.1 Lemmas

We first prove the following lemmas.

Lemma D.1. *If a set $g \in \mathcal{X} \times \mathcal{Y}$ can be formed by a set of sets $\{g_i\}_{i \in \mathcal{I}} \subset \mathcal{G}^{\mathcal{Y}}$ under set union, then $\forall g', g'' \in \mathcal{G}$*

$$\mathbb{P}(f(X)|g', Y = y) = \mathbb{P}(f(X)|g'', Y = y), \forall y$$

induces $\forall g' \in \mathcal{G}$

$$\mathbb{P}(f(X)|g, Y = y) = \mathbb{P}(f(X)|g', Y = y), \forall y$$

Proof. We only need to prove the case when for any y, $\exists g_1, g_2 \in \mathcal{G}^{\mathcal{Y}}$, $g \cap \{Y = y\} = g_1 \cup g_2$. As

$$\mathbb{P}(f(X)|g, Y = y) = \mathbb{P}(f(X)|g_1, Y = y)\frac{\mathbb{P}(g_1, Y = y)}{\mathbb{P}(g, Y = y)}) + \mathbb{P}(f(X)|g_2, Y = y)\frac{\mathbb{P}(g_2, Y = y)}{\mathbb{P}(g, Y = y)})$$

By $\mathbb{P}(f(X)|g_1, Y = y) = \mathbb{P}(f(X)|g_2, Y = y)$, $\mathbb{P}(g, Y = y) = \mathbb{P}(g_1, Y = y) + \mathbb{P}(g_2, Y = y)$, we have $\mathbb{P}(f(X)|g, Y = y) = \mathbb{P}(f(X)|g_1, Y = y) = \mathbb{P}(f(X)|g_2, Y = y)$. □

Lemma D.2. *Suppose the following conditions are satisfied:*
(a) $\mathbb{P}(Y = y|g)/\mathbb{P}(Y = y'|g) = \mathbb{P}(Y = y|g')/\mathbb{P}(Y = y'|g'), \forall g, g' \in \mathcal{G}$ and $\forall y, y' \in \mathcal{Y}$ satisfying $\mathbb{P}(Y = y|g), \mathbb{P}(Y = y'|g), \mathbb{P}(Y = y|g'), \mathbb{P}(Y = y'|g') \neq 0$.
(b) $\mathcal{G}$ only depends on X_{sp}, and $\forall g \in \mathcal{G}$, $\exists c_{g,y}$ s.t. $\mathbb{P}[X_{sp} = x_{sp}, Y = y] = c_{g,y}, \forall x \in g, y \in \mathcal{Y}$.
(c) $f(X) \perp\!\!\!\perp X_{sp}|g$, and $f(X)$ differs with different $\mathbb{P}(Y|X_{inv})$ given g.
Then EIC induces SFC.

Proof. Denote $\mathcal{S}^b_{\mathbf{a}} := \{s \in \mathcal{S}|f(r(s, b)) = \mathbf{a}\}$, $\mathcal{B}_g := \{b \in \mathcal{B}|\exists s, r(s, b) \in g\}$. As $\forall b, b' \in \mathcal{B}_b, \mathbb{P}(X_{sp} = b, Y = y) = \mathbb{P}(X_{sp} = b', Y = y), \forall y$, we have $\mathbb{P}(X_{sp} = b) = \mathbb{P}(X_{sp} = b')$. Then $\mathbb{P}(Y = y|X_{sp} = b) = \mathbb{P}(Y = y|X_{sp} = b'), \forall y$. As

$$\mathbb{P}(Y = y|g) = \sum_{b \in \mathcal{B}_g} \mathbb{P}(Y = y|X_{sp} = b)\mathbb{P}(X_{sp} = b|g) = \mathbb{P}(Y = y|X_{sp} = b).$$

Suppose $\forall y, b, \mathbb{P}(Y = y|X_{sp} = b) \neq 0$. Then we have $\forall b, b' \in \mathcal{B}$,

$$\mathbb{P}(Y = y|X_{sp} = b) = \mathbb{P}(Y = y|X_{sp} = b') = \mathbb{P}(Y = y), \forall y.$$

Now

$$\mathbb{P}(Y=y|f(X)=\mathbf{a},g)=\mathbb{P}(Y=y|\cup_{b\in\mathcal{B}_g}\{X_{inv}\in\mathcal{S}_{\mathbf{a}}^b,X_{sp}=b\})$$
$$=\frac{\sum_{b\in\mathcal{B}_g}\mathbb{P}(Y=y,X_{inv}\in\mathcal{S}_{\mathbf{a}}^b,X_{sp}=b)}{\sum_{b\in\mathcal{B}_g}\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^b,X_{sp}=b)}.$$

As $X_{inv}\perp\!\!\!\perp X_{sp}|Y$, we have

$$\mathbb{P}(Y=y,X_{inv}\in\mathcal{S}_{\mathbf{a}}^b,X_{sp}=b)=\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^b,X_{sp}=b|Y=y)\mathbb{P}(Y=y)$$
$$=\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^b|Y=y)\mathbb{P}(X_{sp}=b,Y=y)$$

As a result,

$$\mathbb{P}(Y=y|f(X)=\mathbf{a},g)=\frac{\sum_{b\in\mathcal{B}_g}\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^b|Y=y)\mathbb{P}(X_{sp}=b,Y=y)}{\sum_{b\in\mathcal{B}_g}\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^b,X_{sp}=b)}$$
$$=\frac{\mathbb{P}(Y=y,X_{sp}=b_g)\sum_{b\in\mathcal{B}_g}\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^b|Y=y)}{\sum_{b\in\mathcal{B}_g}\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^b,X_{sp}=b)}$$
$$=\frac{\mathbb{P}(Y=y,X_{sp}=b_g)\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^g|Y=y)}{\sum_{b\in\mathcal{B}_g}\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^b,X_{sp}=b)}$$

where b_g is any element in $\mathcal{B}_g$, $\mathcal{S}_{\mathbf{a}}^g:=\cup_{b\in\mathcal{B}_g}\mathcal{S}_{\mathbf{a}}^b$. Note that condition (c) induces $\mathcal{S}_{\mathbf{a}}^b=\mathcal{S}_{\mathbf{a}}^{b'}=\mathcal{S}_{\mathbf{a}}^g$, $\forall b,b'\in g$, and the condition (b) induces $X_{sp}\perp\!\!\!\perp X_{inv}|g$. We have

$$\mathbb{P}(Y=y|f(X)=\mathbf{a},g)=\frac{\mathbb{P}(Y=y,X_{sp}=b_g)\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^g|Y=y)}{\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^g)\mathbb{P}(g)}$$
$$=\frac{\mathbb{P}(Y=y,X_{sp}=b_g)\mathbb{P}(Y=y|X_{inv}\in\mathcal{S}_{\mathbf{a}}^g)}{\mathbb{P}(Y=y)\mathbb{P}(g)}$$
$$=\frac{\mathbb{P}(X_{sp}=b_g)}{\mathbb{P}(g)}\mathbb{P}(Y=y|X_{inv}\in\mathcal{S}_{\mathbf{a}}^g)$$

We have

$$\frac{\mathbb{P}(X_{sp}=b_g)}{\mathbb{P}(g)}=\frac{\mathbb{P}(X_{sp}=b_{g'})}{\mathbb{P}(g')},\mathbb{P}(Y=y|X_{inv}\in\mathcal{S}_{\mathbf{a}}^g)=\mathbb{P}(Y=y|X_{inv}\in\mathcal{S}_{\mathbf{a}}^{g'})$$

As $f(X)\perp\!\!\!\perp X_{sp}|g$, we have $\mathcal{S}_{\mathbf{a}}^b=\mathcal{S}_{\mathbf{a}}^g$, $\forall b\in g$. Then we have $\forall b,b'\in\mathcal{B}$, $\mathbb{P}(Y=y|X_{inv}\in\mathcal{S}_{\mathbf{a}}^b)=\mathbb{P}(Y=y|X_{inv}\in\mathcal{S}_{\mathbf{a}}^{b'})$. As $\mathbb{P}(Y|X_{inv}=s)$ is constant, $\forall s\in\mathcal{S}_{\mathbf{a}}^b$, we have $f(X)\perp\!\!\!\perp X_{sp}$. As a result $\mathcal{S}_{\mathbf{a}}^b=\mathcal{S}_{\mathbf{a}}$. As

$$\mathbb{P}(f(X)=\mathbf{a}|g,Y=y)=\sum_b\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}^b|X_{sp}=b,Y=y)\mathbb{P}(X_{sp}=b|g,Y=y)$$
$$=\frac{\mathbb{P}(X_{sp}=b_g)}{\mathbb{P}(g)}\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}|Y=y)$$

We have

$$\mathbb{P}(f(X)=\mathbf{a}|g,Y=y)=\mathbb{P}(f(X)=\mathbf{a}|g',Y=y).$$

And $\mathbb{P}(f(X)=\mathbf{a}|X_{sp}=b,Y=y)=\mathbb{P}(X_{inv}\in\mathcal{S}_{\mathbf{a}}|Y=y)=\mathbb{P}(f(X)=\mathbf{a}|X_{sp}=b',Y=y)$, i.e. SFC is satisfied. □

D.2 Proof for the statement in Section 3

The statement in Section 3. When the causal model of the data generating process follows the causal graph in Figure 1(d) in the main paper, whether the invariant mechanism holds on each group is indeterminate without additional assumptions on the mechanisms between X_{sp} and X_{inv}.

Proof. Consider $\mathbb{P}(Y|X_{sp}, X_{inv})$, we have

$$\mathbb{P}(Y|X_{sp}, X_{inv}) = \frac{\mathbb{P}(X_{sp}|Y, X_{inv})\mathbb{P}(Y|X_{inv})}{\mathbb{P}(X_{sp}|X_{inv})} \propto \mathbb{P}(X_{sp}|Y, X_{inv})\mathbb{P}(Y|X_{inv})$$

It shows that the relation between $\mathbb{P}(Y|X_{sp}, X_{inv})$ and $\mathbb{P}(Y|X_{inv})$ is affected by $\mathbb{P}(X_{sp}|Y, X_{inv})$. However,

$$\mathbb{P}(X_{sp}|Y, X_{inv}) = \sum_{e \in \mathcal{E}_{all}} \mathbb{P}(X_{sp}|Y, E)\mathbb{P}(E|Y, X_{inv})$$

As the mechanisms between Y, E and X_{sp}, and between E and X_{inv} are unknown, so does $\mathbb{P}(X_{sp}|Y, X_{inv})$. As a result, the relation between $\mathbb{P}(Y|X_{inv}, X_{sp})$ and $\mathbb{P}(Y|X_{inv})$ is indeterminate. As $g \in \mathcal{G}$ is an event in $\sigma(X_{sp}, Y)$, we have the relation between $\mathbb{P}(Y|X_{inv}, g)$ and $\mathbb{P}(Y|X_{inv})$ is indeterminate. □

D.3 Proof for Theorem 4.2

Theorem 4.2 in the main paper states as follows.

Theorem D.3. *Suppose the falsity exposure criterion is violated, i.e. $\exists h$ satisfies $\mathbb{P}(Y|h(X_{sp}), g) = \mathbb{P}(Y|h(X_{sp}), g') \neq \mathbb{P}(Y), \forall g, g' \in \mathcal{G}$. Then the optimal solution of group-IL is $f(X) = \mathbb{P}[Y|X_{inv}, h(X_{sp})]$, which fails to generalize when $\mathbb{P}(Y|X_{sp})$ shifts.*

Proof. We first prove that $\Phi(X) = (X_{inv}, h(X_{sp}))$ satisfies EIC. As X_{inv} and X_{sp} are conditionally independent given Y, and groups are only defined by X_{sp} and Y, we have $\forall g, g' \in \mathcal{G}$,

$$\begin{aligned}\mathbb{P}[Y|X_{inv}, h(X_{sp}), g] &= \frac{\mathbb{P}[X_{inv}, h(X_{sp}), g|Y]\mathbb{P}(Y)}{\mathbb{P}[X_{inv}, h(X_{sp}), g]} = \frac{\mathbb{P}[X_{inv}|Y]\mathbb{P}[h(X_{sp}), g|Y]\mathbb{P}(Y)}{\mathbb{P}[X_{inv}, h(X_{sp})]} \\ &= \frac{\mathbb{P}[Y, X_{inv}]\mathbb{P}[Y|h(X_{sp}), g]\mathbb{P}[h(X_{sp}), g]}{\mathbb{P}[X_{inv}, h(X_{sp}), g]\mathbb{P}(Y)} \propto \frac{\mathbb{P}[Y|X_{inv}]\mathbb{P}[Y|h(X_{sp}), g]}{\mathbb{P}(Y)}\end{aligned}$$

As a result, $\mathbb{P}[Y|X_{inv}, h(X_{sp}), g] = \mathbb{P}[Y|X_{inv}, h(X_{sp}), g'] = \mathbb{P}[Y|X_{inv}, h(X_{sp})]$, and $\mathbb{P}[Y|X_{inv}, h(X_{sp})] \neq \mathbb{P}[Y|X_{inv}]$. Without the loss of generality, we suppose any other h' which satisfies $\mathbb{P}(Y|h(X_{sp}), g) = \mathbb{P}(Y|h(X_{sp}), g') \neq \mathbb{P}(Y), \forall g, g' \in \mathcal{G}$ is a function of h, i.e. $\exists l$ s.t. $h'(x) = l(h(x))$. In the objective function of group-IL, the optimal predictor is optimized with the cross-entropy loss. By the Jensen-Inequality, among all the functions of $\Phi(X)$, $\mathbb{P}[Y|X_{inv}, h(X_{sp})]$ minimizes the loss. When $\mathbb{P}(Y|X_{sp})$ encounters arbitrary changes, so does $\mathbb{P}(Y|h(X_{sp}))$. As $\mathbb{P}[Y|X_{inv}, h(X_{sp})]$ is propositional to $\mathbb{P}(Y|h(X_{sp}))$, it can also change in a new domain. □

D.4 Proof for the sufficiency of SFC

The statement in Section 4.2. SFC is a sufficient condition for a function $f(X)$ to be invariant to the intervention [29] on X_{sp}.

Proof. Specifically, the condition "$f(X)$ is invariant to the intervention on X_{sp}" writes as

$$\mathbb{P}(f(X)|do(X_{sp} = b)) = \mathbb{P}(f(X)|do(X_{sp} = b')), \forall b, b' \in \mathcal{B}.$$

Equivalently, we can say X_{sp} has no causal effects on $f(X)$. We consider the causal structures (a) and (b) shown in Figure 3. In both graphs, Y is a backdoor variable from X_{sp} to X, as it blocks all the backdoor path from X_{sp} to X with an arrow into X_{sp}, and it is not a child of X_{sp} (note that, in (b), we assume the arrows between (E, Y) points to Y only when E is the con-founder of Y and X_{sp}). Then by the Back-door criterion [29],

$$\mathbb{P}(X|do(X_{sp} = b)) = \sum_y \mathbb{P}(X|X_{sp} = b, Y = y)\mathbb{P}(Y = y).$$

As a result, for any function f,

$$\mathbb{P}(f(X)|do(X_{sp} = b)) = \sum_y \mathbb{P}(f(X)|X_{sp} = b, Y = y)\mathbb{P}(Y = y).$$

It is straightforward that when

$$\mathbb{P}(f(X)|X_{sp} = b, Y) = \mathbb{P}(f(X)|X_{sp} = b', Y), \forall b, b' \in \mathcal{B}, \tag{SFC}$$

we have $\forall b, b' \in \mathcal{B}$,

$$\mathbb{P}(f(X)|do(X_{sp} = b)) = \sum_y \mathbb{P}(f(X)|X_{sp} = b', Y = y)\mathbb{P}(Y = y) = \mathbb{P}(f(X)|do(X_{sp} = b')).$$

That ends our proof. □

D.5 Proof for Theorem 4.4

We repeat Theorem 4.4 here.

Theorem D.4. *With a set of groups $\mathcal{G}$ inferred by (X_{sp}, Y), if the label balance criterion is violated, functions satisfying EIC can not satisfy SFC.*

Proof. Suppose a function f satisfies EIC, i.e.

$$\mathbb{P}(Y|f(X) = \mathbf{a}, g) = \mathbb{P}(Y|f(X) = \mathbf{a}, g'), \forall g, g' \in \mathcal{G}. \tag{EIC}$$

where $\mathcal{G}$ is defined by some function h_G of (X_{sp}, Y), i.e. $\forall g \in \mathcal{G}$, $g := \{(x, y)|h_G(x_{sp}, y) \in S_g\}$ for some set S_g. Note that here we do not distinguish whether f is the predictor or the feature extractor Φ, because the two has no clear theoretical distinction.

Recall that SFC is stated as

$$\mathbb{P}(f(X)|X_{sp} = b, Y = y) = \mathbb{P}(f(X)|X_{sp} = b', Y = y), \forall b, b' \in \mathcal{B}. \tag{SFC}$$

Suppose f also satisfies SFC. Define $S_g^B := \{b \in \mathcal{B}|\exists y, h_G(b, y) \in S_g\}$, we have

$$\mathbb{P}(f(X) = \mathbf{a}|g, Y = y) = \mathbb{P}(f(X) = \mathbf{a}|X_{sp} \in S_g^B, Y = y) = \mathbb{P}(f(X) = \mathbf{a}|X_{sp} = b, Y = y),$$

we have for $\forall y \in \mathcal{Y}, g, g' \in \mathcal{G}$, satisfying $\mathbb{P}(g, Y = y) \neq 0, \mathbb{P}(g', Y = y) \neq 0$,

$$\mathbb{P}(f(X) = \mathbf{a}|g, Y = y) = \mathbb{P}(f(X) = \mathbf{a}|g', Y = y), \forall y$$

by EIC, we have

$$\frac{\mathbb{P}(Y = y|g)}{\mathbb{P}(Y = y|g')} = \frac{\mathbb{P}(f(X) = \mathbf{a}|g)}{\mathbb{P}(f(X) = \mathbf{a}|g')}, \forall y.$$

Then for another $y' \in \mathcal{Y}$ satisfying $\mathbb{P}(g, Y = y') \neq 0, \mathbb{P}(g', Y = y') \neq 0$,

$$\frac{\mathbb{P}(Y = y|g)}{\mathbb{P}(Y = y'|g)} = \frac{\mathbb{P}(Y = y|g')}{\mathbb{P}(Y = y'|g')}.$$

As a result, if f satisfies SFC, the above condition must be satisfied. □

D.6 Proof for the statement in Section 4.3

The statement in Section 4.3. On both colored-MNIST [4] and coloured-MNIST [1], Y has a uniform distribution, and the spurious correlation has the same ratio for any spurious features, e.g. $\mathbb{P}(Y = 0|\text{color} = \textit{green}) = \mathbb{P}(Y = 1|\text{color} = \textit{red}))$ on colored-MNIST. It can be proved that in this case the majority/minority groups satisfy both criteria.

Proof. Denote

$$\mathbb{P}(Y = 0|\text{color} = \textit{green}) = \mathbb{P}(Y = 1|\text{color} = \textit{red})) = p > 0.5$$

As $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1)$, we have

$$\mathbb{P}(\text{color}|Y) = \frac{\mathbb{P}(Y|\text{color})\mathbb{P}(\text{color})}{\mathbb{P}(Y)} \propto \mathbb{P}(Y|\text{color})$$

then $\mathbb{P}(\text{color} = \textit{green}|Y = 0) = \mathbb{P}(\text{color} = \textit{red}|Y = 1) = p$. The majority group g_{maj}, as proved in Proposition 1 in [8], consists of $\{\text{color} = \textit{green}, Y = 0\}$ and $\{\text{color} = \textit{red}, Y = 1\}$. As a result,

$$\frac{\mathbb{P}(Y = 0|g_{maj})}{\mathbb{P}(Y = 1|g_{maj})} = \frac{\mathbb{P}(\text{color} = \textit{green}, Y = 0)}{\mathbb{P}(\text{color} = \textit{red}, Y = 1)} = 1$$

Similarly, for the minority group g_{min},

$$\frac{\mathbb{P}(Y = 0|g_{maj})}{\mathbb{P}(Y = 1|g_{maj})} = \frac{\mathbb{P}(\text{color} = \textit{red}, Y = 0)}{\mathbb{P}(\text{color} = \textit{green}, Y = 1)} = \frac{1-p}{1-p} = 1$$

The label-balance criterion is thus satisfied. As any function h of the color feature satisfies $\mathbb{P}(Y = 0|h, g_{maj}) = \mathbb{P}(Y = 0|g_{maj}) = \mathbb{P}(Y = 0)$, the falsity exposure criterion is satisfied. □

D.7 Proof for Proposition 4.5

The proposition states as follows.

Proposition D.5. *Suppose we have $(X, Y) \sim \mathbb{P}(X, Y)$. Y takes value in $\{0, 1\}$. X is formed with spurious feature variables $X_{sp} = (B_0, B_1)$, and invariant feature variable S, i.e. $X = r(B_0, B_1, S)$, for some injective function r. B_0 and B_1 are both binary variables, which take values in $\{b_0^0, b_0^1\}$ and $\{b_1^0, b_1^1\}$ respectively. B_0, B_1 and S are conditionally independent given Y. Denote $\mathbb{P}(Y = j|B_i = b_i^j) = p_i, \forall j = 0, 1$. Suppose $p_0 > p_1$. Then we have 1) the majority/minority split e_{maj}, e_{min} violates falsity exposure criterion. 2) the optimal classifier under invariant learning objectives depends on B_1.*

Proof. Without the loss of generality, we suppose $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1)$. Denote B_i takes value in $\{b_i^0, b_i^1\}$. We have $\mathbb{P}(Y = j|B_i = b_i^j) = p_i$, $i = 0, 1$, $j = 0, 1$. As B_i are conditionally independent given Y, we have $\mathbb{P}(Y|B_0, B_1) \propto \mathbb{P}(Y|B_1)\mathbb{P}(Y|B_0)$. As $p_0 > p_1$, we have

$$\mathbb{P}(Y = 0|B_0 = b_0^0, B_1 = b_1^1) > \mathbb{P}(Y = 1|B_0 = b_0^0, B_1 = b_1^1)$$
$$\mathbb{P}(Y = 1|B_0 = b_0^1, B_1 = b_1^0) > \mathbb{P}(Y = 0|B_0 = b_0^0, B_1 = b_1^1)$$

The majority group e_{maj} then consists of the following data sets:

$$\{B_0 = b_0^0, B_1 = b_1^0, Y = 0\}, \{B_0 = b_0^0, B_1 = b_1^1, Y = 0\},$$
$$\{B_0 = b_0^1, B_1 = b_1^0, Y = 1\}, \{B_0 = b_0^1, B_1 = b_1^1, Y = 1\}$$

In the majority set,

$$\begin{aligned}\mathbb{P}(Y = 0|B_1 = b_1^0, e_{maj}) &= \frac{\mathbb{P}(Y = 0, B_1 = b_1^0, B_0 = b_0^0)}{\mathbb{P}(Y = 1, B_1 = b_1^0, B_0 = b_1^1)} \\ &\propto \frac{\mathbb{P}(Y = 0, B_1 = b_1^0)}{\mathbb{P}(Y = 1, B_1 = b_1^1)} = \frac{\mathbb{P}(Y = 0|B_1 = b_1^0)}{\mathbb{P}(Y = 1|B_1 = b_1^1)}\end{aligned}$$

As a result, $\mathbb{P}(Y = 0|B_1 = b_1^0, e_{maj}) = \mathbb{P}(Y = 0|B_1 = b_1^0)$. Similarly,

$$\mathbb{P}(Y = 1|B_1 = b_1^1, e_{maj}) = \frac{\mathbb{P}(Y = 1, B_1 = b_1^1, B_0 = b_0^1)}{\mathbb{P}(Y = 0, B_1 = b_1^0, B_0 = b_0^0)} \propto \frac{\mathbb{P}(Y = 0|B_1 = b_1^1)}{\mathbb{P}(Y = 0|B_1 = b_1^1)}$$

As a result $\mathbb{P}(Y = 1|B_1 = b_1^1, e_{maj}) = \mathbb{P}(Y = 1|B_1 = b_1^1)$. Then we have $\mathbb{P}(Y|B_1, e_{maj}) = \mathbb{P}(Y|B_1, e_{min})$, which means B_1 satisfies EIC. As S is invariant, according to the proof of Theorem. 4.2 in D.3, we have $f^* = \mathbb{P}(Y|S, B_1)$. □

D.8 Proof for Theorem 5.1

We repeat the theorem as follows.

Theorem D.6. *If $\mathcal{G}$ satisfies $f_r^*(X) \perp\!\!\!\perp Y|g$, $\forall g \in \mathcal{G}$, where $f_r^* : \mathcal{X} \to \mathcal{Y}$ is spurious-only, i.e. $\sigma(X_{sp})$-measurable, and minimizes the prediction loss $\mathcal{L}_{ce}^r = \mathbb{E}[\sum_y \mathbb{P}(Y = y|X) \log f_r(X)_y]$, the*

optimal model minimizing the following objective satisfies SFC.

$$\begin{aligned}\mathcal{L}(f) &:= \sum_{g\in\mathcal{G}} \mathbb{E}[w^g(Y)\mathcal{L}^g(f(X),Y)] + \lambda \cdot \textit{penalty}(\{S_g(f)\}_{g\in\mathcal{G}}) \\ &=: \sum_{g\in\mathcal{G}} \tilde{\mathcal{R}}^g(f) + \lambda \cdot \textit{penalty}(\{S_g(f)\}_{g\in\mathcal{G}})\end{aligned}$$

where $\tilde{\mathcal{R}}^g(f)$ is defined as

$$\tilde{\mathcal{R}}^g(f) = \mathbb{E}_{x,y\in g}\, \omega^g(y)\mathcal{L}_{ce}(y, f(X)), \omega^g(y) := \mathbb{P}(Y=y)/\mathbb{P}(Y=y|g) \tag{4}$$

Proof. For convenience, in the following we denote $w_y^g := \omega^g(y)$, $x(b,s) := r(X_{sp}=b, X_{inv}=s)$, $x = r(X_{sp}=x_b, X_{inv}=x_s)$. We use p with lower-cased letters to denote the probability of the event that the corresponding random variable denoted by the upper-cased letter equals that value, e.g. $p(x,y) := \mathbb{P}(X=x, Y=y)$, $p(x|g) = \mathbb{P}(X=x|g)$. $f(\cdot)_y$ denotes the component of $f(\cdot)$ on the dimension corresponding to class y.

Denote f_r as the reference model, which satisfies $f_r(x) = l(x_b)$, where l is a classifier $l : \mathcal{B} \to \mathcal{Y}$. Denote θ_r as the parameter of f_r, the training loss of f_r is defined as

$$\begin{aligned}\mathcal{L}_{ce}(f_r,\theta_r) &= \sum_x p(x) \sum_y p(y|x) \log f_r(x,\theta_r)_y \\ &= \sum_{x_b,x_s} p(x_b,x_s) \sum_y p(y|x_b,x_s) \log l(x_b,\theta_r)_y \\ &= \sum_{x_b,y} p(y,x_b) \log l(x_b,\theta_r)_y\end{aligned}$$

The above equation induces that for $f_r^* := f_r(x,\theta^*)$, where $\theta^* := \arg\min \mathcal{L}_{ce}(f_r,\theta)$, $f_r^*(x) = f_r^*(x')$ if and only if $p(y,x_b) = p(y,x_b')$. If we define $g_{\mathbf{a}} := \{x | f_r^*(x) = \mathbf{a}\}$, we have $\mathbb{P}(X_{sp} = b|g_{\mathbf{a}}) = \mathbb{P}(X_{sp} = b'|g_{\mathbf{a}}), \forall b, b' \in \mathcal{B}$, and $p(y|g_{\mathbf{a}}) = a_y$. We denote the set of strata of f_r as $\mathcal{G}_{f_r}$. Now for any $g \in \mathcal{G}_{f_r}$,

$$\begin{aligned}\tilde{\mathcal{R}}^g(f) &= \sum_x p(x|g) \sum_y w_y^g p(y|x,g) \log(f(x)_y) \\ &= \sum_{b,s} \mathbb{P}(X_{inv}=s, X_{sp}=b|g) \sum_y w_y^g \mathbb{P}(y|X_{inv}=s, X_{sp}=b, g) \log(f(x(b,s))_y)\end{aligned}$$

By the conditional independence of X_{inv} and X_{sp} given Y,

$$\mathbb{P}(X_{inv}=s, X_{sp}=b|Y=y) = \mathbb{P}(X_{inv}=s|Y=y)\mathbb{P}(X_{sp}=b|Y=y) \tag{5}$$

As e is a function of X_{sp}, Y, we have

$$\mathbb{P}(X_{inv}=s, X_{sp}=b, g|Y=y) = \mathbb{P}(X_{inv}=s|Y=y)\mathbb{P}(X_{sp}=b, g|Y=y) \tag{6}$$

By the above, we have

$$\begin{aligned}\tilde{\mathcal{R}}^g(f) &= \sum_{b,s} \mathbb{P}(X_{inv}=s, X_{sp}=b|g) \sum_y w_y^g \mathbb{P}(Y=y|X_{inv}=s, X_{sp}=b, g) \log(f(x(b,s))_y) \\ &= \sum_{b,s} \mathbb{P}(X_{inv}=s, X_{sp}=b|g) \sum_y w_y^g \frac{\mathbb{P}(Y=y, X_{inv}=s, X_{sp}=b, g)}{\mathbb{P}(X_{inv}=s, X_{sp}=b, g)} \log(f(x(b,s))_y) \\ &= \sum_{b,s} \frac{1}{\mathbb{P}(g)} \sum_y w_y^g \mathbb{P}(X_{inv}=s, X_{sp}=b, g|Y=y) p(y) \log(f(x(b,s))_y) \\ &= \sum_{b,s} \frac{1}{\mathbb{P}(g)} \sum_y w_y^g \mathbb{P}(X_{inv}=s|Y=y)\mathbb{P}(X_{sp}=b, g, Y=y) \log(f(x(b,s))_y)\end{aligned}$$

By the definition of g, we have

$$\mathbb{P}(Y = y|X_{sp} = b, g) = a_y \tag{7}$$

So we have

$$\begin{aligned}\tilde{\mathcal{R}}^g(f) &= \sum_{b,s} \frac{1}{\mathbb{P}(g)} \sum_y w_y^g a_y \mathbb{P}(X_{inv} = s|Y = y)\mathbb{P}(X_{sp} = b, g) \log(f(x(b, s))_y) \\ &= \sum_b \mathbb{P}(X_{sp} = b|g) \sum_y \sum_s \mathbb{P}(X_{inv} = s|Y = y) \log(f(x(b, s))_y)\end{aligned}$$

As $\mathbb{P}(X_{sp} = b|g)$ is uniform, we have $\tilde{\mathcal{R}}^g(f) \propto \sum_y \sum_{b\in g,s} \mathbb{P}(X_{inv} = s|y) \log(f_y(x(b, s)))$. As a result

$$\mathcal{L}_{ce}(f) = \sum_g C_g p(g) \sum_y \sum_{x\in g} \mathbb{P}(x_{inv}|Y = y) \log(f(x)_y) \tag{8}$$

where C_g is a constant depend on g. This equation indicates that the loss on $f(\cdot)_y$ only depends on g and $\mathbb{P}(X_{inv} = s|Y = y)$. By imposing invariance constraints on $\mathbb{P}(Y|f, g)$, by Lemma D.2, we have SFC is satisfied, which ends the proof. □