Appendix: When Does Group Invariant Learning Survive Spurious Correlations?

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 [12] or Kruskal-Wallis Test [9].

In our main experiments on PC-MNIST and MNLI, we set the threshold for T_B to 10. It can be seen in Table 1 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 1.

B Experimental Details

B.1 Model Selection

It has been argued that model selection is at the heart of domain generalization [7]. 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 [22] or oracle validation using test data [6] 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 [7]. 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) [7] 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 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.

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

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

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 [4]. 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 [3]. 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 [5; 22]. In our experiments, we follow the practice to utilize MNLI [25] as the training data and HANS (Heuristic Analysis for NLI Systems) [17] 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 [17]. 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

Table 1: Robustness study on PC-MNIST. It shows the performance of SCILL-IRM with threshold $5, 10, 15, 20$ 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 ± 0.56	50.64 ± 0.56	89.95 ± 0.45	54.53 ± 0.60	-	-
SCILL-thr-20	6	83.15 ± 0.47	60.14 ± 1.12	73.37 ± 0.65	**67.95** ± 0.66	72.59 ± 0.33	**67.79** ± 0.57
SCILL-thr-15	7	82.84 ± 0.61	59.79 ± 1.00	73.07 ± 0.69	**68.17** ± 0.56	72.31 ± 0.32	**67.87** ± 0.37
SCILL-thr-10	9	79.65 ± 0.76	**62.49** ± 0.55	71.54 ± 0.35	67.46 ± 0.19	71.54 ± 0.35	67.46 ± 0.19
SCILL-thr-5	15	76.91 ± 0.60	55.50 ± 1.78	66.29 ± 13.1	58.81 ± 2.35	60.29 ± 9.97	61.89 ± 3.96
SCILL-p-0.01	23	79.75 ± 0.32	**62.47** ± 0.93	69.63 ± 0.54	66.78 ± 0.64	72.63 ± 1.21	67.46 ± 0.46

Table 2: Classification accuracy on HANS. Results of methods marked with dagger are cited from [22].

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

set. We train each model with 800 epochs. Following Arjovsky et al. [3], the penalty is applied after 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 [22] 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 [22] 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 [22]. These methods use the same reference model adopted in our experiments,

Table 3: 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 4: 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

but adjust the training objective directly based on its outputs. For example, PoE [5] 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 4. The groups then satisfy the falsity exposure criterion. We construct the ground truth majority/minority split, which violates the falsity exposure, and experiment 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 5, 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 3 show that SCILL$_{uw}$ performs worse than SCILL, demonstrating the importance of the instance reweight step in SCILL.

Table 5: 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$_{gt}$ -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

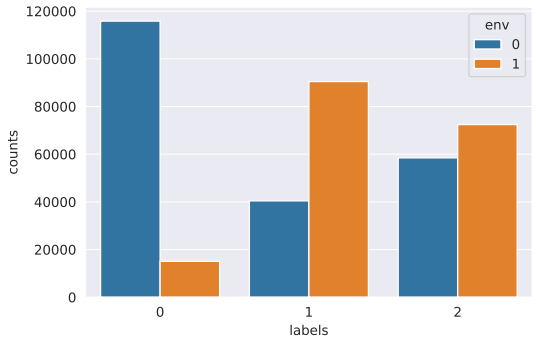


Figure 1: The label proportion varies between the two groups (denoted as 0 and 1 in the figure) inferred by EI on MNLI. The horizontal axis shows 3 labels on MNLI. The vertical axis shows the counts of the instance with the corresponding label in two groups.

B.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 1, 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 1 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.

B.8 Penalties

We experiment with 4 kinds of invariant learning penalties: IRM [3], REx [8], cMMD [10; 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.

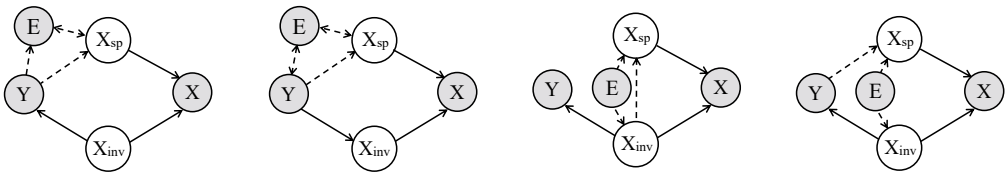
With the same notations, the penalty in V-REx writes as follows.

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

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.

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



(a) [3; 23] (b) [21; 16; 23] (c) [2; 15] (d) [24; 13]

Figure 2: 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).

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 [23].

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

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

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 2 (a), (b). Compared with Figure 1 in the main paper, Figure 2 (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 [23]. Figure 2 (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 [23; 5; 26; 20]. For example, it is required in the proof of conditions of Theorem 4.2 in [23]. The nuisance-varying family defined in [20] 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 2 (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 [16; 20]. 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 [20], and are also discrete as assumed in [16]. Also, their feature invariance penalty is different from that in IL. Specifically, Makar et al. [16] 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. [20] 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 [5; 14; 22; 18; 11; 26], 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 [5; 22] 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

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

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

$$\begin{aligned}\mathbb{P}(Y = y, X_{inv} \in \mathcal{S}^b_{\mathbf{a}}, X_{sp} = b) &= \mathbb{P}(X_{inv} \in \mathcal{S}^b_{\mathbf{a}}, X_{sp} = b | Y = y)\mathbb{P}(Y = y) \\ &= \mathbb{P}(X_{inv} \in \mathcal{S}^b_{\mathbf{a}} | Y = y)\mathbb{P}(X_{sp} = b, Y = y)\end{aligned}$$

As a result,

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

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

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

We have

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

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

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

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 [19] 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 2. 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 [19],

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

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 [3] 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 [6], 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

$$\tilde{\mathcal{R}}^g(f) = \mathbb{E}_{x,y\in g}\, \omega^g(y)\mathcal{L}_{ce}(y, f(X)), \omega^g(y) := \mathbb{P}(Y=y)/\mathbb{P}(Y=y|g) \tag{1}$$

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,

$$\mathbb{P}(X_{inv}=s, X_{sp}=b|Y=y) = \mathbb{P}(X_{inv}=s|Y=y)\mathbb{P}(X_{sp}=b|Y=y) \tag{2}$$

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

$$\mathbb{P}(X_{inv}=s, X_{sp}=b, g|Y=y) = \mathbb{P}(X_{inv}=s|Y=y)\mathbb{P}(X_{sp}=b, g|Y=y) \tag{3}$$

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

$$\mathbb{P}(Y = y|X_{sp} = b, g) = a_y \tag{4}$$

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

$$\mathcal{L}_{ce}(f) = \sum_g C_g p(g) \sum_y \sum_{x\in g} \mathbb{P}(x_{inv}|Y = y) \log(f(x)_y) \tag{5}$$

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. □

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

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

[4] 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`.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.