Fairness-Preserving Regularizer: Balancing Core and Spurious Features

Jiawei Feng [1] **Ancong Wu** [1] **Yuhan Yao** [1] **Wei-Shi Zheng** [1]

Abstract

Real world visual data contains multiple attributes, *e.g.*, color, shape, foreground, background, etc. To solve a specific learning task, machine learning models should use a specific set of attributes. In principle, selecting which set of attributes as the core (non-spurious) ones is determined by the task regardless of how heavily other attributes are (spuriously) correlated with the label. Without prior knowledge for identifying the core attribute or spurious one, we can hardly tell a learned correlation is spurious or not in real-world scenarios. In this work, we dive into this realistic setting and since there is no prior knowledge to determine which feature is core or spurious, we aim to learn a regularized predictor to *fairly balance both core and spurious features*. To achieve this, we start by formalizing fairness of learned features in a linear predictor under multi-view data distribution assumption (Allen-Zhu & Li, 2023). We prove that achieving this fairness can be bounded by a simple regularization term and finally design fairness-preserving regularizer. Experiments on Waterbirds, CelebA and Wilds-FMOW datasets validate the effectiveness of our method.

1. Introduction

Convolutional neural networks (CNNs) can successfully predict ground-truth label under a data distribution similar to the given training set, but heavily fail when the distribution is shifted by some spurious attributes—correlated with the label in the training set but independent in the test set. This unwanted correlation happens in many real-life scenarios, and it has been observed that empirical risk minimization (ERM) trained models could fail to exclude these spurious correlations (Liu et al., 2015a; Sagawa et al., 2020). This phenomenon can be viewed as a consequence of the multi-view nature of visual data structure (Allen-Zhu & Li, 2023; Chen et al., 2023). For instance, a human face image will contain different attributes: gender, ethnicity, nose, hair color, etc. However, without telling a model to focus on a specific attribute, *e.g.* hair color, rather than extract the core feature of hair color, it may inevitably rely on other spurious but still discriminative features that are highly biased towards a particular gender or region, leading to poor performance on the minority groups.

When group information is available, an intuitive solution to mitigate spurious correlation is to learn an invariant representation (Sagawa et al., 2020; Zhou et al., 2021). Motivated by this solution, one line of group-annotation-free works aims to identify each group in a training set and regard them as pseudo group labels (Nam et al., 2020; Liu et al., 2021; Creager et al., 2021; Liu et al., 2023). The key inductive bias of them is the first trained ERM model is biased and can be used to detect spurious correlation, which can completely fail in some cases pointed by (Yong et al., 2022).

As attribute identification can be ill-posed without predefined assumptions, another line of group-annotation-free works focus on learning rich features through different levels of ensemble learning (Wang et al., 2019; Zhang et al., 2022; Asgari et al., 2022). Recently, however, a nontrivial founding in (Izmailov et al., 2022; Ye et al., 2023) is that ERM in fact has the capacity to learn both spurious and core (non-spurious) features. Based on this, recent work (Kirichenko et al., 2023) claims that while both features are learned, the spurious feature can be highly weighted in the final prediction head of the model, leading to poor performance on the minority groups. They utilized an effective two-stage framework that first learned an ERM model and then retrained its prediction head with an additional group-balance dataset.

Compared to the above previous works, we try to mitigate spurious correlation without (1) group information at the stages of training and validation, or (2) predefined assump-

[1] School of Computer Science and Engineering, Sun Yat-sen University, China. Jiawei Feng <fengjw3@mail2.sysu.edu.cn>, Yuhan Yao <yaoyh28@mail2.sysu.edu.cn>, Wei-Shi Zheng <wszheng@ieee.org>. Correspondence to: Ancong Wu <wuanc@mail.sysu.edu.cn>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

Different from the conventional concept of *fairness* wherein the error rates of both majority and minority groups are comparable, we denote it in this paper as a compromise to balance core and spurious features when prior knowledge for identifying them is absent.

tion on what spurious correlation will be, *e.g.* ERM trained models will be biased to spurious features in the first place. These can be unavoidable constraints in realistic applications, for instance, it will be a difficult and time-consuming investigation to discover reliable information of what spurious correlation is when there is a large mount of images and tags from social media for a classification task. Our *key intuition* is that real-world data contain multiple attributes, and in principle, since there is no prior knowledge to identify the core feature from all learned ones, learning all of them including core and spurious ones *in a fair way* can be regarded as a vital trade-off.

Motivated by the above, we first follow the empirical founding that ERM trained model captures both spurious and core features (Kirichenko et al., 2023), but focus on how to learn a fair prediction head without group information or predefined assumption on spurious correlation. To answer that, we first formalize the fairness of learned features in a linear predictor under multi-view data distribution assumption (Allen-Zhu & Li, 2023). Both core and spurious features are defined during the data distribution generation process. After being trained with a gradient-based learning algorithm, the linear predictor can be represented as a linear combination of core feature, spurious feature, and negligible other noise. We finally define a fair predictor as the inner products with core and spurious features being equal. To bring this property into practice, we prove that given an ERM trained feature extractor and the extracted features are under multi-view data distribution assumption, achieving this fairness can be bounded by a simple regularization term without knowing what a core or spurious feature is.

2. Preliminary

We consider supervised visual classification task in the presence of spurious correlation. Given data $\boldsymbol{x} \in \mathcal{X}$ and its corresponding label $y \in \mathcal{Y}$, the goal is to learn a predictor $f : \mathcal{X} \to \mathcal{Y}$. However, for each pair of $(\boldsymbol{x}, y)$ from the training set $\mathcal{D}_{tr}$ with size n, there exists one unknown attribute s, (one for notation simplicity), that is *spuriously correlated* with y, *e.g.*, $(y\!=\!cow, s\!=\!grassland)$ and $(y\!=\!camel, s\!=\!desert)$ would dominate the training set instead of $(y = cow, s = desert)$ or $(y = camel, s = grassland)$. This discrete attribute, which in fact splits $\mathcal{D}_{tr}$ into several groups, could result in a dramatic proportion shift within the test set $\mathcal{D}_{te}$. Specifically, following recent advances (Nam et al., 2020; Liu et al., 2021; Kirichenko et al., 2023), we consider the scenario that each group will be presented *equally* during evaluation, *i.e.*, the attribute will not be correlated with the label in $\mathcal{D}_{te}$.

Following (Izmailov et al., 2022; Ye et al., 2023), we assume that both core and spurious features can be obtained by an ERM model. And we utilize a two-stage training framework where an ERM model is trained in stage one, and the last layer—linear classifier, is retrained from random initialization in stage two with the backbone frozen and without any group information. In the following, we will focus on the second stage.

3. Fairness of Learned Features in a Linear Predictor

In this section, we discuss how to define the fairness of learned features in a linear predictor. Intuitively, we hope that the linear predictor can predict equally well based on spurious or core features. To elaborate this property in a formal and detailed way, we first consider the training data as a variation of multi-view data distribution introduced in (Allen-Zhu & Li, 2023). Under this assumption, core feature and spurious feature can be regarded as different views of data. After training the classifier with a gradient-based learning algorithm, it can be regarded as a linear combination of core feature and spurious feature, and other noise, which can induce a fair way to balance them.

Data distribution We define the training data distribution as a variation of multi-view data distribution. $\boldsymbol{x}$ is composed of P patches $\{\boldsymbol{x}_k | \boldsymbol{x}_k \in \mathbb{R}^d\}_{k=1}^P$, and we represent $\boldsymbol{x}$ as average of patches (for ease of subsequent derivation), $\boldsymbol{x} = \frac{1}{P}\sum_{k=1}^P \boldsymbol{x}_k$. We focus on binary classification problem where $\mathcal{Y} = \{-1, +1\}$, and assume that there are two sets of features, spurious and core ones, associated with the problem. Similar to (Shen et al., 2022; Chen et al., 2023), we define two *orthogonal unit* vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d,\ \boldsymbol{v}_1^T \boldsymbol{v}_2 = 0,\ \|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\|_2 = 1$, as core and spurious features respectively for simplicity of math, such that the core (or spurious) feature of a given class y is $y\boldsymbol{v}_1$ (or $y\boldsymbol{v}_2$). We consider the following data distribution generation mechanism for $\mathcal{D}_{tr}$:

1. Sample $y \in \mathcal{Y}$ uniformly.

2. Given y, each patch $\boldsymbol{x}_k$ of $\boldsymbol{x}$ is defined as

$$\boldsymbol{x}_k = \alpha_k y \boldsymbol{v}_1 + \beta_k y \boldsymbol{v}_2 + \gamma_k \boldsymbol{\xi}_k, \tag{1}$$

   where $\alpha_k, \beta_k, \gamma_k$ are non-negative scalars and $\boldsymbol{\xi}_k \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{I}_d - \boldsymbol{v}_1\boldsymbol{v}_1^T - \boldsymbol{v}_2\boldsymbol{v}_2^T))$. For each k, there is one and only one scalar of the three can be nonzero. The scalars are determined by sampling $\boldsymbol{x}_k$ from a distribution with unknown parameters:

   either a core feature patch, with $\alpha_k > 0, \beta_k = 0, \gamma_k = 0$ while regarding α_k as the magnitude (l^2 norm) of $\boldsymbol{x}_k$,

   or a spurious feature patch, with $\alpha_k = 0, \beta_k > 0, \gamma_k = 0$ while regarding β_k as the magnitude (l^2 norm) of $\boldsymbol{x}_k$,

   or a noise patch, with $\alpha_k = 0, \beta_k = 0, \gamma_k = 1$.

Finally $\boldsymbol{x}$ can be represented as:

$$\boldsymbol{x} = ay\boldsymbol{v}_1 + by\boldsymbol{v}_2 + \boldsymbol{\epsilon}, \tag{2}$$

where $a = \frac{1}{P}\sum_{k=1}^{P}\alpha_k, b = \frac{1}{P}\sum_{k=1}^{P}\beta_k$ can be regarded as l^2 norm of $\boldsymbol{x}$ in the direction of $y\boldsymbol{v}_1$ and $y\boldsymbol{v}_2$ respectively, and $\boldsymbol{\epsilon} = \frac{1}{P}\sum_{k=1}^{P}\gamma_k\boldsymbol{\xi}_k$. We assume that the noise parameters σ are small enough compared to the size of the training set n and number of patch P, such that $\boldsymbol{v}_1, \boldsymbol{v}_2$ will have the relatively major contribution to the training process.

Learning algorithm As $\{\boldsymbol{x}_k | \boldsymbol{x}_k \in \mathbb{R}^d\}_{k=1}^{P}$ can also be viewed as the intermediate output of the previous convolution layer in a CNN (Allen-Zhu & Li, 2023), we simply define a global average pooling layer and a linear classifier $\boldsymbol{W} = [\boldsymbol{w}_{+1}, \boldsymbol{w}_{-1}] \in \mathbb{R}^{d\times|\mathcal{Y}|}$ after the top layer of the CNN backbone to present the model as $F(\boldsymbol{x}) = \boldsymbol{w}_{+1}^T\boldsymbol{x} - \boldsymbol{w}_{-1}^T\boldsymbol{x}$. It is trained by minimizing the following empirical cross-entropy loss:

$$\min_{\boldsymbol{W}} \mathcal{L}_{ce} = \frac{1}{n}\sum_{i=1}^{n} l(y^{(i)} \cdot F(\boldsymbol{x}^{(i)})), \tag{3}$$

where $l(z) = \log(1 + \exp(-z))$. We learn the model using gradient descent starting from Gaussian initialization with learning rate η. At step t, we have

$$\boldsymbol{w}_j^{(t+1)} = \boldsymbol{w}_j^{(t)} - \frac{\eta}{n}\sum_{i=1}^{n} l_i'^{(t)} j y^{(i)}\boldsymbol{x}^{(i)}, \; j \in \mathcal{Y}. \tag{4}$$

Combining Eq. (2), $\boldsymbol{W}$ can be regarded as a linear combination of core feature $\boldsymbol{v}_1$, spurious feature $\boldsymbol{v}_2$, random initialization $\boldsymbol{W}^{(0)}$ and noise of each sample $\boldsymbol{\epsilon}^{(i)}$. Similar to (Cao et al., 2022), it is not difficult to have the following lemma:

Lemma 3.1. *Given the above data distribution and learning algorithm, there exist unique coefficients $\lambda_{j,1}^{(t)}, \lambda_{j,2}^{(t)}$ and $\rho_{j,i}^{(t)}$ such that:*

$$\boldsymbol{w}_j^{(t)} = \boldsymbol{w}_j^{(0)} + \lambda_{j,1}^{(t)} \cdot j\boldsymbol{v}_1 + \lambda_{j,2}^{(t)} \cdot j\boldsymbol{v}_2 + \sum_{i=1}^{n}\rho_{j,i}^{(t)}\boldsymbol{\epsilon}^{(i)}, \tag{5}$$

where $\lambda_{j,1}^{(t)} \approx \langle \boldsymbol{w}_j^{(t)}, j\boldsymbol{v}_1\rangle, \lambda_{j,2}^{(t)} \approx \langle \boldsymbol{w}_j^{(t)}, j\boldsymbol{v}_2\rangle$.

Finally, we can give the formulation of fairness for core and spurious features in linear predictor as follows:

Definition 3.2. Given a linear classifier $\boldsymbol{w}_j$ for class $j \in \mathcal{Y}$ satisfying Lemma 3.1. We say $\boldsymbol{w}_j$ is *fair* if

$$\langle \boldsymbol{w}_j^{(t)}, j\boldsymbol{v}_1\rangle = \langle \boldsymbol{w}_j^{(t)}, j\boldsymbol{v}_2\rangle. \tag{6}$$

4. Fairness-Preserving Regularizer

In Definition 3.2, there are two unknown vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$. To bring fairness of learned features into practice, we first convert it into an optimization problem by introducing a predefined constant $C > 0$:

$$\min_{\boldsymbol{w}_j} |\langle \boldsymbol{w}_j, j\boldsymbol{v}_1\rangle - C| + |\langle \boldsymbol{w}_j, j\boldsymbol{v}_2\rangle - C|, \; j \in \mathcal{Y}. \tag{7}$$

Based on the conditions introduced in Section 3, we give our main result in the following theorem.

Theorem 4.1. *Consider an ERM trained backbone and a linear classifier $\boldsymbol{W} = [\boldsymbol{w}_{-1}, \boldsymbol{w}_{+1}]$ satisfy Lemma 3.1, then minimizing Eq. (7) can be upper-bounded by the following objective given a pair of $(\boldsymbol{x}, y)$ from training set $\mathcal{D}_{tr}$:*

$$\min_{\boldsymbol{w}_y} \frac{1}{P}\sum_{k=1}^{P} |\langle \boldsymbol{w}_y, \boldsymbol{x}_k\rangle - \|\boldsymbol{x}_k\|_2 \cdot C| \tag{8}$$

We give a formal proof in Appendix A. Intuitively, we use each data sample as a bridge to feature disentanglement due to its multi-view data distribution nature. At each patch of data, the feature is purified enough to be a local and stronger fairness regularizer for its classifier, thus achieving a weaker yet global fairness. In practical, both $\langle \boldsymbol{w}_y, \boldsymbol{x}_k\rangle$ and $\|\boldsymbol{x}_k\|_2$ are not difficult to obtain, and we will show how to eliminate the constant C in the following.

How to decide C According to Eq. (7), C can be regarded as the magnitude of the classifier. Since it would have a huge influence on decision boundaries, we replace this hyper-parameter with a max-min normalization process, and finally achieve our fairness-preserving regularizer:

$$l_{fp} = \frac{1}{P}\sum_{k=1}^{P} |\widehat{\langle \boldsymbol{w}_y, \boldsymbol{x}_k\rangle} - \widehat{\|\boldsymbol{x}_k\|_2}|, \tag{9}$$

where $\widehat{\langle \boldsymbol{w}_y, \boldsymbol{x}_k\rangle}$ is max-min normalized version of $\langle \boldsymbol{w}_y, \boldsymbol{x}_k\rangle$ and $\widehat{\|\boldsymbol{x}_k\|_2}$ as well. Note that in Eq. (8), assigning different C for different pair of $(\boldsymbol{x}, y)$ will not change the optimization direction to fairness property. Therefore, it is not difficult to find that Eq. (8) can be upper-bounded by Eq. (9).

Overall training Recall that in the second stage, we only retrain the last layer—linear classifier $\boldsymbol{W}$ with the ERM trained backbone frozen. The classifier is optimized by the following objective:

$$\min_{\boldsymbol{W}} \mathcal{L}_{ce} + \delta\frac{1}{n}\sum_{i=1}^{n} l_{fp}(\boldsymbol{x}^{(i)}, y^{(i)}), \tag{10}$$

where δ is the weighting factor of the fairness-preserving regularization term and is simply set to 1 through all experiments.

Table 1. Comparison to other methods. The number is averaged over 3 independent running seeds and the number in brackets represents standard deviation. “Group” means whether used group information. “WGA” denotes worst group accuracy and “MGA” represents mean group accuracy.

Method	Group	Waterbirds		CelebA		FMOW	
		WGA	MGA	WGA	MGA	WGA	MGA
GDRO	✓	68.5(6.0)	-	66.3(7.8)	-	30.2	-
ERM-DFR	✓	91.1(0.8)	-	89.4(0.9)	-	41.6(0.6)	-
ERM		70.0(2.1)	90.4(0.2)	44.3(1.8)	**95.5**(0.1)	31.8(0.9)	52.4(0.1)
RWY		65.4(0.6)	-	46.1(2.1)	-	30.5(0.6)	-
ERM-FRR		71.2(2.7)	92.1(0.2)	42.2(1.6)	**95.5**(0.1)	31.6(0.6)	52.5(0.2)
ERM-FP		**87.3**(1.8)	**93.7**(1.2)	**47.3**(1.5)	**95.5**(0.1)	**32.4**(0.8)	**52.8**(0.1)

5. Results

In this section, we evaluate our fairness-preserving regularizer on three image classification datasets under spurious correlation compared to upper-bound methods with group supervision and existing methods without prior knowledge for identifying groups. We also illustrate visualization results of our method.

Datasets We consider three image classification under spurious correlation benchmarks. **Waterbirds** (Sagawa et al., 2019) is used for binary image classification of bird types (landbird or waterbird). The background(land or water) is spuriously correlated with the class, because most landbirds are shown on land, and most waterbirds are shown over water. **CelebA hair color** (Liu et al., 2015b) aims at the binary classification of whether a person in an image is blond or not. Gender is a spurious attribute in the dataset, as most images labeled ”blonde” depict females. **WILDS-FMOW** (Christie et al., 2018; Koh et al., 2021; Sagawa et al., 2021) consists of satellite images, and the goal is to classify images according to the type of building or land use. The spurious attribute lies in the region (Asia, Europe, Africa, America, and Oceania). The training data additionally contains another group Other, which is dropped during evaluation. Following the WILDS benchmark for this dataset, the groups are defined by just the value of the spurious attribute. In particular, the worst group accuracy corresponds to the worst accuracy across regions. Images of these regions are unevenly represented in the data, resulting in uneven representation. In addition, the test images were taken several years later than the train images, so there is an additional type of domain shift.

The Waterbirds and CelebA datasets are commonly used to benchmark the performance of group robustness methods, while the FMOW dataset presents challenging real-world problems with spurious correlations. In the above datasets, the inputs do not resemble natural images from data sets such as ImageNet (Russakovsky et al., 2015), so models cannot simply rely on feature transfer, but must learn the relevant features from data to obtain good performance.

Implementation Details According to prior works (Sagawa et al., 2019; Idrissi et al., 2022), we use a ResNet-50 (He et al., 2016) pre-trained on ImageNet-1K (Russakovsky et al., 2015) on three benchmark datasets. In the first stage (ERM training), we set the number of epochs to 100 for Waterbirds and 20 for both CelebA and FMOW. In the second stage, we re-initialize the final linear classifier and set the number of epochs to 10% of the first stage.

Methods We consider five methods for learning the features. GDRO (Sagawa et al., 2019) is a state-of-the-art method that uses the group information on the training data to minimize the worst group loss instead of the average loss, which is often considered an oracle method or upper-bound on the worst group performance under spurious correlations (Liu et al., 2021; Creager et al., 2021). ERM-DFR (Kirichenko et al., 2022) is based on the ERM method, simply retraining the last layer of the model on a small held-out dataset where the spurious correlation does not hold. ERM (Empirical Risk Minimization) is the standard training on the original training data, without any techniques targeted at improving the worst group performance. RWY reweights the loss in each class according to the size of the class (Idrissi et al., 2022). One similar work to ours is FRR (Addepalli et al., 2023). They adopted two-stage training and regularized the classifier by reconstructing the final output feature assuming that simplicity bias could make neural network brittle. We do not hold such assumption.

Comparison to other methods Comparison results to other related methods are illustrated in Table 1. Compared with the method without group information, our method performs better than the one-stage method of RWY on all three datasets. Also as a two-stage method, ERM-FRR has a gap of 16.1% on WGA and a gap of 1.6% on MGA from our method on Waterbirds. While on CelebA, WGA is 5.1% lower than our method. On FMOW, our method outperforms ERM-FRR by 0.8% on WGA.

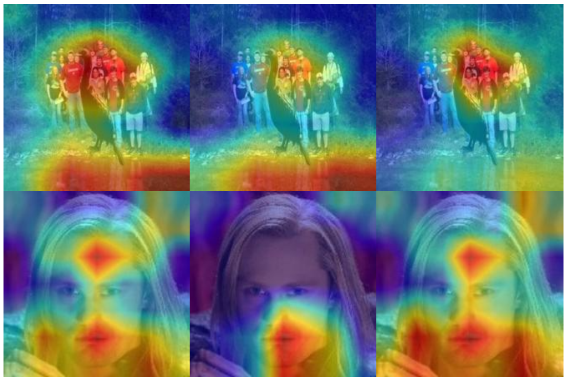

Figure 1. Visualization results. The first (second) line are from Waterbirds (CelebA).

Visualization For an image-label pair, we can easily visualize both $\widehat{\|\boldsymbol{x}_k\|_2}$ and $\widehat{\langle \boldsymbol{w}_y, \boldsymbol{x}_k \rangle}$ by up-sampling them to the image size. We illustrate $\widehat{\|\boldsymbol{x}_k\|_2}$ right after ERM training, $\widehat{\langle \boldsymbol{w}_y, \boldsymbol{x}_k \rangle}$ before the second stage retraining, and $\widehat{\langle \boldsymbol{w}_y, \boldsymbol{x}_k \rangle}$ after retraining, as Figure 1. From left to right, each column represent $\widehat{\|\boldsymbol{x}_k\|_2}$ after training, $\widehat{\langle \boldsymbol{w}_y, \boldsymbol{x}_k \rangle}$ before retraining and $\widehat{\langle \boldsymbol{w}_y, \boldsymbol{x}_k \rangle}$ after retraining. It is shown that our regularization term can make the classifier focus on activated features relatively fairly, rather than focus some of them and ignore the other.

6. Conclusion

In this work, we discussed how to achieve a compromise to balance core and spurious features when there is no prior knowledge for identifying each of them. We aimed to learn a regularized predictor to fairly balance both core and spurious features. We designed fairness-preserving regularizer by proving that achieving fairness of learned features in a predictor can be bounded by this regularizer.

References

Addepalli, S., Nasery, A., Radhakrishnan, V. B., Netrapalli, P., and Jain, P. Feature reconstruction from outputs can mitigate simplicity bias in neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zH9GcZ3ZGXu.

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Uuf2q9TfXGA.

Asgari, S., Khani, A., Khani, F., Gholami, A., Tran, L., Mahdavi-Amiri, A., and Hamarneh, G. Masktune: Mitigating spurious correlations by forcing to explore. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=hMGSz9PNQes.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=pF8btdPVTL_.

Chen, Y., Huang, W., Zhou, K., Bian, Y., Han, B., and Cheng, J. Towards understanding feature learning in out-of-distribution generalization. *arXiv preprint arXiv:2304.11327*, 2023.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. On feature learning in the presence of spurious correlations. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=wKhUPzqVap6.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Liu, S., Zhang, X., Sekhar, N., Wu, Y., Singhal, P., and Fernandez-Granda, C. Avoiding spurious correlations via logit correction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5BaqCFVh5qL.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015a.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015b.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Sagawa, S., Koh, P. W., Lee, T., Gao, I., Xie, S. M., Shen, K., Kumar, A., Hu, W., Yasunaga, M., Marklund, H., et al. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, 2021.

Shen, R., Bubeck, S., and Gunasekar, S. Data augmentation as feature manipulation. In *International Conference on Machine Learning*, pp. 19773–19808. PMLR, 2022.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3eefceb8087e964f89c2d59e8a249915-Paper.pdf.

Ye, H., Zou, J., and Zhang, L. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8968–8990. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/ye23a.html.

Yong, L., Zhu, S., Tan, L., and Cui, P. ZIN: When and how to learn invariance without environment partition? In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=pUPFRSxfACD.

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

Zhou, C., Ma, X., Michel, P., and Neubig, G. Examining and combating spurious features under distribution shift. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12857–12867. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhou21g.html.

A. Proof of Theorem 4.1

Proof. We start by minimizing the following equation, which is a combination of two absolute value terms:

$$\min_{\boldsymbol{w}_j} |\langle \boldsymbol{w}_j, j\boldsymbol{v}_1\rangle - C| + |\langle \boldsymbol{w}_j, j\boldsymbol{v}_2\rangle - C|. \tag{11}$$

Consider a data point $(\boldsymbol{x}, y)$ from training set $\mathcal{D}_{tr}$ and its corresponding classifier weight $\boldsymbol{w}_y$, multiplying the two terms by a, b respectively does not change the optima:

$$\min_{\boldsymbol{w}_y} |a \cdot \langle \boldsymbol{w}_y, y\boldsymbol{v}_1\rangle - a \cdot C| + |b \cdot \langle \boldsymbol{w}_y, y\boldsymbol{v}_2\rangle - b \cdot C|. \tag{12}$$

Using Eq.(2), we have:

$$\begin{aligned}
&\min_{\boldsymbol{w}_y} |\frac{1}{P}\sum_{k=1}^{P}\alpha_k \cdot \langle \boldsymbol{w}_y, y\boldsymbol{v}_1\rangle - \frac{1}{P}\sum_{k=1}^{P}\alpha_k \cdot C| + |\frac{1}{P}\sum_{k=1}^{P}\beta_k \cdot \langle \boldsymbol{w}_y, y\boldsymbol{v}_2\rangle - \frac{1}{P}\sum_{k=1}^{P}\beta_k \cdot C|. \\
=&|\frac{1}{P}\sum_{k=1}^{P}\langle \boldsymbol{w}_y, \alpha_k y\boldsymbol{v}_1\rangle - \alpha_k \cdot C| + |\frac{1}{P}\sum_{k=1}^{P}\langle \boldsymbol{w}_y, \beta_k y\boldsymbol{v}_2\rangle - \beta_k \cdot C| \\
\leq&\frac{1}{P}\sum_{k=1}^{P}|\langle \boldsymbol{w}_y, \alpha_k y\boldsymbol{v}_1\rangle - \alpha_k \cdot C| + |\langle \boldsymbol{w}_y, \beta_k y\boldsymbol{v}_2\rangle - \beta_k \cdot C| \\
\leq&\frac{1}{P}\sum_{k=1}^{P}|\langle \boldsymbol{w}_y, \alpha_k y\boldsymbol{v}_1\rangle - \alpha_k \cdot C| + |\langle \boldsymbol{w}_y, \beta_k y\boldsymbol{v}_2\rangle - \beta_k \cdot C| + |\langle \boldsymbol{w}_y, \gamma_k\boldsymbol{\xi}_k\rangle - \|\gamma_k\boldsymbol{\xi}_k\|_2 \cdot C| \\
=&\frac{1}{P}\sum_{k=1}^{P}|\langle \boldsymbol{w}_y, \boldsymbol{x}_k\rangle - \|\boldsymbol{x}_k\|_2 \cdot C|,
\end{aligned} \tag{13}$$

while the first inequality holds for triangle inequality, and the last equation holds for Eq. (1). Therefore, minimizing Eq. (7) can be upper-bounded by Eq. (8).

□