On the Effect of Key Factors in Spurious Correlation: A Theoretical Perspective

Yipei Wang **Xiaoqian Wang**
Elmore Family School of Electrical and Computer Engineering
Purdue University

Abstract

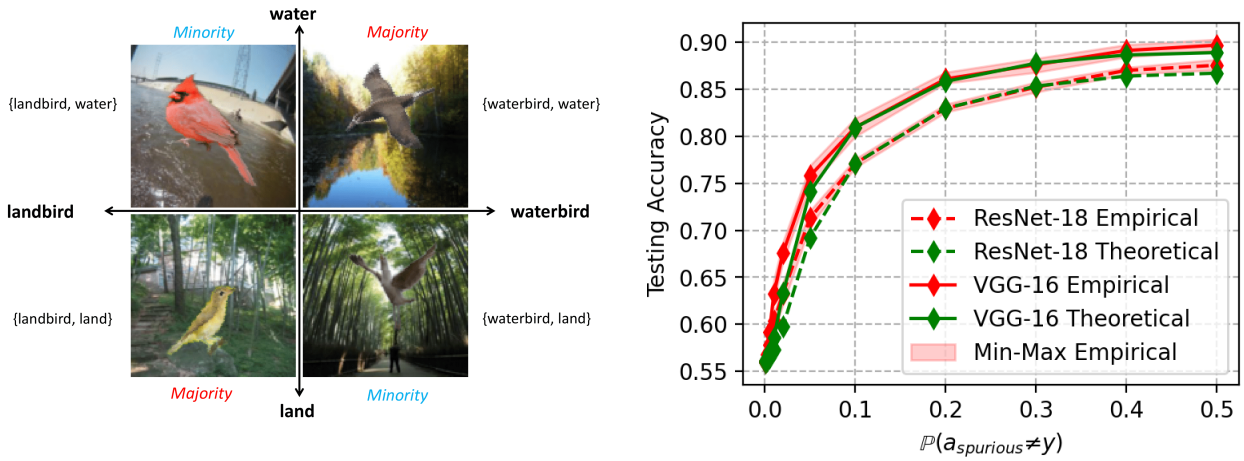
Spurious correlations arise when irrelevant patterns in input data are mistakenly associated with labels, compromising the generalizability of machine learning models. While these models may be confident during the training stage, they often falter in real-world testing scenarios due to the shift of these misleading correlations. Current solutions to this problem typically involve altering the correlations or regularizing latent representations. However, while these methods show promise in experiments, a rigorous theoretical understanding of their effectiveness and the underlying factors of spurious correlations is lacking. In this work, we provide a comprehensive theoretical analysis, supported by empirical evidence, to understand the intricacies of spurious correlations. Drawing on our proposed theorems, we investigate the behaviors of classifiers when confronted with spurious features, and present our findings on how various factors influence these correlations and their impact on model performances, including the Mahalanobis distance of groups, and training/testing spurious correlation ratios. Additionally, by aligning empirical outcomes with our theoretical discoveries, we highlight the feasibility of assessing the degree of separability of intertwined real-world features. This research paves the way for a nuanced comprehension of spurious correlations, laying a solid theoretical groundwork that promises to steer future endeavors toward crafting more potent mitigation techniques.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

With the rapidly increasing computational power, modern machine learning studies have achieved great success. Complex deep models, benefitting from this surge, are capable of extracting refined and structural features from complicated training data (Bengio et al., 2017). Their power has led to massive deployment across a spectrum of applications, from autonomous vehicles (Janai et al., 2020) to critical medical imaging analysis (Litjens et al., 2017; Saab et al., 2022), etc. Given the wide applications, the reliability of these models becomes paramount, especially in these high-stakes areas. However, when trained with limited data, the performance of these deep models is significantly influenced by the distribution shift between the training and testing distributions (Bishop and Nasrabadi, 2006; LeCun et al., 2015; Wang and Deng, 2018; Zhou et al., 2022; Wang et al., 2022).

While distribution shifts present various challenges, one particularly pervasive issue is the spurious correlation. It arises when there exists correlation between non-causal features with causal features. A model trained on such features is likely to rely on these non-causal features in the data that seem to correlate with the desired output. For example, in image recognition tasks, it is not uncommon for the models be deceived by the correlation between the background and the primary subject (Beery et al., 2018). If a dataset consistently shows waterbirds with water as the background, the model will start to associate the water background with waterbirds. Beyond object recognition, natural language processing (NLP) models can be misled by nuances like negations and vagueness (Williams et al., 2017; Gururangan et al., 2018). In facial recognition, there is a risk of the model inadvertently focusing on demographic features rather than the intended characteristics (Liu et al., 2015a, 2021; Ming et al., 2022). Such dependencies not only compromise accuracy in the testing scenario, but also raise great concerns regarding safety and fairness in real-world applications (Wang et al., 2022; Van de Poel and Royakkers, 2023).



(a) Waterbird dataset

(b) Testing Accuracy

Figure 1: The demonstration of the (a) waterbird dataset and (b) the empirical results and our theoretical results on the waterbird data with respect to the change of the correlation ratio across different models.

In the context of studies of spurious correlation, there are two types of features. An **invariant feature** is assumed to be perfectly correlated with the ground truth labeling, and the models are expected to rely on these features for accurate predictions. Conversely, a **spurious feature** is correlated with the label (i.e. the invariant feature) in the training data but independent from it in the testing scenarios. To illustrate, consider the waterbird dataset (Sagawa et al., 2019). As shown in fig. 1 (a), here the invariant feature determines whether the bird is a waterbird or a landbird, while the spurious feature refers to whether the background is water or land. Assuming that the only distribution shift from the training scenario to the testing scenario is the correlation shift, the decline in performance can be attributed to $\mathbb{P}(\texttt{spurious feature!=label}) = 1 - \mathbb{P}(\texttt{spurious feature==label})$, which is the **spurious correlation ratio** in the training data. In essence, a training dataset with a spurious correlation ratio distant from 0.5 is likely to result in poorer performance when tested on unbiased data.

Spurious correlations introduce two factors into the training data that affect the models, from both the statistical and the geometric aspects (Nagarajan et al., 2020). Naturally, to mitigate the issue caused by the spurious correlation, existing methods can be roughly aligned with the two aspects. From a statistical point of view, the spurious correlation can be mitigated by altering the correlation ratio. Since the ratio is directly determined by the number of samples in each group, the intuitive and straightforward way is to change *correlation ratio* by changing the number of data samples from each group. As an alternative, various methods are proposed to mitigate this issue by reweighting/resampling schemes (such as concentrating more on the worst-group data) in training (Buda et al., 2018; Duchi et al., 2019; Sagawa et al., 2019; Nam et al., 2020; Liu et al., 2021). On the other hand, in the geometric aspect, recent studies also reveal that the separability of the invariant feature and the spurious feature also significantly affect the robustness of the trained model (Geirhos et al., 2020; Shah et al., 2020; Shi et al., 2022). However, since *separability* of features is an abstract concept compared to the correlation ratio, this aspect is paid less attention to in the study of spurious correlation problems or related algorithms.

Although the two categories of methods are intuitive and direct, the rigorous analyses on how the two aspects *quantitatively* affect the data distribution and the trained models are lacking. For example, it is already well-known that the testing scenario can achieve improved performance when the spurious correlation decreases since the training and testing distributions become more similar (Bishop and Nasrabadi, 2006; LeCun et al., 2015). However, it remains unknown to what extent can the change in the correlation ratio affect the model's performance over unbiased data quantitatively. As shown in fig. 1 (b), across different models, the testing performance changes consistently w.r.t. the change of the spurious correlation ratio $\mathbb{P}(\texttt{spurious feature!=label})$, following a similar trend. It can be observed that the performance first has a boosting increase, and then reaches the plateau very quickly. This nonlinearity suggests that the initial value of the correlation ratio can also significantly impact the extent of performance improvement after the ratio is altered. Although noticed before as the statistical skews (Nagarajan et al., 2020), the studies on the

influence of the ratio rely on the training steps, and do not provide comprehensive understandings. Instead, we derive analytical studies through the influence of the correlation ratio on the Bayesian optimal classifier and propose tight bounds of the ratio dynamics. The results hold even for DNNs and real datasets such as waterbird. Our theoretical results (green curves) can estimate the empirically trend (red curves) precisely.

As for the geometric properties of the spurious correlation, Nagarajan et al. (2020) use the max-margin classifier to bound the scalar weight of the spurious feature. In this work, we delve deeper into this aspect, and propose analytical form to characterize the reliance of the Bayesian optimal classifier on the spurious feature. The results also theoretically inspire a simple regularization term that changes the *separability* of the penultimate-layer embedding of the data. This essentially corresponds to the feature alignment in the domain generalization problem. Unlike the reweighting schemes, it does not require any manipulation to the correlation ratio, but can still improve the group robustness of the models and is comparable to the state-of-the-art algorithm Group-DRO (Sagawa et al., 2019). This is a verification of the importance of the overlooked factors in spurious correlations. The implementation of the experiments is open source in GitHub. Our contribution can be summarized as follows:

- We theoretically study the two factors spurious, correlation ratio and feature separability, that affect the model performance.
- From our theorem, we can *quantitatively* measure the effect of the two factors in model performance. The theorems are also applicable to generalized real-world applications.
- In terms of spurious correlation ratio, we propose an approach to quantitatively estimate model performance w.r.t. the change of ratio on real data.
- In terms of feature separability, we verify the results with a regularization technique that does not require any reweighting schemes but can still improve the group robustness of the models to the level of the state-of-the-art (SOTA) Group-DRO.

The structure of this paper is organized as follows. Related work is reviewed in section 2. We set up assumptions and present theoretical results in section 3. All proofs can be found in the appendix. Then we introduce for the experiments in section 4. Experimental results are presented in section 5. The significance and limitations are discussed in section 6.

2 RELATED WORK

While we carry out theoretical studies that quantify the factors of spurious features: the correlation ratio and the feature properties, there are other works that offer insightful results of the two factors.

The Correlation Ratio. As the straightforward factor of spurious correlations, most of the methods are essentially focused on this. Resampling methods either oversample the minor groups or undersample the major groups (Japkowicz and Stephen, 2002; He and Garcia, 2009; Buda et al., 2018; Byrd and Lipton, 2019). Reweighting methods, on the other hand, assign different weights to samples of certain properties (Duchi et al., 2019; Sagawa et al., 2019; Nam et al., 2020; Liu et al., 2021; Ahmed et al., 2021; Zhou et al., 2021). The effect of the correlation ratio is also studied through controllable ratio (Sagawa et al., 2020; Kirichenko et al., 2022; Zhang et al., 2022).

Feature Properties. Different from the work that focuses on the correlation ratio, the feature properties are much less studied in the spurious correlation problems. Cao et al. (2019) point out that not only changing the correlation ratio, but also regularization over the model can lead to performance improvement. Shah et al. (2020) show that DNNs have a tendency to rely on extremely simple features. Sagawa et al. (2020) identify the signal-to-noise ratio as one of the major properties of spurious correlation. Shi et al. (2022) argue that the feature representations of unsupervised learning can outperform supervised models in an extreme correlation ratio. Kirichenko et al. (2022) introduce the concept of "core features", which are always learned even by the models that underperform minority groups. Ming et al. (2022) assume the invariant and spurious features are linearly combined as the input representation and develop the optimal classifier theoretically. This is also similar to the alignment of latent features in domain generalization problems (Muandet et al., 2013; Erfani et al., 2016; Ghifary et al., 2016; Hu et al., 2020; Jin et al., 2020). However, this aspect is not sufficiently explored in handling spurious correlation problems.

3 ANALYSIS ON TWO FACTORS

3.1 Problem Setup

Let $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$ denote the dataset, where $\mathcal{X} \subset \mathbb{R}^d$ is determined by the combination of N independent high-level feature representation $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N) \in \mathcal{Z} \subseteq \mathbb{R}^p$ through the bijection $\Phi : \mathcal{Z} \to \mathcal{X}$. The dimension of N features are $p_n|_{n=1}^N$ respectively, s.t. $\sum_{n=1}^N p_n = p$. The feature attributions are defined as

$\boldsymbol{a} \in \mathcal{A} \subseteq \{-1, 1\}^N$. Then the mean of each group is defined by $\boldsymbol{\mu}_{\boldsymbol{a}} = [a_1\boldsymbol{\mu}_1^T, \cdots, a_N\boldsymbol{\mu}_N^T]^T \in \mathbb{R}^p$. For example, when $N = 3, \boldsymbol{a} = [1, 1, -1]$, then $\boldsymbol{\mu}_{\boldsymbol{a}} = [\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, -\boldsymbol{\mu}_3^T]$, indicating that the 1st and 2nd features positive while the third one is negative. Let $\boldsymbol{t}_{\boldsymbol{a}} \in \{0, 1\}^N$ be the indicator s.t. $t_{\boldsymbol{a},n} = \mathbf{1}_{\{a_n=1\}}$. We have $y \equiv a_1$ since the invariant feature is consistent with the label. Conventionally, we assume that in the latent representation $\boldsymbol{z}$, the features are Gaussian mixtures and orthogonal (Nagarajan et al., 2020; Sagawa et al., 2020; Yao et al., 2022; Idrissi et al., 2022; Ming et al., 2022):

$$y \sim \text{Uniform}\{-1, 1\} \tag{1}$$

$$\begin{aligned} \boldsymbol{z}_n|y \sim &\mathbb{P}(a_n = y)\mathcal{N}(y \cdot \boldsymbol{\mu}_n|\Sigma_n) + \\ &\mathbb{P}(a_n \neq y)\mathcal{N}(-y \cdot \boldsymbol{\mu}_n|\Sigma_n) \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_n, \Sigma_n$ are the mean and covariance terms of corresponding features, such that $\|\boldsymbol{\mu}_n\| > 0, \Sigma_n \succ 0$. We denote by $\Sigma = \text{diag}(\Sigma_1, \cdots, \Sigma_N) \in$ the covariance matrix of the entire latent representation, and naturally $\Sigma \succ 0$. Besides, let $\alpha_n = \mathbb{P}(a_n = y|y = 1) = \mathbb{P}(a_n = 1|y = 1)$, $\beta_n = \mathbb{P}(a_n = y|y = -1) = \mathbb{P}(a_n = -1|y = -1)$ denote the correlation ratios in the positive and negative classes. WLOG, we assume that the first feature is the invariant feature, and thus $\alpha_1 = \beta_1 = 1$. For example, in the waterbird dataset demonstrated in fig. 1(a), $N = 2$ and the invariant feature $\boldsymbol{z}_1$ (that is, $\boldsymbol{z}_{\text{inv}}$) refers to the taxa of birds as in {`waterbird, landbird`}, while $\boldsymbol{z}_2$ (i.e. $\boldsymbol{z}_{\text{spur}}$) refers to the habitats shown in the background as in {`water, land`}.

According to the symmetry, WLOG, we assume that $\alpha_n, \beta_n \geq 0.5$, $\forall n \leq N$. $\alpha_n, \beta_n \to 1$ indicates a high correlation between y and $\boldsymbol{z}_n$, while $\alpha_n, \beta_n \to 0.5$ indicates a low correlation. Over the latent space $\mathcal{Z} \times \mathcal{Y}$, We consider linear classifiers defined by $\hat{y} = \boldsymbol{w}^T\boldsymbol{z}$.

3.2 Bounding the Performance Shift

As shown in fig. 1 (b), there is a potential consistent trend of model performance with respect to the correlation ratio $\mathbb{P}(a_n = y)$. When the correlation ratio changes, both the amount of the ratio shift and the original ratio have significant influence to the performance. Therefore, we study this universal trend analytically for how it is affected, starting by determining the Bayesian optimal classifier. We first have

Lemma 1. ***(Accuracy)*** *Given the parameters* $\Theta = (\boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$*, the accuracy of the classifier* $\boldsymbol{w}$ *is*

$$A(\boldsymbol{w}; \Theta) = \frac{1}{2}\Big(1 + \sum_{\boldsymbol{a} \in \{\pm 1\}^N} \gamma_{\boldsymbol{a}} \text{erf}\big(\frac{\boldsymbol{\mu}_{\boldsymbol{a}}^T\boldsymbol{w}}{\sqrt{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}}\big)\Big) \tag{3}$$

where $\gamma_{\boldsymbol{a}} = \big(\prod_{n=1}^N [\alpha_n^{t_{\boldsymbol{a},n}}(1-\alpha_n)^{(1-t_{\boldsymbol{a},n})}] + \prod_{n=1}^N [\beta_n^{t_{\boldsymbol{a},n}}(1-\beta_n)^{(1-t_{\boldsymbol{a},n})}]\big)/2$

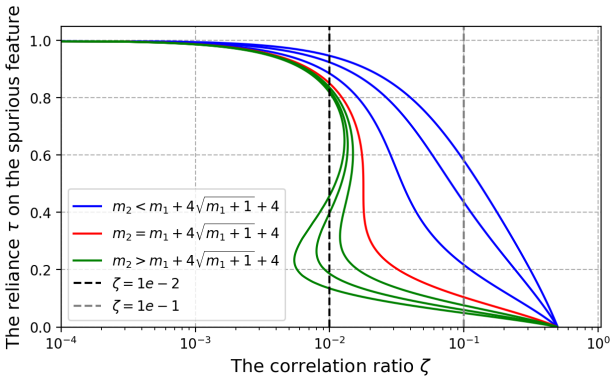


Figure 2: The relation between the reliance τ on the spurious feature and the correlation ratio ζ. Note that this is symmetric w.r.t. $\zeta = 0.5$, where $\tau = 0$.

The proof is presented in appendix A. This demonstrates how the accuracy of the classifier $\boldsymbol{w}$ is affected by the data distributions Θ, through (i) the spurious correlation ratios determined by $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and (ii) the separability through $\boldsymbol{\mu}, \Sigma$. Let $\boldsymbol{w}^* = \arg\max_{\boldsymbol{w} \in \mathbb{R}^p} A(\boldsymbol{w}; \Theta)$ denote the Bayesian optimal classifier w.r.t. the training data. We then prove in appendix B that it satisfies

Lemma 2. *The Bayesian optimal classifier* $\boldsymbol{w}* = [(\boldsymbol{w}_1^*)^T, \cdots, (\boldsymbol{w}_N^*)^T]^T$ *given* $\Theta = (\boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ *can be written as* $\boldsymbol{w}_i^* = \eta c_i \Sigma_i^{-1}\boldsymbol{\mu}_i, i = 1, \cdots, N$*, such that*

$$c_n = \sum_{\substack{\boldsymbol{a} \in \{\pm 1\}^N \\ a_n = 1}} (\gamma_{\boldsymbol{a}} - \gamma_{-\boldsymbol{a}}) \exp\Big(-\frac{(\sum_{n=1}^N t_{\boldsymbol{a},n} c_n m_n)^2}{2\sum_{n=1}^N c_n^2 m_n}\Big)$$

where $\eta > 0$ *is a positive constant that determines the norm of the classifier. And here* $m_n = \boldsymbol{\mu}_n^T\Sigma^{-1}\boldsymbol{\mu}_n > 0$ *is the Mahalanobis distance of the n-th feature.*

The Mahalanobis distance m_n defines the separability of the groups, and is independent of the spurious correlation ratio. Larger m_n indicates higher separability of the feature $\boldsymbol{z}_n$. Furthermore, the spurious correlations $\boldsymbol{z}_{2,\cdots,N}$ are independent, we thus focus on the scenario where there's one spurious feature ($N = 2$). The analysis generalizes to other spurious features automatically. When $N = 2$, $\alpha_1 = \beta_1 = 1, (1 - \alpha_1) = (1 - \beta_1) = 0$. Thus $\gamma_{\boldsymbol{a}}$ is simplified to $\gamma_{(+1,+1)} = (1 \cdot \alpha_2 + 1 \cdot \beta_2)/2 =: \zeta$, $\gamma_{(+1,-1)} = (1 \cdot (1 - \alpha_2) + 1 \cdot (1 - \beta_2))/2 = 1 - \zeta$, $\gamma_{(-1,-1)} = \gamma_{(-1,+1)} = (0 + 0)/2 = 0$. In the 4 groups, we omit "1" and denote attributions by $++, +-, -+, --$ for simplications. Then $\gamma_{\boldsymbol{a}}$ is simplified as $\gamma_{--} = \gamma_{-+} = 0, \gamma_{+-} = 1 - \zeta, \gamma_{++} = \zeta$. where $\zeta = \frac{\alpha_{\text{spur}} + \beta_{\text{spur}}}{2}$ is the average **correlation ratio** of the entire dataset. We can then prove that

Corollary 2.1. *Given* $\boldsymbol{m}, \mathcal{Z}, \mathcal{Y}, \zeta$*, let* $\eta = 1/c_{\text{inv}}, \tau = c_{\text{spur}}\eta = c_{\text{spur}}/c_{\text{inv}}$*, then the optimal classifier is denoted by* $\boldsymbol{w}_{\text{inv}}^* = \Sigma_{\text{inv}}^{-1}\boldsymbol{\mu}_{\text{inv}}, \boldsymbol{w}_{\text{spur}}^* = \tau\Sigma_{\text{spur}}^{-1}\boldsymbol{\mu}_{\text{spur}}$.

From lemma 2, the extent to which the optimal classifier relies on the spurious feature can be quantified by $\|\boldsymbol{w}^*_{\text{spur}}\|/\|\boldsymbol{w}^*_{\text{inv}}\| = \tau\|\Sigma^{-1}_{\text{spur}}\boldsymbol{\mu}_{\text{spur}}\|/\|\Sigma^{-1}_{\text{inv}}\boldsymbol{\mu}_{\text{inv}}\| = \tau \cdot \texttt{const}$. Therefore, it is desired that τ to be small. In fact, the relation between the optimal classifier's reliance on the spurious feature and the spurious correlation can be quantified by the following Corollary.

Corollary 2.2. *(informal) (i) When $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}}+1} + 4$, $\exists g \in \mathcal{C}^1$ s.t. $\tau = g(\zeta)$, and g is monotonically increasing. (ii) When $m_{\text{spur}} \geq m_{\text{inv}} + 4\sqrt{m_{\text{inv}}+1} + 4$, τ is monotonically increasing w.r.t. ζ when ζ is not very close to 1 or 0. There can be multiple optimal classifiers when ζ is close to 1 or 0.*

As visualized in fig. 2, the relation between the optimal classifier's reliance on the spurious feature and the ratio ζ of the data is affected by the threshold (red curve). When the spurious correlation is not severe, there is a unique optimal classifier given $m_{\text{inv}}, m_{\text{spur}}, \zeta$. However, as $\zeta \approx 0, 1$, the classifier becomes unstable, resulting in multiple optimal classifiers. This is because when the spurious correlation is very strong, and yet the spurious feature is very easy to separate (e.g. the presence of watermark, different colors, etc.), the classifier can either rely on the invariant feature or the spurious feature to achieve equivalent optimality in the training phase. Such instability suggests that it is important to study the influence of correlation change with different original ratio ζ.

With the optimal classifier $\boldsymbol{w}^*$ available, it's performance can be obtained by substituting it back to lemma 1. Besides, the performance on the testing data where the original spurious correlation does not hold can be obtained through $\boldsymbol{w}^*$. To avoid ambiguity, we use $\zeta_{\text{tr}}, \zeta_{\text{te}}$ to denote the correlation ratio of the training/testing data. From eq. (9), we thus have the following expression of the classifier's performance.

Lemma 3. ***(Optimal Accuracy.)*** *Let $\zeta_{\text{tr}}, \zeta_{\text{te}}$ be the correlation ratios in the training and testing data. And let the $m_{\text{inv}}, m_{\text{spur}}$ be the Mahalanobis distance of the invariant and spurious features and satisfy $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}}+1} + 4$. Then the training and testing accuracy of the optimal classifier can be written as:*

$$\begin{aligned} A(\zeta_{\text{tr}}) =& \frac{1}{2}\big[1 + \zeta_{\text{tr}} R\big(g(\zeta_{\text{tr}})\big) + r\big(g(\zeta_{\text{tr}})\big)\big] \\ A(\zeta_{\text{te}}; \zeta_{\text{tr}}) =& \frac{1}{2}\big[1 + \zeta_{\text{te}} R\big(g(\zeta_{\text{tr}})\big) + r\big(g(\zeta_{\text{tr}})\big)\big] \end{aligned} \tag{4}$$

$$\text{where} \quad \begin{cases} R(\tau) = \text{erf}\Big(\dfrac{m_{\text{inv}} + \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\Big) \\ \qquad\quad - \text{erf}\Big(\dfrac{m_{\text{inv}} - \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\Big) \\ r(y) = \text{erf}\Big(\dfrac{m_{\text{inv}} - \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\Big) \end{cases}$$

From the formulae of the training and testing accuracy, it can be noticed that under certain constraints of $m_{\text{inv}}, m_{\text{spur}}$, the testing accuracy can be written as a continuously differentiable function of ζ_{tr}. Therefore, with the goal of quantifying how the shift of spurious correlation ratio (ζ_{tr}) influences the model performance, we propose the following theorem.

Theorem 4. *Given the Mahalanobis distances of the two features $m_{\text{inv}}, m_{\text{spur}} > 0$ such that $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}}+1} + 4$, and the training correlations $\zeta_{\text{tr}}, \zeta'_{\text{tr}} \in (0, 1)$, the performance shift over the testing set with correlation ratio $\zeta_{\text{te}} \in (0, 1)$ is bounded by*

$$\begin{aligned} &|A(\zeta_{\text{te}}; \zeta_{\text{tr}}) - A(\zeta_{\text{te}}; \zeta'_{\text{tr}})| \\ \leq& \frac{m_{\text{spur}}}{2\sqrt{2\pi m_{\text{inv}}}} \frac{M}{\zeta(1-\zeta)} (\zeta_{\text{te}} + 2)|\zeta_{\text{tr}} - \zeta'_{\text{tr}}| \end{aligned} \tag{5}$$

where ζ is between $\zeta_{\text{tr}}, \zeta'_{\text{tr}}$ and $M > 0$ is a constant.

The detailed proof is presented in appendix F.

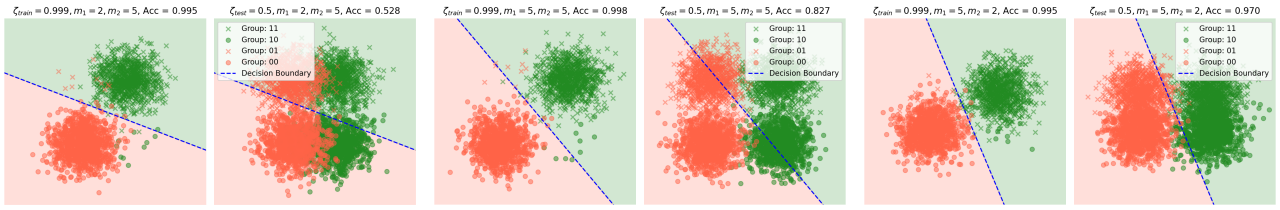
Quantitative effect of correlation ratio. In theorem 4, we show an important conclusion that when the change of the correlation $\delta = |\zeta'_{\text{tr}} - \zeta_{\text{tr}}|$ in the training set is small, the upper bound is related to the location $\zeta \in (\zeta_{\text{tr}}, \zeta'_{\text{tr}})$. Taking the limitation $\delta \to 0$ and letting $\zeta_{\text{te}} = 0.5$, $|\frac{\partial A}{\partial \zeta_{\text{tr}}}| \leq \frac{5 m_{\text{spur}}}{4\sqrt{2\pi m_{\text{inv}}}} \frac{M}{\zeta_{\text{tr}}(1-\zeta_{\text{tr}})}$. Therefore, the testing performance shift is bounded to be as small when ζ_{tr} approaches 0.5 – the shift is constant. But the shift is affected by ζ_{tr} significantly near $0, 1$ since the upperbound $\to \infty$ as $\zeta_{\text{tr}} \to 0, 1$. This is also consistent with the experimental results shown in fig. 1. In section 5.1, we further verify that such consistency is not only qualitative, but also quantitative. As a result, when adding samples at the plateau (i.e., $\zeta_{\text{tr}} \to 0.5$) of the curve, the effectiveness is marginal. This result serves as a guideline for deciding the expense of mitigating spurious correlation problems. On the opposite, adding samples before reaching the plateau improves the model performance significantly.

In addition, according to the formula of the training accuracy shown in eq. (4), testing accuracy $A(\zeta_{\text{te}}; \zeta_{\text{tr}})$ is linear w.r.t. ζ_{te}. Although when $\boldsymbol{z}_2$ is the spurious feature, ζ_{te} is fixed as 0.5, this observation can still be interesting for general distribution shift problems. More specifically, we have the following corollary.

Corollary 4.1. *(i) When $N = 2$, the testing accuracy shift linearly w.r.t. ζ_{te}. (ii) For general N, the testing accuracy is multi-linear with the ratios of all features.*

3.3 Separability of Features

Now that the spurious correlation does not hold in the testing scenario, a model $\boldsymbol{w} = [\boldsymbol{w}_1^T \ \boldsymbol{w}_2^T]^T$ is considered robust to the correlation shift when the model relies



(a) $m_{\text{inv}} = 2, m_{\text{spur}} = 5, A \approx 52.8\%$ (b) $m_{\text{inv}} = 5, m_{\text{spur}} = 5, A \approx 82.7\%$ (c) $m_{\text{inv}} = 5, m_{\text{spur}} = 2, A \approx 97.0\%$

Figure 3: The demonstration of the effectiveness of changing separability on the synthetic data. It is assumed that $\zeta_{\text{tr}} = 0.999$, and $\zeta_{\text{te}} = 0.5$ across three scenarios, where only $m_{\text{inv}}, m_{\text{spur}}$ are changed. A denotes the testing accuracy, which increases from 52.8% to 82.7% and finally 97.0% from (a) to (c). Higher $m_{\text{inv}}, m_{\text{spur}}$ indicate higher separability of the corresponding features.

more on the invariant feature than spurious feature. In the prediction is $\hat{y} = (\boldsymbol{w}^*)^T \boldsymbol{z} = (\boldsymbol{w}_1^*)^T \boldsymbol{z}_1 + (\boldsymbol{w}_2^*)^T \boldsymbol{z}_2$, the contributions of the spurious and the invariant features are $\mathbb{E}_{\boldsymbol{z} \in \mathcal{Z}} (\boldsymbol{w}_i^*)^T \boldsymbol{z}_i = (\boldsymbol{w}_i^*)^T \mu_i, i = 1, 2$. Hence the instability to the spurious correlation is quantified by

$$L(\boldsymbol{w}^*) = \frac{(\boldsymbol{w}_2^*)^T \mu_2}{(\boldsymbol{w}_1^*)^T \mu_1} = \frac{c_2 m_{\text{spur}}}{c_1 m_{\text{inv}}} = \tau \frac{m_{\text{spur}}}{m_{\text{inv}}} \tag{6}$$

However, note that τ is dependent on $m_{\text{inv}}, m_{\text{spur}}$, hence the relationship between $L(\boldsymbol{w})$ and $m_{\text{inv}}, m_{\text{spur}}$ is not trivial. In fact, we have the following results

Theorem 5. *(i) When* $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4$, $\frac{\tau}{m_1}$ *decreases monotonically with respect to* m_1. *(ii) When* $m_{\text{spur}} \leq 2(\sqrt{m_{\text{inv}} + 1} + 2)$, τm_2 *increases monotonically with respect to* m_2.

We prove this in appendix G. This result suggests that larger m_{inv} (more separable invariant feature) always leads to more robust classifiers. On the other hand, counterintuitively, the robustness is *only* guaranteed to decrease as the spurious feature becomes less separable *when* m_{spur} *is not too large*. Similar as discussed in corollary 2.2, when the spurious feature is much more separable, the optimal classifier becomes unstable and hard to justify when the correlation ratio is also high. The empirical results on synthetic data are demonstrated in fig. 3. When increasing m_{inv} or decreasing m_{spur}, the resulting classifier is more robust to the correlation shift in the testing data. Further experiments of the separability of real data and DNNs are presented in section 5.2.

Real Data and DNNs. While the conclusion in theorem 5 is established with Gaussian assumptions, the orthogonality assumption can be imposed through the mapping $\Phi(\boldsymbol{x}) = \boldsymbol{z}$ of the input data. Inspired by the theoretical results, we apply a direct approach to bridge the theoretical results and the application on real-world data and DNNs by feature alignment. To our best knowledge, such feature alignment trick, direct as it is, however, is not a common approach in handling spurious correlations. Re-weighting schemes are usually preferred. Specifically, we propose a regularization method to improve the robustness of models by manipulating only the separability of the data.

Let $X = X_{++} \cup X_{+-} \cup X_{-+} \cup X_{--} \in \mathbb{R}^{|X| \times d}$ be the training data, consisting of 4 groups. A model is defined as $\hat{y} = \sigma(\Phi(\boldsymbol{x})^T \boldsymbol{w})$, where $\Phi : \mathbb{R}^d \to \mathbb{R}^p$ is the backbone and σ is the sigmoid activation. Note that estimating and regularizing the inverse of the covariance matrices are computationally expensive in practice. Therefore, instead of regularizing $m_{\text{inv}}, m_{\text{spur}}$, we simplify the regularization term to

$$L_{reg} = \frac{\|\boldsymbol{\mu}_{\text{spur}}\|}{\|\boldsymbol{\mu}_{\text{inv}}\|} = \frac{\|\boldsymbol{\mu}_{\text{spur}}^+ - \boldsymbol{\mu}_{\text{spur}}^-\|}{\|\boldsymbol{\mu}_{\text{inv}}^+ - \boldsymbol{\mu}_{\text{inv}}^-\|} \tag{7}$$

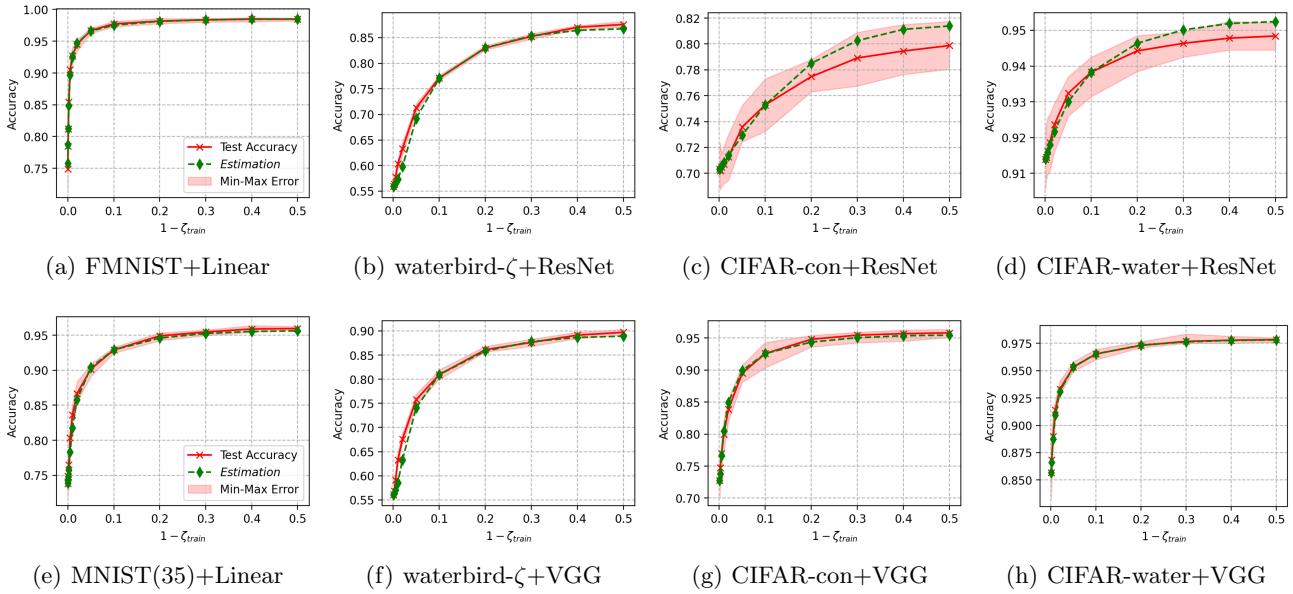
where $\boldsymbol{\mu}_{\text{inv,spur}}^+$ represents the average of the means of the groups where $a_{\text{inv,spur}} = 1$, and $\boldsymbol{\mu}_{\text{inv,spur}}^-$ represents that for groups where $a_{\text{inv,spur}} = -1$. For instance

$$\begin{aligned} \boldsymbol{\mu}_{\text{inv}}^+ &= (\mathbb{E}_{\boldsymbol{x} \in X_{++}} \Phi(\boldsymbol{x}) + \mathbb{E}_{\boldsymbol{x} \in X_{+-}} \Phi(\boldsymbol{x}))/2 \\ \boldsymbol{\mu}_{\text{inv}}^- &= (\mathbb{E}_{\boldsymbol{x} \in X_{-+}} \Phi(\boldsymbol{x}) + \mathbb{E}_{\boldsymbol{x} \in X_{--}} \Phi(\boldsymbol{x}))/2. \end{aligned} \tag{8}$$

The objective is defined as $\min_{\Phi, \boldsymbol{w}} \{L_{cls} + L_{reg}\}$ where L_{cls} is the classification loss. And the optimization is implemented batch-wise. The experiments are carried out to validate effectiveness in section 5.2.

4 DATASETS

To construct spurious correlations, we modify existing datasets including simple datasets MNIST (LeCun et al., 1998), Fashion-MNIST (FMNIST) (Xiao et al., 2017) for linear models, and complex datasets CIFAR-10 (Krizhevsky et al., 2009), Caltech-UCSD Birds-200-2011 (CUB) (Wah et al., 2011), Places (Zhou et al., 2017) for DNNs. Inspired by Dominoes (Shah et al., 2020; Pagliardini et al., 2022) we concatenate images from the opposite class as the spurious feature. In this way, it can be justified in advance that the separability of the invariant and spurious features should



(a) FMNIST+Linear (b) waterbird-ζ+ResNet (c) CIFAR-con+ResNet (d) CIFAR-water+ResNet

(e) MNIST(35)+Linear (f) waterbird-ζ+VGG (g) CIFAR-con+VGG (h) CIFAR-water+VGG

Figure 4: Accuracy trend w.r.t. the ratio ζ_{tr}. Red solid curves show the raw testing accuracy, while green dashed curves show the estimated results based on our theorems. The axis is set as $1-\zeta_{\text{tr}}$ to have a right-to-left trend. The results show that our theorem-based estimation closely matches the actual accuracy trend across multiple datasets and models.

be equivalent. This is applied to MNIST, FMNIST, and CIFAR-10. In order to also study the difference of artificial spurious features, we add a watermarks in CIFAR-10, where a 4×4 back square is attached in the bottom right corner of the image. We also follow (Sagawa et al., 2019) to construct waterbird, where birds from CUB-200 are divided into water birds and land birds. They are cropped and attached over water and land background scenes from Places. To balance the two classes and augment the data, we re-sample the waterbird images and resize the birds to 15% percent of the entire image and randomly rotate them by $[-15°, 15°]$. The resulting dataset is named waterbird-ζ as in all datasets, the correlation ratio ζ is controllable. We also include a subset of CelebA (Liu et al., 2015b). The informative features are the hair colors and the spurious features are the gender of the subject. Given correlation ratio ζ, we fix seeds and randomly generate the largest subset that satisfies the group relations.

We conducted binary classification on two sub-classes of the MNIST, FMNIST, and CIFAR-10 datasets. The first two classes of FMNIST and CIFAR-10 are taken into consideration without selection, which are "top vs. trouser" and "airplane vs automobile". As for MNIST, the sub-classes are 3 vs 5 and 5 vs 8[1] since distinguishing between 0, 1 digit images can be too trivial. Samples from the datasets are demonstrated in the appendix H. The input of MNIST and FMNIST is of size 1568. CIFAR-concate is of size $6 \times 32 \times 32$. CIFAR-watermark is of size $3 \times 32 \times 32$. And the input of waterbirds and CelebA is resized to $3 \times 128 \times 128$.

[1]These are the most indistinguishable pairs in MNIST

5 EXPERIMENTS

In this section, we conduct various experiments to demonstrate that, although the theoretical results are proposed under constraints, they possess the capacity to generalize to real data and DNNs. MNIST and FMNIST are tested with linear models and others are tested with DNNs. The experiments are implemented using Intel Core i9-9960X CPU @ 3.10GHz, paired with Quadro RTX 6000 GPUs.

5.1 Estimating the Accuracy Shift – Using the Correlation Ratio

We let ζ_{tr} vary in $[0.5, 1]$. Models are implemented using `sklearn` and `PyTorch`. VGG-16 (Simonyan and Zisserman, 2014), ResNet-18 (He et al., 2016), and AlexNet (Krizhevsky, 2014) are tested. DNNs are optimized using the `SGD` solver with `lr=1e-3,momentum=0.9,weight_decay=5e-4` for 200 epochs. In order to demonstrate our theoretical results generalize to complex scenarios, we estimate the accuracy trend of complex models on real data.

Note that from eqs. (4) and (5), the testing accuracy

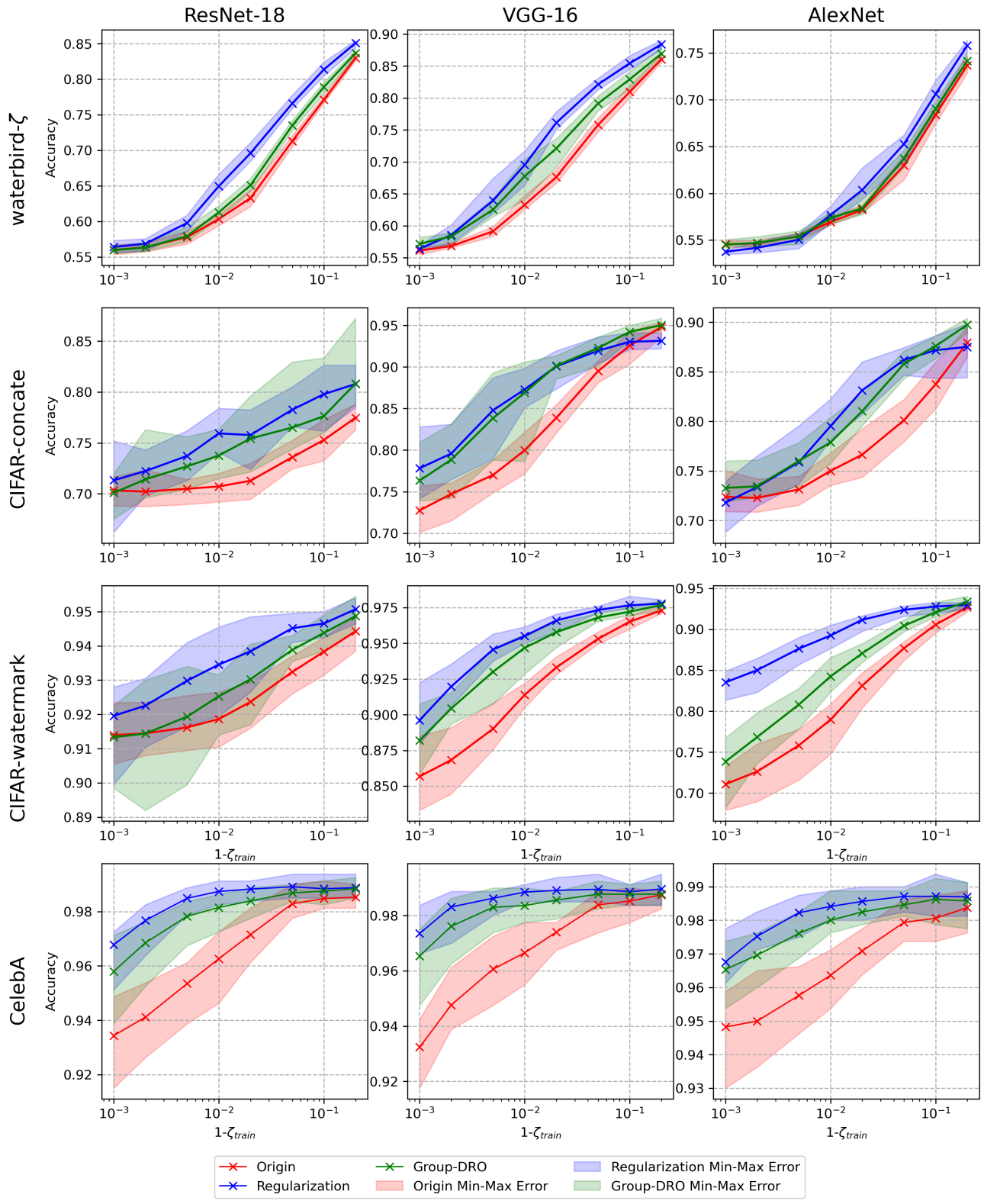


Figure 5: The testing accuracy of the origin (red), separability regularization (blue), and Group-DRO (green). The light-colored regions are the min-max error region of 10 rounds of experiments, seeds from 0 to 9.

is determined by the separability $m_{\text{inv}}, m_{\text{spur}}$ of features and the correlation ratio ζ_{tr}. Therefore, given the estimated $\hat{m}_1, \hat{m}_2$, the accuracy will be determined solely by the ratio ζ_{tr}. Although m_{inv} and m_{spur} are unknown for real data, it is feasible to estimate the separability by indicators $\hat{m}_1, \hat{m}_2$ from the data accuracy given the initial ζ_{tr} values that are near 1, and then $\hat{m}_1, \hat{m}_2$ are used to estimate the accuracy with other values of ζ_{tr} (that are close to 0.5). The separability indicators are estimated at $\zeta_{\text{tr}}^1 = 0.9, 0.999$. The experiments are repeated 10 times with seeds 0 to 9. And the estimations are shown in fig. 4. Please refer to appendix I for the detailed algorithm and the results for AlexNet and MNIST-5 vs 8. For the results of VGG-16, AlexNet, and MNIST. Results show that the estimation of the accuracy trend w.r.t. ζ_{tr} locates very close to the average accuracy in 10 rounds. This suggests that (i) Even for non-linear models and complicated real datasets, the testing accuracy is still highly dependent on the separability of features and the training correlation ratio, and that (ii) The trend of testing accuracy follows the theoretical results.

The result serves as guidelines for dataset construction and augmentations – given a highly biased dataset with the initial testing accuracy, we can *estimate the accuracy improvement resulting from adding a specific number of samples to balance the data.*

5.2 Separability vs Correlation Ratio– Using the Separability

The theoretical insights show that both the initial ratio in the training data and the separability of the features affect the models' robustness toward the correlation shift. However, existing work always focuses on schemes that mitigate the correlation ratio of the training data, while the separability is paid little attention to. In section 3.3, it is established that increasing the separability of the invariant features and decreasing the separability of the spurious features can both improve the robustness of the classifier towards correlation shift when there exists a spurious correlation.

We then compare the results of the feature alignment regularization (blue) in eq. (7) with the raw accuracy (red) and the SOTA reweighting method Group-DRO (green), which is implemented precisely following the original implementation [2] Still, experiments are carried out on CIFAR-concate, CIFAR-watermark, waterbird-ζ and CelebA using the three DNNs. The results are shown in fig. 5. It can be found that, although implemented with completely different schemes, both feature alignment and Group-DRO manage to improve the models' robustness to the spurious correlation. It should be noticed that since the size of the CIFAR-concate and CIFAR-watermark datasets is much smaller than that of waterbird-ζ, the CIFAR-related results are not as smooth. Especially, when $\zeta_{\text{tr}} \to 1$, the minor groups are almost empty, making the optimization difficult. The results demonstrate that the separability of features should be paid more attention when resolving spurious correlation problems.

[2] Group-DRO is implemented following the official release, which is licensed under the MIT License (Copyright (c) 2022 The authors).

6 CONCLUSIONS

In conclusion, this study provides a comprehensive theoretical analysis of the factors of the spurious correlation problem and their impact on machine learning models. We theoretically study the two critical factors affecting the model performance in the testing distribution: the correlation ratio and the feature separability. Our research quantitatively underscores how the correlation ratio in the training dataset and its changes influence the model's performance. It is also revealed that the change in the separability of the data can affect the model performance when the correlation ratio remains invariant. For the correlation ratio, we propose a method for estimating model performance concerning correlation ratio changes, which has proven effective on real data under both linear and non-linear models. To improve feature separability, we introduced a simple regularization term that enhances model robustness without the need for reweighting schemes based on the theoretical analysis. It is admitted that the proposed regularization technique is defined as simple as possible to primarily validate theoretical findings. Our study suggests that regularization techniques and reweighting schemes can both improve the robustness. Therefore, it would be interesting for future work to dive deeper to develop more sophisticated techniques that combine regularization and reweighting together. These efforts could further strengthen the handling of spurious correlations in machine learning studies

Acknowledgements

This work was partially supported by the EMBRIO Institute, contract #2120200, a National Science Foundation (NSF) Biology Integration Institute, Purdue's Elmore ECE Emerging Frontiers Center, and NSF IIS #1955890, IIS #2146091.

References

Ahmed, F., Bengio, Y., Van Seijen, H., and Courville, A. (2021). Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473.

Bengio, Y., Goodfellow, I., and Courville, A. (2017). *Deep learning*, volume 1. MIT press Cambridge, MA, USA.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259.

Byrd, J. and Lipton, Z. (2019). What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pages 872–881. PMLR.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Duchi, J. C., Hashimoto, T., and Namkoong, H. (2019). Distributionally robust losses against mixture covariate shifts. *Under review*, 2:1.

Erfani, S., Baktashmotlagh, M., Moshtaghi, M., Nguyen, X., Leckie, C., Bailey, J., and Kotagiri, R. (2016). Robust domain generalisation by enforcing distribution invariance. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 1455–1461. AAAI Press.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Ghifary, M., Balduzzi, D., Kleijn, W. B., and Zhang, M. (2016). Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

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

Hu, S., Zhang, K., Chen, Z., and Chan, L. (2020). Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pages 292–302. PMLR.

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

Janai, J., Güney, F., Behl, A., Geiger, A., et al. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Jin, X., Lan, C., Zeng, W., and Chen, Z. (2020). Feature alignment and restoration for domain generalization and adaptation. *arXiv preprint arXiv:2006.12009*.

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

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn,

C. (2021). Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

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

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

Ming, Y., Yin, H., and Li, Y. (2022). On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10051–10059.

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR.

Nagarajan, V., Andreassen, A., and Neyshabur, B. (2020). Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. (2020). Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684.

Pagliardini, M., Jaggi, M., Fleuret, F., and Karimireddy, S. P. (2022). Diversity through disagreement for better transferability. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Saab, K., Hooper, S., Chen, M., Zhang, M., Rubin, D., and Ré, C. (2022). Reducing reliance on spurious features in medical image classification with spatial specificity.

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

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. (2020). An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585.

Shi, Y., Daunhawer, I., Vogt, J. E., Torr, P., and Sanyal, A. (2022). How robust is unsupervised representation learning to distribution shift? In *The Eleventh International Conference on Learning Representations*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Van de Poel, I. and Royakkers, L. (2023). *Ethics, technology, and engineering: An introduction*. John Wiley & Sons.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. (2022). Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, M. and Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. (2022). Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. (2022). Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

Zhou, C., Ma, X., Michel, P., and Neubig, G. (2021). Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, pages 12857–12867. PMLR.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [No]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Proof of Lemma 1.

Lemma 1. (Accuracy) Given the parameters $\Theta = (\boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$, the accuracy of the classifier $\boldsymbol{w}$ is

$$A(\boldsymbol{w}; \Theta) = \frac{1}{2}\Big(1 + \sum_{\boldsymbol{a}\in\{\pm 1\}^N} \gamma_{\boldsymbol{a}} \mathrm{erf}\big(\frac{\boldsymbol{\mu}_{\boldsymbol{a}}^T \boldsymbol{w}}{\sqrt{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}}\big)\Big) \tag{9}$$

where $\gamma_{\boldsymbol{a}} = \big(\prod_{n=1}^N [\alpha_n^{t_{\boldsymbol{a},n}}(1-\alpha_n)^{(1-t_{\boldsymbol{a},n})}] + \prod_{n=1}^N [\beta_n^{t_{\boldsymbol{a},n}}(1-\beta_n)^{(1-t_{\boldsymbol{a},n})}]\big)/2$.

Proof. Considering the combination, there are 2^N groups, which are denoted by $\boldsymbol{a} \in \{\pm 1\}^N$. Depending on the corresponding group, the mean of the entire latent representation is denoted as

$$\boldsymbol{\mu}_{\boldsymbol{a}} = \begin{bmatrix} a_1\boldsymbol{\mu}_1 \\ a_2\boldsymbol{\mu}_2 \\ \vdots \\ a_N\boldsymbol{\mu}_N \end{bmatrix} \tag{10}$$

where $t_{\boldsymbol{a},n} = 1$ when $a_n = 1$ and $t_{\boldsymbol{a},n} = 0$ when $a_n = -1$. Note that as the attribution of the informative feature, a_1 is equivalent to y. Besides, the correlation ratio satisfies that $\alpha_1 = \beta_1 = 1$. Then the joint distribution of the data conditioned on y can be written as

$$\begin{aligned} p(\boldsymbol{z}|y=1) &= \prod_{n=1}^N p(z_n|y=1) = p(z_1|y=1)\prod_{n=2}^N p(z_n|y=1) \\ &= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \sum_{a_2\cdots a_N\in\{0,1\}} \Big\{ \Big[\prod_{n=2}^N \alpha_n^{t_{\boldsymbol{a},n}}(1-\alpha_n)^{(1-t_{\boldsymbol{a},n})}\Big] \exp\Big(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{a}})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{a}})^T\Big)\Big\} \end{aligned} \tag{11}$$

$$\begin{aligned} p(\boldsymbol{z}|y=-1) &= \prod_{n=1}^N p(z_n|y=-1) = p(z_1|y=-1)\prod_{n=2}^N p(z_n|y=-1) \\ &= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \sum_{a_2\cdots a_N\in\{0,1\}} \Big\{ \Big[\prod_{n=2}^N \beta_n^{t_{\boldsymbol{a},n}}(1-\beta_n)^{(1-t_{\boldsymbol{a},n})}\Big] \exp\Big(-\frac{1}{2}(\boldsymbol{z}+\boldsymbol{\mu}_{\boldsymbol{a}})^T\Sigma^{-1}(\boldsymbol{z}+\boldsymbol{\mu}_{\boldsymbol{a}})^T\Big)\Big\} \end{aligned} \tag{12}$$

Note that $y \sim \mathrm{Uniform}\{-1, 1\}$, hence TN+FP = TP+FN, and the accuracy is the average of the recall and the true negative rate. Let $\Omega(\boldsymbol{z}) = \{\boldsymbol{w}|\boldsymbol{w}^T\boldsymbol{z} > 0\}$ be the half-space of the positive prediction. Then the recall can be computed as

$$\begin{aligned} &\int_{\Omega(\boldsymbol{w})} \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \sum_{a_2\cdots a_N\in\{0,1\}} \Big\{ \Big[\prod_{n=2}^N \alpha_n^{t_{\boldsymbol{a},n}}(1-\alpha_n)^{(1-t_{\boldsymbol{a},n})}\Big] \exp\Big(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{a}})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{a}})^T\Big)\Big\} \\ =&\frac{1}{2} \sum_{a_2\cdots a_N\in\{0,1\}} \Big\{ \Big[\prod_{n=2}^N \alpha_n^{t_{\boldsymbol{a},n}}(1-\alpha_n)^{(1-t_{\boldsymbol{a},n})}\Big] \Big(1 - \mathrm{erf}\big(-\frac{\boldsymbol{\mu}_{\boldsymbol{a}}^T\boldsymbol{w}}{\sqrt{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}}\big)\Big)\Big\} \\ =&\frac{1}{2}\Big(1 + \sum_{a_2\cdots a_N\in\{0,1\}} \Big[\prod_{n=2}^N \alpha_n^{t_{\boldsymbol{a},n}}(1-\alpha_n)^{(1-t_{\boldsymbol{a},n})}\Big] \mathrm{erf}\big(\frac{\boldsymbol{\mu}_{\boldsymbol{a}}^T\boldsymbol{w}}{\sqrt{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}}\big)\Big) \end{aligned} \tag{13}$$

Similarly, the true negative rate can be written as

$$\int_{\mathbb{R}^d\setminus\Omega(\boldsymbol{w})} p(\boldsymbol{z}|y=-1)\mathrm{d}\boldsymbol{z} = \frac{1}{2}\Big(1+\sum_{a_2\cdots a_N\in\{0,1\}}\Big[\prod_{n=1}^{N}\beta_n^{t_{\boldsymbol{a},n}}(1-\beta_n)^{(1-t_{\boldsymbol{a},n})}\Big]\mathrm{erf}(\frac{\boldsymbol{\mu}_{\boldsymbol{a}}^T\boldsymbol{w}}{\sqrt{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}})\Big) \tag{14}$$

Then as the average, the accuracy is

$$\begin{aligned} A(\boldsymbol{w};\boldsymbol{\mu},\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}) =& \frac{1}{2}\Big(\int_{\Omega(\boldsymbol{w})} p(\boldsymbol{z}|y=1)\mathrm{d}\boldsymbol{z} + \int_{\mathbb{R}^d\setminus\Omega(\boldsymbol{w})} p(\boldsymbol{z}|y=-1)\mathrm{d}\boldsymbol{z}\Big) \\ =& \frac{1}{2}\Big(1+\sum_{\boldsymbol{a}\in\{\pm1\}^N}\gamma_{\boldsymbol{a}}\mathrm{erf}\Big(\frac{\boldsymbol{\mu}_{\boldsymbol{a}}^T\boldsymbol{w}}{\sqrt{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}}\Big)\Big) \end{aligned} \tag{15}$$

where $\gamma_{\boldsymbol{a}} = \frac{\left[\prod_{n=1}^N \alpha_n^{t_{\boldsymbol{a},n}}(1-\alpha_n)^{(1-t_{\boldsymbol{a},n})}\right]+\left[\prod_{n=1}^N \beta_n^{t_{\boldsymbol{a},n}}(1-\beta_n)^{(1-t_{\boldsymbol{a},n})}\right]}{2}$.□

B Proof of Lemma 2.

Lemma 2. The Bayesian optimal classifier $\boldsymbol{w}* = [(\boldsymbol{w}_1^*)^T,\cdots,(\boldsymbol{w}_N^*)^T]^T$ given $\Theta = (\boldsymbol{\mu},\Sigma,\boldsymbol{\gamma})$ can be written as $\boldsymbol{w}_i^* = \eta c_i\Sigma_i^{-1}\boldsymbol{\mu}_i, i=1,\cdots,N$, such that

$$c_n = \sum_{\substack{\boldsymbol{a}\in\{\pm1\}^N\\ a_n=1}}(\gamma_{\boldsymbol{a}}-\gamma_{-\boldsymbol{a}})\exp\Big(-\frac{(\sum_{n=1}^N t_{\boldsymbol{a},n}c_n m_n)^2}{2\sum_{n=1}^N c_n^2 m_n}\Big)$$

where $\eta>0$ is a positive constant that determines the norm of the classifier. And here $m_n = \boldsymbol{\mu}_n^T\Sigma^{-1}\boldsymbol{\mu}_n > 0$ is the Mahalanobis distance of the n-th feature.

proof. In order to maximize $A(\boldsymbol{w})$, the stationary point can be computed by $\nabla_{\boldsymbol{w}}A(\boldsymbol{w})=0$

$$0 = \nabla_{\boldsymbol{w}}A(\boldsymbol{w}) = \frac{1}{2}\nabla_{\boldsymbol{w}}\Big(\sum_{\boldsymbol{a}\in\{\pm1\}^N}\gamma_{\boldsymbol{a}}\mathrm{erf}\Big(\frac{\boldsymbol{\mu}_{\boldsymbol{a}}^T\boldsymbol{w}}{\sqrt{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}}\Big)\Big) \tag{16}$$

$$= \frac{1}{\sqrt{\pi}}\frac{\Sigma\boldsymbol{w}\boldsymbol{w}^T-(\boldsymbol{w}^T\Sigma\boldsymbol{w})I}{(\boldsymbol{w}^T\Sigma\boldsymbol{w})^{3/2}}\sum_{\boldsymbol{a}\in\{\pm1\}^N}\Big\{\gamma_{\boldsymbol{a}}\exp\Big(-\frac{(\boldsymbol{\mu}_{\boldsymbol{a}}^T\boldsymbol{w})^2}{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}\Big)\boldsymbol{\mu}_{\boldsymbol{a}}\Big\} \tag{17}$$

It suffices to solve $\Sigma\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{q} = (\boldsymbol{w}^T\Sigma\boldsymbol{w})\boldsymbol{q}$, where

$$\boldsymbol{q} = \sum_{\boldsymbol{a}\in\{\pm1\}^N}\Big\{\gamma_{\boldsymbol{a}}\exp\Big(-\frac{(\boldsymbol{\mu}_{\boldsymbol{a}}^T\boldsymbol{w})^2}{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}\Big)\boldsymbol{\mu}_{\boldsymbol{a}}\Big\} \tag{18}$$

can be seen as the weighted summation of all $\boldsymbol{\mu}_{\boldsymbol{a}}$. Note that $\boldsymbol{q}$ is an eigenvector of the 1-rank matrix $\Sigma\boldsymbol{w}\boldsymbol{w}^T$. Therefore, the eigenvector $\boldsymbol{q}$ is colinear with $\Sigma\boldsymbol{w}$. That is, $\Sigma\boldsymbol{w}\boldsymbol{w}^T(\Sigma\boldsymbol{w}) = \Sigma\boldsymbol{w}(\boldsymbol{w}^T\Sigma\boldsymbol{w})$. So $\exists\eta\in\mathbb{R}$ s.t. $\Sigma\boldsymbol{w}=\eta\boldsymbol{q}$, which can be written in detail as

$$\begin{bmatrix}\boldsymbol{w}_1\\ \vdots\\ \boldsymbol{w}_N\end{bmatrix} = \eta\sum_{\boldsymbol{a}\in\{\pm1\}^N}\gamma_{\boldsymbol{a}}\exp\Big(-\frac{(\sum_{n=1}^N a_n\boldsymbol{\mu}^T\boldsymbol{w}_n)^2}{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}\Big)\begin{bmatrix}a_1\Sigma_1^{-1}\boldsymbol{\mu}_1\\ \vdots\\ a_N\Sigma_N^{-1}\boldsymbol{\mu}_N\end{bmatrix} =: \eta\begin{bmatrix}c_1\Sigma_1^{-1}\boldsymbol{\mu}_1\\ \vdots\\ c_N\Sigma_N^{-1}\boldsymbol{\mu}_N\end{bmatrix} \tag{19}$$

Substituting $\boldsymbol{w}_n = \eta c_n \Sigma_n^{-1} \boldsymbol{\mu}_n$ back, we have

$$c_k = \sum_{\boldsymbol{a}\in\{\pm 1\}^N} \gamma_{\boldsymbol{a}} a_k \exp\Big(-\frac{(\sum_{n=1}^N a_n \boldsymbol{\mu}^T \boldsymbol{w}_n)^2}{2\boldsymbol{w}^T\Sigma\boldsymbol{w}}\Big) \tag{20}$$

$$= \sum_{\boldsymbol{a}\in\{\pm 1\}^N} \gamma_{\boldsymbol{a}} a_k \exp\Big(-\frac{(\sum_{n=1}^N a_n \boldsymbol{\mu}^T c_n \Sigma_n^{-1}\boldsymbol{\mu}_n \eta)^2}{2\sum_{n=1}^N (c_n\Sigma_n^{-1}\boldsymbol{\mu}_n\eta)^T\Sigma_n(c_n\Sigma_n^{-1}\boldsymbol{\mu}_n\eta)}\Big) \tag{21}$$

$$= \sum_{\boldsymbol{a}\in\{\pm 1\}^N} \gamma_{\boldsymbol{a}} a_k \exp\Big(-\frac{(\sum_{n=1}^N a_n c_n m_n)^2}{2\sum_{n=1}^N c_n^2 m_n}\Big) \tag{22}$$

$$= \sum_{\boldsymbol{a}\in\{\pm 1\}^N, a_k=1} (\gamma_{\boldsymbol{a}} - \gamma_{-\boldsymbol{a}}) \exp\Big(-\frac{1}{2}\frac{(\sum_{n=1}^N a_n c_n m_n)^2}{\sum_{n=1}^N c_n^2 m_n}\Big) \tag{23}$$

and hence prove the statement.□

C Proof of Corollary 2.1

Corollary 2.1 Given $\boldsymbol{m}, \mathcal{Z}, \mathcal{Y}, \zeta$, let $\eta = 1/c_{\text{inv}}, \tau = c_{\text{spur}}\eta = c_{\text{spur}}/c_{\text{inv}}$, then the optimal classifier is denoted by $\boldsymbol{w}^*_{\text{inv}} = \Sigma^{-1}_{\text{inv}}\boldsymbol{\mu}_{\text{inv}}, \boldsymbol{w}^*_{\text{spur}} = \tau\Sigma^{-1}_{\text{spur}}\boldsymbol{\mu}_{\text{spur}}$.

proof. When $N = 2$, we denote the subscripts $_{1,2}$ by $_{\text{inv,spur}}$ for clarity. Note that $\boldsymbol{w}^*_{\text{inv}} = c_{\text{inv}}\eta\Sigma^{-1}_{\text{inv}}\boldsymbol{\mu}_{\text{inv}}, \boldsymbol{w}^*_{\text{spur}} = c_{\text{spur}}\eta\Sigma^{-1}_{\text{spur}}\boldsymbol{\mu}_{\text{spur}}$. Let $\eta = 1/c_{\text{inv}}$, and $\tau = c_{\text{spur}}/c_{\text{inv}}$, then $\boldsymbol{w}^*_{\text{inv}} = \Sigma^{-1}_{\text{inv}}\boldsymbol{\mu}_{\text{inv}}, \boldsymbol{w}^*_{\text{spur}} = \tau\Sigma^{-1}_{\text{spur}}\boldsymbol{\mu}_{\text{spur}}$.□

D Proof of Corollary 2.2

Corollary 2.2 (informal) (i) When $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}}+1} + 4$, $\exists g \in \mathcal{C}^1$ s.t. $\tau = g(\zeta)$, and g is monotonically increasing. (ii) When $m_{\text{spur}} \geq m_{\text{inv}} + 4\sqrt{m_{\text{inv}}+1} + 4$, τ is monotonically increasing w.r.t. ζ when ζ is not very close to 1 or 0. There can be multiple optimal classifiers when ζ is close to 1 or 0.

proof. We first show the existence of $g(\zeta) \in \mathcal{C}^1$.

When $N = 2$, the expression for $\boldsymbol{c}$ can be written as

$$\begin{cases} c_{\text{inv}} = \zeta\exp\Big(-\frac{1}{2}\frac{(c_{\text{inv}}m_{\text{inv}} + c_{\text{spur}}m_{\text{spur}})^2}{c_{\text{inv}}^2 m_{\text{inv}} + c_{\text{spur}}^2 m_{\text{spur}}}\Big) + (1-\zeta)\exp\Big(-\frac{1}{2}\frac{(c_{\text{inv}}m_{\text{inv}} - c_{\text{spur}}m_{\text{spur}})^2}{c_{\text{inv}}^2 m_{\text{inv}} + c_{\text{spur}}^2 m_{\text{spur}}}\Big) \\ c_{\text{spur}} = \zeta\exp\Big(-\frac{1}{2}\frac{(c_{\text{inv}}m_{\text{inv}} + c_{\text{spur}}m_{\text{spur}})^2}{c_{\text{inv}}^2 m_{\text{inv}} + c_{\text{spur}}^2 m_{\text{spur}}}\Big) - (1-\zeta)\exp\Big(-\frac{1}{2}\frac{(c_{\text{inv}}m_{\text{inv}} - c_{\text{spur}}m_{\text{spur}})^2}{c_{\text{inv}}^2 m_{\text{inv}} + c_{\text{spur}}^2 m_{\text{spur}}}\Big) \end{cases} \tag{24}$$

Recall we defined that $\tau = c_{\text{spur}}/c_{\text{inv}}$, then

$$\tau = \frac{\zeta\exp\Big(-\frac{1}{2}\frac{(c_{\text{inv}}m_{\text{inv}}+c_{\text{spur}}m_{\text{spur}})^2}{c_{\text{inv}}^2 m_{\text{inv}}+c_{\text{spur}}^2 m_{\text{spur}}}\Big) - (1-\zeta)\exp\Big(-\frac{1}{2}\frac{(c_{\text{inv}}m_{\text{inv}}-c_{\text{spur}}m_{\text{spur}})^2}{c_{\text{inv}}^2 m_{\text{inv}}+c_{\text{spur}}^2 m_{\text{spur}}}\Big)}{\zeta\exp\Big(-\frac{1}{2}\frac{(c_{\text{inv}}m_{\text{inv}}+c_{\text{spur}}m_{\text{spur}})^2}{c_{\text{inv}}^2 m_{\text{inv}}+c_{\text{spur}}^2 m_{\text{spur}}}\Big) + (1-\zeta)\exp\Big(-\frac{1}{2}\frac{(c_{\text{inv}}m_{\text{inv}}-c_{\text{spur}}m_{\text{spur}})^2}{c_{\text{inv}}^2 m_{\text{inv}}+c_{\text{spur}}^2 m_{\text{spur}}}\Big)} \tag{25}$$

$$= \tanh\big(\frac{\log\zeta - \log(1-\zeta)}{2} - \frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}} + m_{\text{spur}}\tau^2}\big) \tag{26}$$

Let $\omega = \frac{\log\zeta - \log(1-\zeta)}{2}$, and $\phi(\omega,\tau) = \tanh(\omega - \frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}}+m_{\text{spur}}\tau^2}) - \tau$. Then the solutions form a curve $\phi(\omega,\tau) = 0$. Consider the partial derivative w.r.t. τ as

$$\frac{\partial\phi}{\partial\tau} = \frac{m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})\text{sech}^2(\omega - \frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}}+m_{\text{spur}}\tau^2})}{(m_{\text{inv}} + m_{\text{spur}}\tau^2)^2} - 1 = 0 \tag{27}$$

where $\text{sech}^2(\omega - \frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}}+m_{\text{spur}}\tau^2}) = 1 - \tanh^2(\omega - \frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}}+m_{\text{spur}}\tau^2}) = 1 - \tau^2 \in (0,1)$. Substitute this back, we have

$$m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})(1-\tau^2) = (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2 \tag{28}$$

It can be solved by

$$\tau_{1,2} = \sqrt{\frac{m_{\text{inv}}(m_{\text{inv}} + m_{\text{spur}} - 2 \pm \sqrt{(m_{\text{inv}} - m_{\text{spur}})^2 - 8(m_{\text{inv}} + m_{\text{spur}})})}{2m_{\text{spur}}(1 + m_{\text{inv}})}}, \tau_1 < \tau_2 \tag{29}$$

In order for this to be valid, the area of $m_{\text{inv}}, m_{\text{spur}}$ can be obtained by solving

$$\begin{cases} (m_{\text{inv}} - m_{\text{spur}})^2 > 8(m_{\text{inv}} + m_{\text{spur}}) \\ m_{\text{inv}} + m_{\text{spur}} - 2 > \sqrt{(m_{\text{inv}} - m_{\text{spur}})^2 - 8(m_{\text{inv}} + m_{\text{spur}})} \\ \dfrac{(m_{\text{inv}} + m_{\text{spur}} - 2)m_{\text{inv}}m_{\text{spur}} \pm \sqrt{m_{\text{inv}}^2 m_{\text{spur}}^2((m_{\text{inv}} - m_{\text{spur}})^2 - 8(m_{\text{inv}} + m_{\text{spur}}))}}{2m_{\text{spur}}^2(1 + m_{\text{inv}})} \\ \dfrac{m_{\text{inv}}(m_{\text{inv}} + m_{\text{spur}} - 2 \pm \sqrt{(m_{\text{inv}} - m_{\text{spur}})^2 - 8(m_{\text{inv}} + m_{\text{spur}})})}{2m_{\text{spur}}(1 + m_{\text{inv}})} < 1 \\ m_{\text{inv}}, m_{\text{spur}} > 0 \end{cases} \tag{30}$$

And the resulting feasible area for $m_{\text{inv}}, m_{\text{spur}}$ is

$$(m_{\text{inv}}, m_{\text{spur}}) \in (0, \infty) \times (m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4, \infty) \tag{31}$$

Hence when $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4$, $\frac{\partial \phi}{\partial \tau} \neq 0$. By the implicit function theorem, there exists a function such that $c_{\text{spur}}/c_{\text{inv}} = \tau = g_{\text{inv}}(\omega)$. Let g be defined by $g(\zeta) = g_{\text{inv}}(\frac{\log \zeta - \log(1-\zeta)}{2}) = c_{\text{spur}}/c_{\text{inv}} = c_{\text{spur}}/\eta$. Then $\boldsymbol{w}^*_{\text{spur}} = g(\zeta)\Sigma^{-1}_{\text{spur}}\boldsymbol{\mu}_{\text{spur}}$.

E Formalized Discussion on the relationship between $m_{\text{inv}}, m_{\text{spur}}$

Corollary 2.3.1 Given $\boldsymbol{m}, \mathcal{Z}, \mathcal{Y}$ defined above, if $m_{\text{spur}} = m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4$, then $\boldsymbol{w}^*$ is still uniquely determined by ζ, through continuous function g, but $g(\zeta)$ is not differentiable at $\zeta^*, 1 - \zeta^*$, where

$$\begin{aligned} \zeta^* = \frac{1}{2}\Bigg(1 + \tanh\Big(&\operatorname{arctanh}\Big(\sqrt{\frac{m_{\text{inv}}(m_{\text{inv}} + 2\sqrt{m_{\text{inv}} + 1} + 1)}{(1 + m_{\text{inv}})(m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4)}}\Big) \\ &+ \frac{(m_{\text{inv}} + 1)(m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4)}{2(m_{\text{inv}} + \sqrt{m_{\text{inv}} + 1} + 1)}\Big)\Bigg) \end{aligned} \tag{32}$$

proof. Since $\tau = \tanh(\omega - \frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}} + m_{\text{spur}}\tau^2})$, ω can be written as

$$\omega = \operatorname{arctanh}(\tau) + \frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}} + m_{\text{spur}}\tau^2} \tag{33}$$

On the other hand, the function $g(\zeta)$ is not differentiable at ζ^* where $\frac{\partial \phi}{\partial \tau} = 0$. And since $m_{\text{spur}} = m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4$

$$\tau = \sqrt{\frac{m_{\text{inv}}(m_{\text{inv}} + m_{\text{spur}} - 2)}{2m_{\text{spur}}(1 + m_{\text{inv}})}} = \sqrt{\frac{m_{\text{inv}}(m_{\text{inv}} + 2\sqrt{m_{\text{inv}} + 1} + 1)}{(m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4)(1 + m_{\text{inv}})}} \tag{34}$$

Substituting this back results in

$$\omega = \operatorname{arctanh}\Big(\sqrt{\frac{m_{\text{inv}}(m_{\text{inv}} + 2\sqrt{m_{\text{inv}} + 1} + 1)}{(1 + m_{\text{inv}})(m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4)}}\Big) + \frac{(m_{\text{inv}} + 1)(m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4)}{2(m_{\text{inv}} + \sqrt{m_{\text{inv}} + 1} + 1)}\Big) \tag{35}$$

Note that $\zeta = \frac{1}{2}(1 + \tanh(\omega))$, we have the statement proved.

It can be easily found that ζ^* is monotonically increasing with m_{inv}, and $\zeta^* > \zeta^*_{m_{\text{inv}}=0} = (1+\tanh(2))/2 \approx 0.9820$, which is already very highly correlated. Besides, in practice, the informative features need to be basically separable, resulting in even larger ζ. As a result, for most of the scenarios where $\zeta \in (1-\zeta^*, \zeta^*)$, there's a unique classifier $\boldsymbol{w}$. In the extreme case, however, multiple classifiers can achieve the optimal classifier at the same time.

Corollary 2.3.2 Given $\boldsymbol{m}, \mathcal{Z}, \mathcal{Y}$ defined above, if $m_{\text{spur}} > m_{\text{inv}} + 4\sqrt{m_{\text{inv}}+1} + 4$, then (i) $\boldsymbol{w}^*$ is uniquely determined by ζ in $(0, 1-\zeta_2^*) \cup (1-\zeta_{\text{inv}}^*, \zeta_{\text{inv}}^*) \cup (\zeta_2^*, 1)$. (ii) When $\zeta \in (1-\zeta_2^*, 1-\zeta_{\text{inv}}^*) \cup (\zeta_{\text{inv}}^*, \zeta_2^*)$, there are three optimal classifiers, where $\zeta_{1,2}^* = (\tanh(\text{arctanh}(\tau_{1,2}^*) + \frac{m_{\text{inv}} m_{\text{spur}}}{m_{\text{inv}} + m_{\text{spur}} \tau_{1,2}^2}) + 1)/2$ and

$$\tau_{1,2}^* = \sqrt{\frac{m_{\text{inv}}(m_{\text{inv}} + m_{\text{spur}} - 2 \pm \sqrt{(m_{\text{inv}} - m_{\text{spur}})^2 - 8(m_{\text{inv}} + m_{\text{spur}})})}{2m_{\text{spur}}(1 + m_{\text{inv}})}}, \tau_1 < \tau_2 \tag{36}$$

(iii) The three classifiers are achieved by $g(\zeta_{\text{inv}}^*) \in (0, \tau_{\text{inv}}^*), g(\zeta_2^*) \in (\tau_{\text{inv}}^*, \tau_2^*), g(\zeta_3^*) \in (\tau_2^*, 1)$, respectively.

proof. Note that $\omega = \text{arctanh}(\tau) + \frac{m_{\text{inv}} m_{\text{spur}} \tau}{m_{\text{inv}} + m_{\text{spur}} \tau^2}$, thus we can see that ω is monotonically increasing in $(0, \tau_1) \cup (\tau_2, 1)$, and monotonically decreasing in (τ_1, τ_2), where $\tau_{1,2}$ are the roots in appendix D. As a consequence, given $\omega > 0$, there is only one possible τ (i.e. only one classifier) if $\omega \in (0, \omega_2) \cup (\omega_1, 1)$, and there are three possible τ (i.e. three different optimal classifiers) when $\omega \in (\omega_2, \omega_1)$. Here $\omega_{1,2} = \text{arctanh}(\tau_{1,2}) + \frac{m_{\text{inv}} m_{\text{spur}} \tau_{1,2}}{m_{\text{inv}} + m_{\text{spur}} \tau_{1,2}^2}$. Since ζ and ω are 1-to-1 mapped, we have $\zeta_{1,2}^* = (\tanh(\omega_{1,2}) + 1)/2 = (\tanh(\text{arctanh}(\tau_{1,2}^*) + \frac{m_{\text{inv}} m_{\text{spur}}}{m_{\text{inv}} + m_{\text{spur}} \tau_{1,2}^2}) + 1)/2$. Finally, from appendix D, the two roots of τ are

$$\tau_{1,2} = \sqrt{\frac{m_{\text{inv}}(m_{\text{inv}} + m_{\text{spur}} - 2 \pm \sqrt{(m_{\text{inv}} - m_{\text{spur}})^2 - 8(m_{\text{inv}} + m_{\text{spur}})})}{2m_{\text{spur}}(1 + m_{\text{inv}})}}, \tau_1 < \tau_2 \tag{37}$$

Counterintuitively, these results suggest that when the spurious features are a lot more separable compared to the informative features, and highly correlated to the labels at the same time, the optimal models can be achieved by either 3 different ratios between the weights of informative features and spurious features. Here, since $\boldsymbol{\mu}, \Sigma$ are fixed, $\frac{\|\boldsymbol{w}_2\|}{\|\boldsymbol{w}_{\text{inv}}\|} = g(\zeta_i^*) \cdot \texttt{Const}, i = 1, 2, 3$. And ideally, it is expected that the ratio $\frac{\|\boldsymbol{w}_2\|}{\|\boldsymbol{w}_{\text{inv}}\|}$ to be smaller, indicating that the informative features are made sufficient use of. In $(1-\zeta_{\text{inv}}^*, \zeta_{\text{inv}}^*)$, the informative features are dominant in the model decisions. As ζ keeps increasing/decreasing to $\zeta \in (1-\zeta_2^*, 1-\zeta_{\text{inv}}^*) \cup (\zeta_{\text{inv}}^*, \zeta_2^*)$, the spurious feature is much more separable than the informative feature, but the correlation to the labeling is not perfect, resulting three possible optimal models that rely on the informative/spurious features differently. Finally, when $\zeta \in (0, 1-\zeta_2^*) \cup (\zeta_2^*, 1)$, the spurious feature becomes dominant in the decision process, since the spurious correlation is close to 1 and the spurious feature is more separable.

F Proof of Lemma 3. and Theorem 4.

Lemma 3. (**Optimal Accuracy.**) Let $\zeta_{\text{tr}}, \zeta_{\text{te}}$ be the correlation ratios in the training and testing data. And let the $m_{\text{inv}}, m_{\text{spur}}$ be the Mahalanobis distance of the invariant and spurious features and satisfy $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}}+1} + 4$. Then the training and testing accuracy of the optimal classifier can be written as:

$$\begin{aligned} A(\zeta_{\text{tr}}) =& \frac{1}{2}\big[1 + \zeta_{\text{tr}} R\big(g(\zeta_{\text{tr}})\big) + r\big(g(\zeta_{\text{tr}})\big)\big] \\ A(\zeta_{\text{te}}; \zeta_{\text{tr}}) =& \frac{1}{2}\big[1 + \zeta_{\text{te}} R\big(g(\zeta_{\text{tr}})\big) + r\big(g(\zeta_{\text{tr}})\big)\big] \\ \text{where} \quad & \begin{cases} R(\tau) = \text{erf}\Big(\frac{m_{\text{inv}} + \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\Big) - \text{erf}\Big(\frac{m_{\text{inv}} - \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\Big) \\ r(y) = \text{erf}\Big(\frac{m_{\text{inv}} - \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\Big) \end{cases} \end{aligned} \tag{38}$$

proof. From appendix A, the general form of the training and testing accuracy can be written as

$$\begin{cases} A(\gamma^{\text{tr}}) = \frac{1}{2}\Big(1 + \sum\limits_{\boldsymbol{a} \in \{\pm 1\}^N} \gamma_{\boldsymbol{a}}^{\text{tr}} \text{erf}\Big(\frac{\boldsymbol{\mu}_{\boldsymbol{a}}^T \boldsymbol{w}^*}{\sqrt{2(\boldsymbol{w}^*)^T \Sigma \boldsymbol{w}^*}}\Big)\Big) \\ A(\gamma^{\text{te}}; \gamma^{\text{tr}}) = \frac{1}{2}\Big(1 + \sum\limits_{\boldsymbol{a} \in \{\pm 1\}^N} \gamma_{\boldsymbol{a}}^{\text{te}} \text{erf}\Big(\frac{\boldsymbol{\mu}_{\boldsymbol{a}}^T \boldsymbol{w}^*}{\sqrt{2(\boldsymbol{w}^*)^T \Sigma \boldsymbol{w}^*}}\Big)\Big) \end{cases} \tag{39}$$

where $\boldsymbol{w}^*$ is the optimal classifier w.r.t. the training correlation γ^{tr}. When $N = 2$ and $\alpha_{\text{inv}} = \beta_{\text{inv}} = 1$, we have

$$\begin{cases} \gamma_{(0,0)} = \dfrac{(1-\alpha_{\text{inv}})(1-\alpha_2) + (1-\beta_{\text{inv}})(1-\beta_2)}{2} = 0 \\ \gamma_{(1,0)} = \dfrac{\alpha_{\text{inv}}(1-\alpha_2) + \beta_{\text{inv}}(1-\beta_2)}{2} = 1 - \dfrac{\alpha_2+\beta_2}{2} =: 1-\zeta \\ \gamma_{(0,1)} = \dfrac{(1-\alpha_{\text{inv}})\alpha_2 + (1-\beta_{\text{inv}})\beta_2}{2} = 0 \\ \gamma_{(1,1)} = \dfrac{\alpha_{\text{inv}}\alpha_2 + \beta_{\text{inv}}\beta_2}{2} =: \zeta \end{cases} \tag{40}$$

Note that the optimal classifier can be written as $\boldsymbol{w}^*_{\text{inv}} = \Sigma_{\text{inv}}^{-1}\boldsymbol{\mu}_{\text{inv}}, \boldsymbol{w}^*_{\text{inv}} = \tau\Sigma_2^{-1}\boldsymbol{\mu}_{\text{spur}}$. Substituting the classifier and the ratio back by τ and ζ, we have the training accuracy formula as

$$A(\zeta_{\text{tr}}) = \frac{1}{2}\Big(1 + (1-\zeta_{\text{tr}})\text{erf}\big(\frac{m_{\text{inv}} - \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\big) + \zeta_{\text{tr}}\text{erf}\big(\frac{m_{\text{inv}} + \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\big)\Big) \tag{41}$$

$$= \frac{1}{2}\Big(1 + \zeta_{\text{tr}}\big(\text{erf}\big(\frac{m_{\text{inv}} + \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\big) - \text{erf}\big(\frac{m_{\text{inv}} - \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\big)\big) \tag{42}$$

$$+ \text{erf}\big(\frac{m_{\text{inv}} - \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\big)\Big) \tag{43}$$

$$= \frac{1}{2}(1 + \zeta_{\text{tr}} R(g(\zeta_{\text{tr}})) + r(g(\zeta_{\text{tr}}))) \tag{44}$$

Similarly, the formula of the testing accuracy is

$$A(\zeta_{\text{te}}; \zeta_{\text{tr}}) = \frac{1}{2}(1 + \zeta_{\text{te}} R(g(\zeta_{\text{tr}})) + r(g(\zeta_{\text{tr}}))) \tag{45}$$

Here R, r are the functions defined in the statement of the lemma.□

Theorem 4. Given the Mahalanobis distances of the two features $m_{\text{inv}}, m_{\text{spur}} > 0$ such that $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4$, and the training correlations $\zeta_{\text{tr}}, \zeta'_{\text{tr}} \in (0,1)$, the performance shift over the testing set with correlation ratio $\zeta_{\text{te}} \in (0,1)$ is bounded by

$$|A(\zeta_{\text{te}}; \zeta_{\text{tr}}) - A(\zeta_{\text{te}}; \zeta'_{\text{tr}})| \le \frac{m_{\text{spur}}}{2\sqrt{2\pi m_{\text{inv}}}} \frac{M}{\zeta(1-\zeta)}(\zeta_{\text{te}} + 2)|\zeta_{\text{tr}} - \zeta'_{\text{tr}}| \tag{46}$$

where ζ is between $\zeta_{\text{tr}}, \zeta'_{\text{tr}}$ and $M > 0$ is a constant.

proof. When $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4$, by implicit function, we have $g \in \mathcal{C}^{infty}$ s.t. $g(\omega) = \tau$ for $\forall \omega \in \mathbb{R}$ and the derivative is

$$\frac{\mathrm{d}\tau}{\mathrm{d}\omega}(\omega_0, \tau_0) = -\frac{\frac{\partial\phi}{\partial x}(\omega_0, \tau_0)}{\frac{\partial\phi}{\partial\tau}(\omega_0, \tau_0)} = -\frac{\text{sech}^2(\omega - \frac{m_{\text{inv}} m_{\text{spur}}}{m_{\text{inv}} + m_{\text{spur}}\tau^2})}{\frac{m_{\text{inv}} m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})\text{sech}^2(\omega - \frac{m_{\text{inv}} m_{\text{spur}}}{m_{\text{inv}} + m_{\text{spur}}\tau^2})}{(m_{\text{inv}} + m_{\text{spur}}\tau^2)^2} - 1} \tag{47}$$

$$= \frac{(m_{\text{inv}} + m_{\text{spur}}\tau^2)^2(1-\tau^2)}{(m_{\text{inv}} + m_{\text{spur}}\tau^2)^2 - m_{\text{inv}} m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})(1-\tau^2)} \tag{48}$$

$$= \frac{4m_{\text{spur}}^2(\frac{m_{\text{inv}}}{2m_{\text{spur}}} + \frac{\tau^2}{2})(\frac{m_{\text{inv}}}{2m_{\text{spur}}} + \frac{\tau^2}{2})(1-\tau^2)}{m_{\text{spur}}^2(1+m_{\text{inv}})\tau^4 + m_{\text{inv}} m_{\text{spur}}(2 - m_{\text{inv}} - m_{\text{spur}})\tau^2 + m_{\text{inv}}^2(1+m_{\text{spur}}^2)} \tag{49}$$

$$\le \frac{4m_{\text{spur}}^2\big(\frac{\frac{m_{\text{inv}}}{2m_{\text{spur}}} + \frac{\tau^2}{2} + \frac{m_{\text{inv}}}{2m_{\text{spur}}} + \frac{\tau^2}{2} + 1 - \tau^2}{3}\big)^3}{\frac{m_{\text{inv}}^2 m_{\text{spur}}^2(8(m_{\text{inv}} + m_{\text{spur}}) - (m_{\text{inv}} - m_{\text{spur}})^2)}{4m_{\text{spur}}^2(1+m_{\text{inv}})}} \tag{50}$$

$$= \frac{16(m_{\text{inv}} + m_{\text{spur}})^3(m_{\text{inv}} + 1)}{27 m_{\text{inv}}^2 m_{\text{spur}}(8(m_{\text{inv}} + m_{\text{spur}}) - (m_{\text{inv}} - m_{\text{spur}})^2)} =: M \tag{51}$$

Therefore

$$|\varphi'(\zeta)| = |\frac{\mathrm{d}\tau}{\mathrm{d}\omega}\frac{\mathrm{d}\omega}{\mathrm{d}\zeta}| = M|\frac{\mathrm{d}\omega}{\mathrm{d}\zeta}| = \frac{M}{2\zeta(1-\zeta)} \tag{52}$$

On the other hand, note that it is imposed that $\tau \in (-1,1)$. Then for $R(\tau) = \mathrm{erf}(\frac{m_{\mathrm{inv}}+\tau m_{\mathrm{spur}}}{\sqrt{2(m_{\mathrm{inv}}+\tau^2 m_{\mathrm{spur}})}}) - \mathrm{erf}(\frac{m_{\mathrm{inv}}-\tau m_{\mathrm{spur}}}{\sqrt{2(m_{\mathrm{inv}}+\tau^2 m_{\mathrm{spur}})}})$, we have

$$\begin{aligned}
|R'(\tau)| =& \frac{2}{\sqrt{\pi}}\Big| \exp\big(-(\frac{m_{\mathrm{inv}}+\tau m_{\mathrm{spur}}}{\sqrt{2(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)}})^2\big)\frac{m_{\mathrm{inv}}m_{\mathrm{spur}}(\tau-1)}{\sqrt{2}(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)^{3/2}} - && (53)\\
& \exp\big(-(\frac{m_{\mathrm{inv}}-\tau m_{\mathrm{spur}}}{\sqrt{2(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)}})^2\big)\frac{m_{\mathrm{inv}}m_{\mathrm{spur}}(\tau+1)}{\sqrt{2}(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)^{3/2}}\Big| && (54)\\
\le& \frac{\sqrt{2}}{\sqrt{\pi}}\big(|\frac{m_{\mathrm{inv}}m_{\mathrm{spur}}(\tau-1)}{(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)^{3/2}}| + |\frac{m_{\mathrm{inv}}m_{\mathrm{spur}}(\tau+1)}{(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)^{3/2}}|\big) && (55)\\
=& \sqrt{\frac{2}{\pi}}\frac{m_{\mathrm{inv}}m_{\mathrm{spur}}}{(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)^{3/2}}\big(|\tau-1|+|\tau+1|\big) && (56)\\
=& \sqrt{\frac{2}{\pi}}\frac{m_{\mathrm{inv}}m_{\mathrm{spur}}}{(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)^{3/2}} \le \sqrt{\frac{2}{\pi}}\frac{m_{\mathrm{inv}}m_{\mathrm{spur}}}{m_{\mathrm{inv}}^{3/2}} = m_{\mathrm{spur}}\sqrt{\frac{2}{\pi m_{\mathrm{inv}}}} && (57)
\end{aligned}$$

and similarly

$$\begin{aligned}
|r'(\tau)| =& \frac{2}{\sqrt{\pi}}\Big| \exp\big(-(\frac{m_{\mathrm{inv}}-\tau m_{\mathrm{spur}}}{\sqrt{2(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)}})^2\big)\frac{m_{\mathrm{inv}}m_{\mathrm{spur}}(\tau+1)}{\sqrt{2}(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)^{3/2}}\Big| && (58)\\
\le& \sqrt{\frac{2}{\pi}}\Big|\frac{m_{\mathrm{inv}}m_{\mathrm{spur}}(\tau+1)}{(m_{\mathrm{inv}}+m_{\mathrm{spur}}\tau^2)^{3/2}}\Big| \le 2m_{\mathrm{spur}}\sqrt{\frac{2}{\pi m_{\mathrm{inv}}}} && (59)
\end{aligned}$$

Finally, the accuracy shift can be written as

$$\begin{aligned}
& |A(\zeta_{\mathrm{te}};\zeta_{\mathrm{tr}}) - A(\zeta_{\mathrm{te}};\zeta'_{\mathrm{tr}})| && (60)\\
=& |\frac{1}{2}(1+\zeta_{\mathrm{te}}R(\varphi(\zeta_{\mathrm{tr}})) + r(\varphi(\zeta_{\mathrm{tr}}))) - \frac{1}{2}(1+\zeta_{\mathrm{te}}R(\varphi(\zeta'_{\mathrm{tr}})) + r(\varphi(\zeta'_{\mathrm{tr}})))| && (61)\\
=& |\frac{\zeta_{\mathrm{te}}}{2}(R(\varphi(\zeta_{\mathrm{tr}})) - R(\varphi(\zeta'_{\mathrm{tr}}))) + \frac{1}{2}(r(\varphi(\zeta_{\mathrm{tr}})) - r(\varphi(\zeta'_{\mathrm{tr}})))| && (62)\\
\le& \frac{\zeta_{\mathrm{te}}}{2}|R(\varphi(\zeta_{\mathrm{tr}})) - R(\varphi(\zeta'_{\mathrm{tr}}))| + \frac{1}{2}|r(\varphi(\zeta_{\mathrm{tr}})) - r(\varphi(\zeta'_{\mathrm{tr}}))| && (63)\\
\le& \frac{\zeta_{\mathrm{te}}}{2}\xi_R\xi_{\varphi,1}|\zeta_{\mathrm{tr}} - \zeta'_{\mathrm{tr}}| + \frac{1}{2}\xi_r\xi_{\varphi,2}|\zeta_{\mathrm{tr}} - \zeta'_{\mathrm{tr}}| && (64)\\
\le& \big(\frac{\zeta_{\mathrm{te}}}{2}\cdot m_{\mathrm{spur}}\sqrt{\frac{2}{\pi m_{\mathrm{inv}}}}\cdot\frac{M}{2\zeta(1-\zeta)} + m_{\mathrm{spur}}\sqrt{\frac{2}{\pi m_{\mathrm{inv}}}}\cdot\frac{M}{2\zeta(1-\zeta)}\big)|\zeta_{\mathrm{tr}} - \zeta'_{\mathrm{tr}}| && (65)\\
=& \frac{m_{\mathrm{spur}}}{2\sqrt{2\pi m_{\mathrm{inv}}}}\frac{M}{\zeta(1-\zeta)}(\zeta_{\mathrm{te}}+2)|\zeta_{\mathrm{tr}} - \zeta'_{\mathrm{tr}}| && (66)
\end{aligned}$$

G Proof of Theorem 5.

Theorem. 5 (i) When $m_{\mathrm{spur}} < m_{\mathrm{inv}} + 4\sqrt{m_{\mathrm{inv}}+1} + 4$, $\frac{\tau}{m_1}$ decreases monotonically with respect to m_1. (ii) When $m_{\mathrm{spur}} \le 2(\sqrt{m_{\mathrm{inv}}+1}+2)$, τm_2 increases monotonically with respect to m_2.

proof. Since we only consider the scenario where $\zeta > 0.5$, which means $\omega > 0$. By symmetry, $\tau \ge 0$. (i) It suffices

to show that $\frac{\mathrm{d}\tau}{\mathrm{d}m_{\text{inv}}} < 0$. Since $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}}+1} + 4$, from appendix D, we know that $|\frac{\partial\phi}{\partial\tau}| > 0$. Thus

$$\frac{\mathrm{d}\tau}{\mathrm{d}m_{\text{inv}}} = -\frac{-\frac{m_{\text{spur}}^2\tau^3\text{sech}^2(\omega-\frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}}+m_{\text{spur}}\tau^2})}{(m_{\text{inv}}+m_{\text{spur}}\tau^2)^2}}{\frac{m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2-m_{\text{inv}})\text{sech}^2(\omega-\frac{m_{\text{inv}}m_{\text{spur}}}{m_{\text{inv}}+m_{\text{spur}}\tau^2})}{(m_{\text{inv}}+m_{\text{spur}}\tau^2)^2} - 1} \tag{67}$$

$$= \frac{m_{\text{spur}}^2\tau^3\text{sech}^2(\omega - \frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}}+m_{\text{spur}}\tau^2})}{m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})\text{sech}^2(\omega - \frac{m_{\text{inv}}m_{\text{spur}}}{m_{\text{inv}}+m_{\text{spur}}\tau^2}) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2} \tag{68}$$

where the numerator is positive, and the denominator can be written as

$$m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})\text{sech}^2(\omega - \frac{m_{\text{inv}}m_{\text{spur}}}{m_{\text{inv}} + m_{\text{spur}}\tau^2}) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2 \tag{69}$$

$$= m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})(1-\tau^2) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2 \tag{70}$$

where the discriminant is

$$\Delta = (m_{\text{inv}} - m_{\text{spur}})^2 - 8(m_{\text{inv}} + m_{\text{spur}}) < 0 \tag{71}$$

Therefore, the determinant is negative, resulting in $\frac{\mathrm{d}\tau}{\mathrm{d}m_{\text{inv}}} < 0$.

(ii) Similarly, we can write $\frac{\mathrm{d}\tau}{\mathrm{d}m_{\text{spur}}}$ as

$$\frac{\mathrm{d}\tau}{\mathrm{d}m_{\text{spur}}} = -\frac{-\frac{m_{\text{inv}}^2\tau\text{sech}^2(\omega-\frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}}+m_{\text{spur}}\tau^2})}{(m_{\text{inv}}+m_{\text{spur}}\tau^2)^2}}{\frac{m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2-m_{\text{inv}})\text{sech}^2(\omega-\frac{m_{\text{inv}}m_{\text{spur}}}{m_{\text{inv}}+m_{\text{spur}}\tau^2})}{(m_{\text{inv}}+m_{\text{spur}}\tau^2)^2} - 1} \tag{72}$$

$$= \frac{m_{\text{inv}}^2\tau\text{sech}^2(\omega - \frac{m_{\text{inv}}m_{\text{spur}}\tau}{m_{\text{inv}}+m_{\text{spur}}\tau^2})}{m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})\text{sech}^2(\omega - \frac{m_{\text{inv}}m_{\text{spur}}}{m_{\text{inv}}+m_{\text{spur}}\tau^2}) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2} \tag{73}$$

$$= \frac{m_{\text{inv}}^2\tau(1-\tau^2)}{m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})(1-\tau^2) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2} > 0 \tag{74}$$

Hence

$$\frac{\mathrm{d}\tau m_{\text{spur}}}{\mathrm{d}m_{\text{spur}}} = \tau + \frac{\mathrm{d}\tau}{\mathrm{d}m_{\text{spur}}}m_{\text{spur}} \tag{75}$$

$$= \tau + \frac{m_{\text{inv}}^2 m_{\text{spur}}\tau(1-\tau^2)}{m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})(1-\tau^2) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2} \tag{76}$$

$$= \frac{m_{\text{inv}}^2 m_{\text{spur}}\tau(1-\tau^2) + \tau(m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})(1-\tau^2) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2)}{m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})(1-\tau^2) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2} \tag{77}$$

where the denominator is negative. And the numerator can be written as

$$\texttt{num} = m_{\text{inv}}^2 m_{\text{spur}}\tau(1-\tau^2) + \tau(m_{\text{inv}}m_{\text{spur}}(m_{\text{spur}}\tau^2 - m_{\text{inv}})(1-\tau^2) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2) \tag{78}$$

$$= m_{\text{inv}}m_{\text{spur}}\tau(1-\tau^2)[m_{\text{inv}} + (m_{\text{spur}}\tau^2 - m_{\text{inv}})] - \tau(m_{\text{inv}} + m_{\text{spur}}\tau^2)^2 \tag{79}$$

$$= \tau\Big[m_{\text{inv}}m_{\text{spur}}^2\tau^2(1-\tau^2) - (m_{\text{inv}} + m_{\text{spur}}\tau^2)^2\Big] \tag{80}$$

Therefore, it can be found that the roots of num are ($\tau = 0$ is trivial)

$$\tau = \sqrt{\frac{m_{\text{inv}}m_{\text{spur}}((m_{\text{spur}}-2) \pm \sqrt{m_{\text{spur}}^2 - 4m_{\text{spur}} - 4m_{\text{inv}}})}{2(m_{\text{inv}}+1)m_{\text{spur}}^2}} \tag{81}$$

Thus when $m_{\text{spur}} \in (0, 2(\sqrt{m_{\text{inv}}+1}+1)]$, $\Delta \le 0$, $\texttt{num} \le 0$, and hence $\frac{\mathrm{d}\tau m_{\text{spur}}}{\mathrm{d}m_{\text{spur}}} \ge 0$.

Table 1: The averaged accuracy over ζ_{tr}s: `[0.999, 0.998, 0.995, 0.99, 0.98, 0.95, 0.9, 0.8]` and 10 rounds.

Dataset	Model	Oringin	Regularization	Group-DRO
waterbird-ζ	ResNet	65.46% ± 9.54%	68.69% ± 10.69%	66.15% ± 10.05%
	VGG	68.70% ± 10.73%	72.54% ± 11.47%	70.70% ± 10.49%
	AlexNet	60.62% ± 6.62%	61.61% ± 7.80%	60.88% ± 6.76%
CIFAR-concate	ResNet	72.80% ± 2.60%	76.37% ± 3.91%	76.35% ± 4.75%
	VGG	83.13% ± 8.25%	86.59% ± 5.74%	86.08% ± 7.57%
	AlexNet	77.83% ± 5.27%	81.97% ± 6.22%	80.88% ± 5.76%
CIFAR-watermark	ResNet	91.98% ± 1.16%	93.20% ± 1.24%	92.34% ± 1.82%
	VGG	91.77% ± 4.36%	94.98% ± 3.13%	94.26% ± 3.40%
	AlexNet	80.73% ± 8.18%	88.97% ± 3.85%	83.87% ± 7.57%
CelebA	ResNet	96.45% ± 1.98%	98.39% ± 0.81%	97.91% ± 1.10%
	VGG	96.72% ± 1.93%	98.60% ± 0.62%	98.21% ± 0.87%
	AlexNet	96.68% ± 1.46%	98.20% ± 0.75%	97.88% ± 0.85%

H Datasets

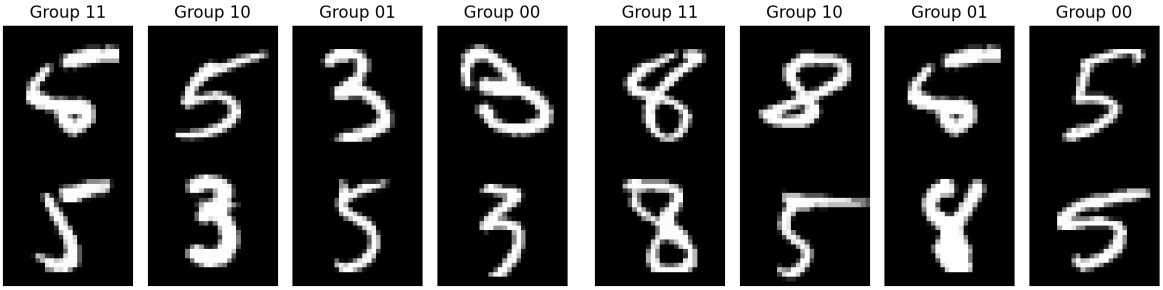
The datasets that are used in the experiments are demonstrated in fig. 6. In the concatenation datasets (a)(b)(c)(d), the upper block represents the informative feature $\boldsymbol{z}_{\text{inv}}$, which is strictly related to the ground truth label. And the lower block represents the informative feature $\boldsymbol{z}_2$. For the watermark dataset (e), the informative feature $\boldsymbol{z}_{\text{inv}}$ is the original image, while the spurious feature $\boldsymbol{z}_2$ is the presence of the black square watermark at the bottom right corner. As for the waterbird-ζ dataset (f), the informative feature is the taxa of the birds while the spurious feature is the background habitats. The subset version of CelebA. As shown in (g).

I The Algorithm for Estimating the Trend of Testing Accuracy w.r.t. the Ratio

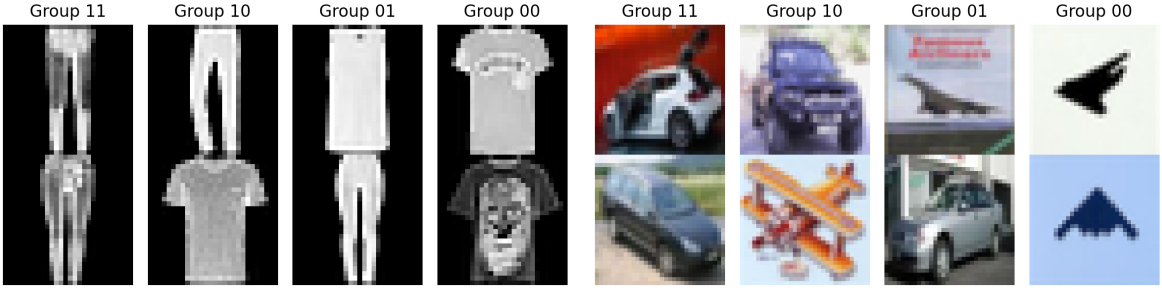
Note that from the theorems, the testing accuracy A is decided by $m_{\text{inv}}, m_{\text{spur}}$ and ζ_{tr} through eqs. (4) and (5). Therefore, now that the Mahalanobis distances $m_{\text{inv}}, m_{\text{spur}}$ are unknown for the model, they can be estimated using two ratio-accuracy pairs $(\zeta_{\text{tr}}^1, A_1), (\zeta_{\text{tr}}^2, A_2)$. The corresponding pseudo-Mahalanobis distances $\hat{m}_{\text{inv}}, \hat{m}_{\text{spur}}$ can be estimated as parts of the roots $(m_{\text{inv}}, m_{\text{spur}}, \tau_1, \tau_2)$ of the system

$$\begin{cases} F_1(m_{\text{inv}}, m_{\text{spur}}, \tau_1, \tau_2) = \operatorname{erf}\big(\frac{m_{\text{inv}} + m_{\text{spur}}\tau_1}{\sqrt{2(m_{\text{inv}} + m_{\text{spur}}\tau_1^2}}\big) + \operatorname{erf}\big(\frac{m_{\text{inv}} - m_{\text{spur}}\tau_1}{\sqrt{2(m_{\text{inv}} + m_{\text{spur}}\tau_1^2}}\big) - 4A_1 + 2 = 0 \\ F_2(m_{\text{inv}}, m_{\text{spur}}, \tau_1, \tau_2) = \tau_1 - \tanh\big(\frac{1}{2}\log(\frac{\zeta_{\text{tr}}^1}{1-\zeta_{\text{tr}}^1}) - \frac{m_{\text{inv}} m_{\text{spur}}\tau_1}{m_{\text{inv}} + m_{\text{spur}}\tau_1^2}\big) = 0 \\ F_3(m_{\text{inv}}, m_{\text{spur}}, \tau_1, \tau_2) = \operatorname{erf}\big(\frac{m_{\text{inv}} + m_{\text{spur}}\tau_2}{\sqrt{2(m_{\text{inv}} + m_{\text{spur}}\tau_1^2}}\big) + \operatorname{erf}\big(\frac{m_{\text{inv}} - m_{\text{spur}}\tau_2}{\sqrt{2(m_{\text{inv}} + m_{\text{spur}}\tau_1^2}}\big) - 4A_2 + 2 = 0 \\ F_4(m_{\text{inv}}, m_{\text{spur}}, \tau_1, \tau_2) = \tau_2 - \tanh\big(\frac{1}{2}\log(\frac{\zeta_{\text{tr}}^1}{1-\zeta_{\text{tr}}^1}) - \frac{m_{\text{inv}} m_{\text{spur}}\tau_2}{m_{\text{inv}} + m_{\text{spur}}\tau_2^2}\big) = 0 \end{cases} \tag{82}$$

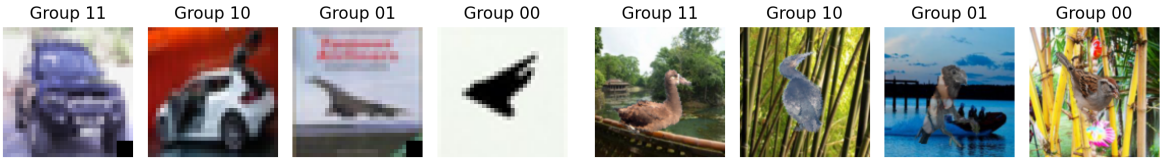
In experiments, this is solved using the `scipy` toolkit to minimize $\sum_{i=1}^4 F_i^2$. Afterwards, the pseudo-Mahalanobis distances $\hat{m}_{\text{inv}}, \hat{m}_{\text{spur}}$ can be substitute back to eqs. (4) and (5) to predict the testing accuracy with given ζ_{tr}s. The part of experimental results is already shown in the main manuscript due to the space limit. Here we demonstrate the results of all tested models and datasets in fig. 7. The first column (a)(e)(i) contains MNIST and FMNIST datasets with linear models. The 2nd-4th rows contain the results with the datasets waterbird-ζ, CIFAR-concate, and CIFAR-watermark respectively. It can be found that the prediction fits the experimental trends of the testing accuracy in all cases except for AlexNet & waterbird-ζ. We deduce that this is due to the limited expressiveness of AlexNet.



(a) MNIST-3 vs 5 (b) MNIST-5 vs 8



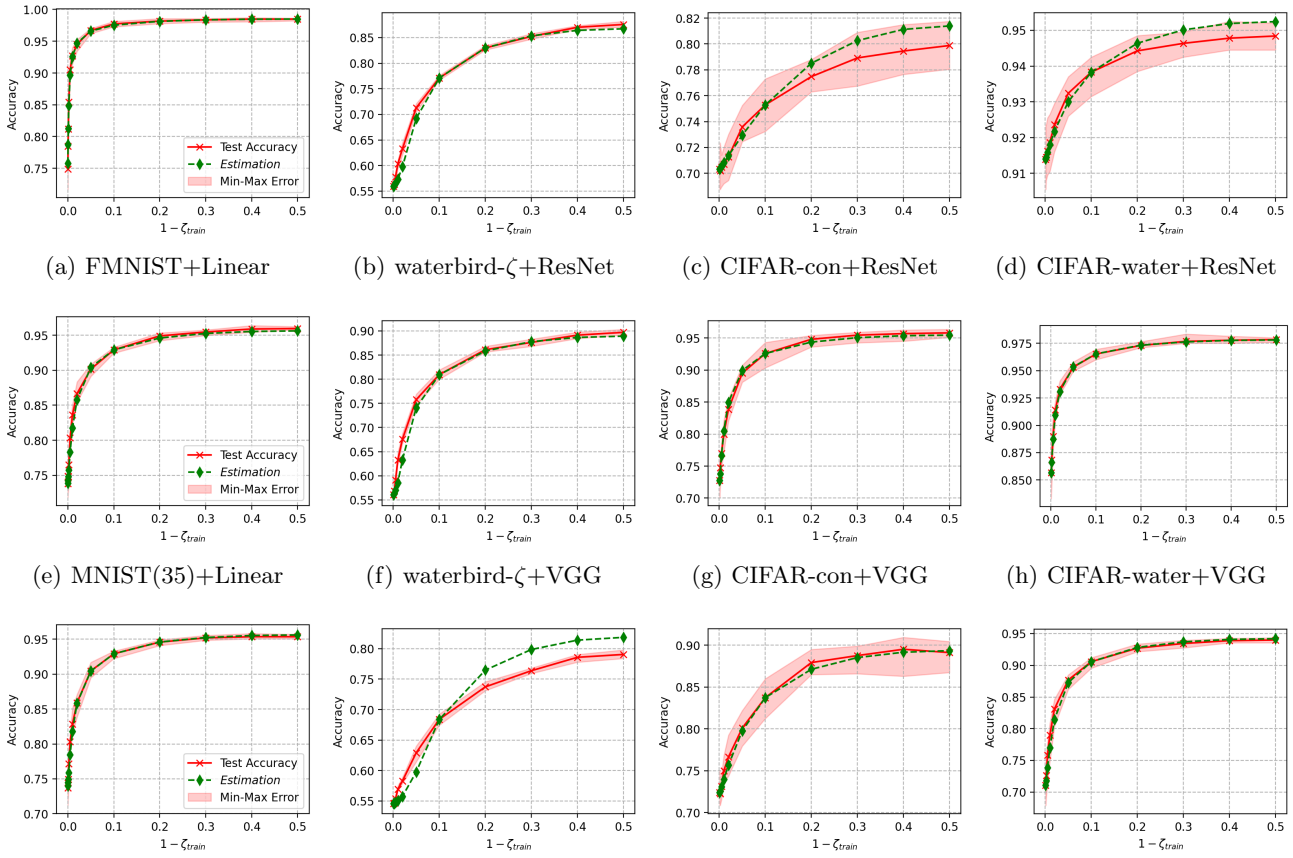
(c) Fashion-MNIST (d) CIFAR-concate



(e) CIFAR-watermark (f) waterbird-ζ

(g) CelebA

Figure 6: A demonstration of the figures of the datasets that are tested. Samples of groups $++$, $+-$, $-+$, and $++$ are shown for each dataset. In the concatenation data, the upper blocks are the informative features, and the lower blocks are the spurious features. In the watermark dataset, the object car vs airplane is the informative feature, while the presence of the watermark in the bottom right corner is the spurious feature. And in the waterbird-ζ dataset, the taxa of the birds is the informative feature, and the habitat shown in the background is the spurious one. Finally, in the CelebA dataset, the hair color black vs. blonde is the informative feature while the gender male vs. female is the spurious feature.



(a) FMNIST+Linear (b) waterbird-ζ+ResNet (c) CIFAR-con+ResNet (d) CIFAR-water+ResNet

(e) MNIST(35)+Linear (f) waterbird-ζ+VGG (g) CIFAR-con+VGG (h) CIFAR-water+VGG

(i) MNIST(58)+Linear (j) waterbird-ζ+AlexNet (k) CIFAR-con+AlexNet (l) CIFAR-water+AlexNet

Figure 7: Accuracy trend w.r.t. the correlation ratio ζ_{tr}. Red solid curves show raw testing accuracy, while green dashed curves show the estimated results based on our theorems. The axis is set as $1-\zeta_{\text{tr}}$ to have a right-to-left trend. The results show that our theorem-based estimation closely matches the actual accuracy trend across multiple datasets and models.

J Supplementary Analysis of the Regularization over the Separability

Tabular Results. As shown in the main manuscript, the effectiveness of the regularization is tested over a list of ζ_{tr}s defined as `[0.999, 0.998, 0.995, 0.99, 0.98, 0.95, 0.9, 0.8]`. The averaged accuracy over tested ζ_{trains} is also demonstrated in table 1. Results demonstrate the fact that the regularization over separability is capable of effectively mitigating the spurious correlation problem.

Remarks. It should be noticed that the spurious correlation problem harms the testing accuracy by overfitting the correlation in the training distribution. Therefore, if the training loss is not minimized (i.e. the training accuracy is still low), the testing accuracy can be higher than the final testing accuracy. That is, with the training process, the training accuracy will increase monotonically, while the testing accuracy increases first, and then decreases when the overfitting happens. Therefore, before the training accuracy approaches 1, there are unstable states where testing accuracy is higher than the accuracy when the training is finalized. Such improvement in the testing accuracy is due to such early stops (before overfitting). In order to avoid these scenarios, and to show the effectiveness of the regularization, we visualize and analyze the detailed training process to demonstrate that the regularization indeed improves the models' robustness even under the overfitting. The training details of the first round (`seed = 0`) on waterbird-ζ are shown in figs. 8 and 9.

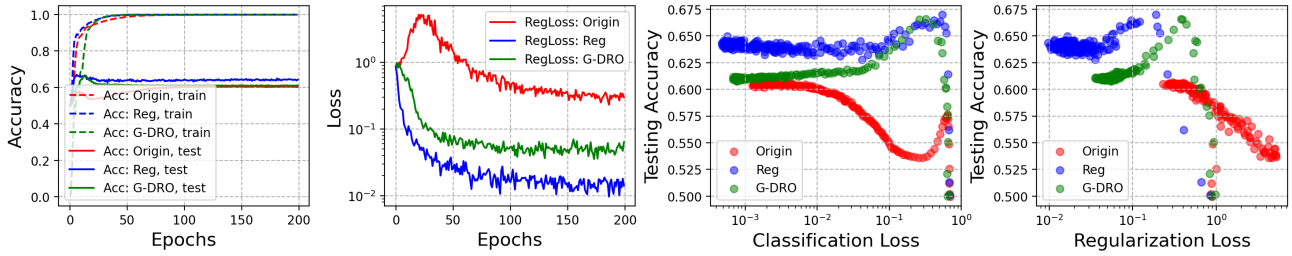
The first column demonstrates the training accuracy and the testing accuracy of origin (Origin), regularization (Reg), and Group-DRO (G-DRO). This demonstrates that all models reach the finalized state where the training accuracy converges to 1. As can be found around the 20th-30th iteration, there are bumps in the testing accuracy

corresponding to the overfitting phenomena.

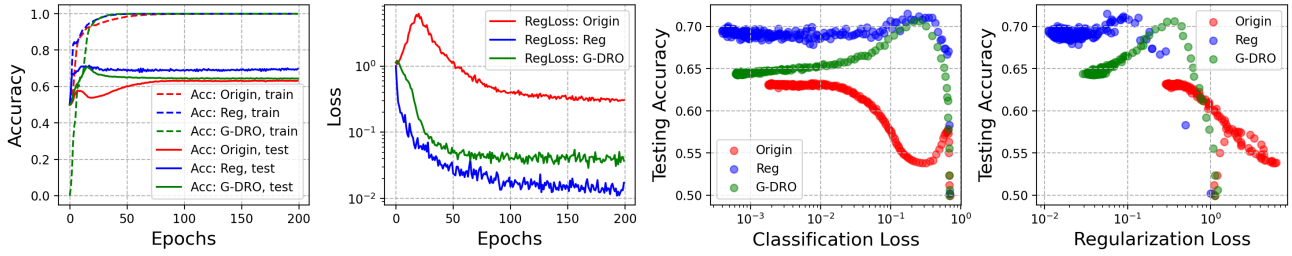
The second column shows the regularization loss-training epochs trend, which is recorded for origin and GroupDRO, too. It shows that (i) even without the regularization, optimized models have the tendency to increase the separability between informative feature groups while decreasing the separability between spurious feature groups. This validates the proposed regularization term. (ii) the regularization method achieves the lowest regularization loss among the three models. (iii) Comparing the regularization loss of the original model (red curves in the second column) and the testing accuracy of the original model (red curves in the first column), we can see that their trends correspond. The small bumps in the regularization loss cause small drops in the testing accuracy around corresponding iterations (around the 20th-30th).

The third column shows the results of testing accuracy vs classification loss. At the same level of classification loss, it is expected that they are of the same level of overfitting. And we can see that the regularization does not prevent the classification loss from dropping in the training process. Instead, it actually improves the separability even when overfitting occurs.

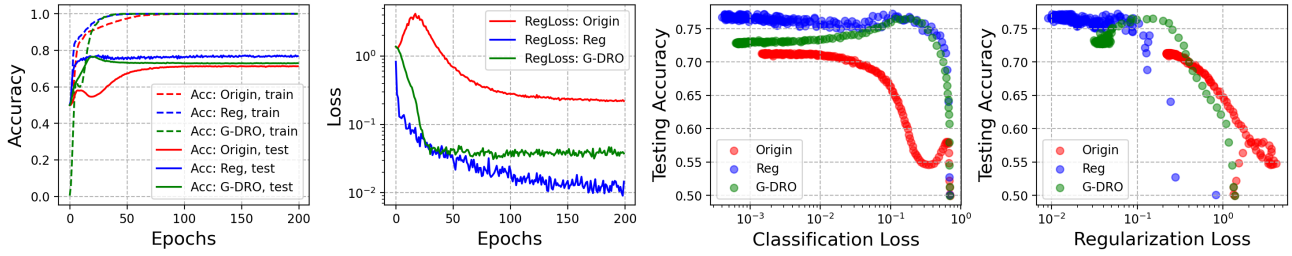
At last, the fourth column shows the results of testing accuracy vs regularization loss. The proposed regularization achieves the lowest regularization loss. Moreover, when ζ_{tr} is close to 1 (which are the difficult extreme cases), accuracy decreases less w.r.t. the decrease of the regularization loss for the proposed method.



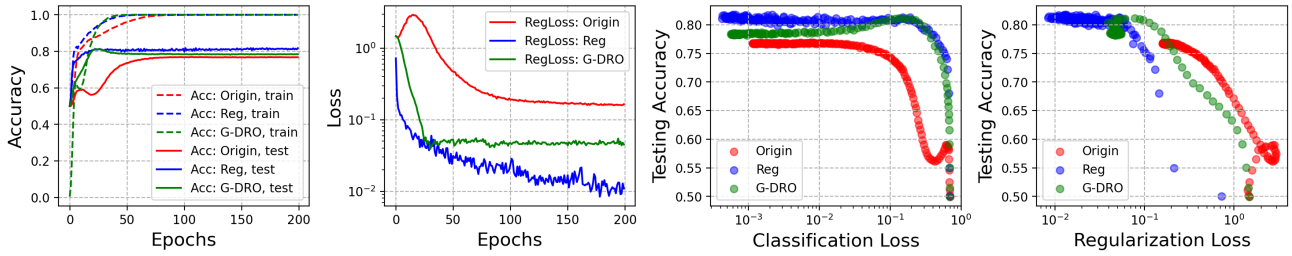
(a) $\zeta_{\text{tr}} = 0.99$



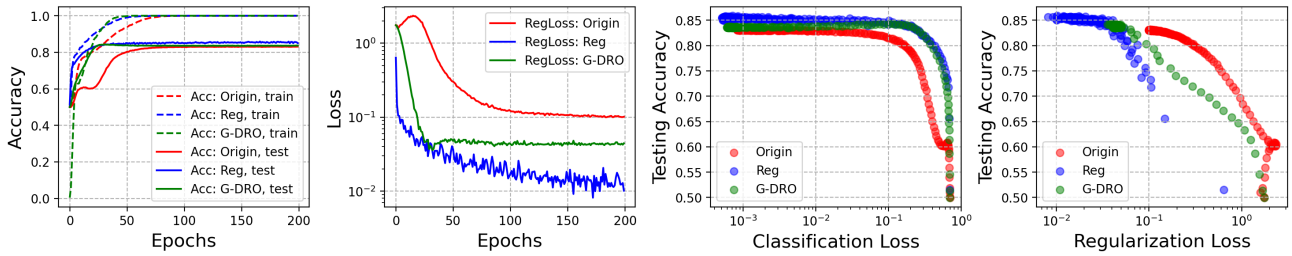
(b) $\zeta_{\text{tr}} = 0.98$



(c) $\zeta_{\text{tr}} = 0.95$

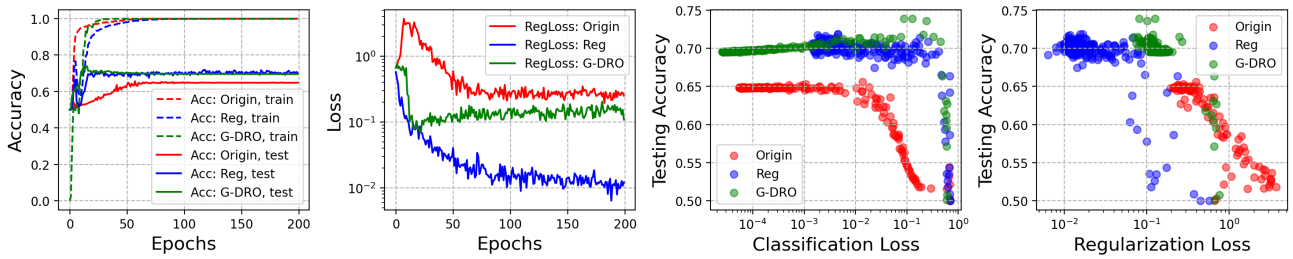


(d) $\zeta_{\text{tr}} = 0.9$



(e) $\zeta_{\text{tr}} = 0.8$

Figure 8: The training details of the first round (`seed=0`) of the waterbird-ζ dataset with ResNet-18. The four columns are (i) (training/testing)Accuracy-Epochs; (ii) Regularization Loss-Epochs; (iii) Testing Accuracy-Classification Loss; (iv) Testing Accuracy-Regularization Loss.



(a) $\zeta_{\text{tr}} = 0.99$

(b) $\zeta_{\text{tr}} = 0.98$

(c) $\zeta_{\text{tr}} = 0.95$

(d) $\zeta_{\text{tr}} = 0.9$

(e) $\zeta_{\text{tr}} = 0.8$

Figure 9: The training details of the first round (`seed=0`) of the waterbird-ζ dataset with ResNet-18. The four columns are (i) (training/testing)Accuracy-Epochs; (ii) Regularization Loss-Epochs; (iii) Testing Accuracy-Classification Loss; (iv) Testing Accuracy-Regularization Loss.