Elastic Representation: Mitigating Spurious Correlations for Group Robustness

Tao Wen
Dartmouth College
tw2672@nyu.edu

Zihan Wang
New York University
zw3508@nyu.edu

Quan Zhang
Michigan State University
quan.zhang@broad.msu.edu

Qi Lei
New York University
ql518@nyu.edu

Abstract

Deep learning models can suffer from severe performance degradation when relying on spurious correlations between input features and labels, making the models perform well on training data but have poor prediction accuracy for minority groups. This problem arises especially when training data are limited or imbalanced. While most prior work focuses on learning invariant features (with consistent correlations to y), it overlooks the potential harm of spurious correlations between features. We hereby propose Elastic Representation (ElRep) to learn features by imposing Nuclear- and Frobenius-norm penalties on the representation from the last layer of a neural network. Similar to the elastic net, ElRep enjoys the benefits of learning important features without losing feature diversity. The proposed method is simple yet effective. It can be integrated into many deep learning approaches to mitigate spurious correlations and improve group robustness. Moreover, we theoretically show that ElRep has minimum negative impacts on in-distribution predictions. This is a remarkable advantage over approaches that prioritize minority groups at the cost of overall performance.

1 INTRODUCTION

Group robustness is critical for deep learning models, particularly when they are deployed in real-world applications like medical imaging and disease diagnosis (Huang et al., 2022; Kirichenko et al., 2023). In practice, data are often limited, and models are frequently exposed to domains or distributions that are not well represented in training data. Group robustness aims to enable models to generalize to unseen domains, addressing challenges such as differing image backgrounds or styles. Standard training procedures, like empirical risk minimization, can result in good performance on average but poor accuracy for certain groups, especially in the presence of spurious correlations (Sagawa et al., 2020; Haghtalab et al., 2022).

Spurious correlations arise when models rely on features that correlate with class labels in the training data but are irrelevant to the true labeling function. This leads to performance degradation for out-of-distribution (OOD) generalization. For example, a model trained to classify objects, like waterbirds and landbirds, might rely on background or textures (Geirhos et al., 2019; Xiao et al., 2021), like water and land, which are irrelevant to the object, resulting in low accuracy for minority groups of waterbirds on land and landbirds on water. Ideally, a deep learning model should learn features that have invariant correlations with labels for all distributions.

While neural-network classification models trained by empirical risk minimization (ERM) may lead to poor group robustness and OOD generalization (Geirhos et al., 2020; Zhang et al., 2022) and be no better than random guessing on minority groups when predictions exclusively depend on spurious features (Shah et al., 2020), recent studies have shown that even standard ERM can well learn both spurious and invariant (non-spurious) features (Kirichenko et al., 2023; Izmailov et al., 2022); the low accuracy of ERM on minority groups results from the classifier, i.e., the linear output layer of a neural network, which tends to overweight spurious features. Based on this finding, we propose Elastic Representation (ElRep) by imposing nuclear-norm and Frobenius-norm penalties on feature representations. This approach not only regularizes the learning of spurious features but also enhances the prominence of invariant features.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

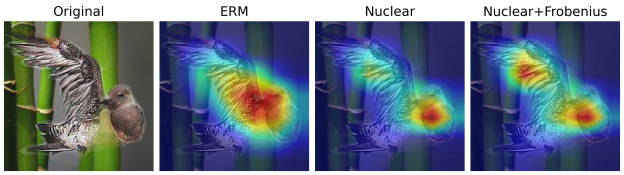

Figure 1: A long-tailed Jaeger, a waterbird on a land background, from the waterbirds dataset (Sagawa et al., 2019). The heat maps depict the pixel contributions to bird type prediction using Grad-CAM (Selvaraju et al., 2019). From left to right are the original image, ERM, ERM with nuclear norm, and ERM with nuclear and Frobenius norms, respectively. ERM learns features including background areas. ERM with nuclear norm focuses on the head, and ERM with both norms evenly emphasizes the head and the wing.

Our approach borrows the idea from the elastic net (Zou and Hastie, 2005) that imposes ℓ_1 and ℓ_2 penalties on regression coefficients. Though we regularize the feature representation rather than the weights of the classifier, the intuition is similar. Specifically, a nuclear norm regularizing the singular values of the representation matrix facilitates a sparse retrieval of the backbone features, and its effectiveness has been underpinned by Shi et al. (2024). However, we have observed that using a nuclear-norm penalty alone can suffer from a problem similar to that of lasso, as it tends to capture only part of the invariant features but omit others if they are highly correlated. The over-regularization can undermine the robustness on minority groups or unseen data where only the omitted features are present.

To address this issue, we introduce a Frobenius-norm penalty to regularize the representation in addition to a nuclear-norm penalty. Analogous to the advantage of the elastic net over lasso, the Frobenius norm tunes down the sparsity and keeps more invariant features when they are correlated. We illustrate this finding in Figure 1 with an image of a waterbird on a land background. ERM without regularization captures features that include the object and background areas. When applying a nuclear norm, the learned features emphasize the bird's head but somewhat overlook the wing. So, the model may fail on images where a bird's head is blocked. With both nuclear and Frobenius norms, the representation captures the head and wing, effectively regularizing the learning of the background and making both invariant features prominent without sacrificing either.

We distinguish ElRep from extant literature by making the following contributions.

1. ElRep mitigates spurious correlation *without relying on group information*, which is often required by many group robustness methods to adjust weights of minority groups. This is essential for real-world applications as group annotations are largely impractical.

2. We theoretically prove and empirically show that ElRep has a minimum sacrifice of the overall performance while *improving the worst-group accuracy*.

3. ElRep is simple yet effective without extra computational cost. It is a *general framework* that can be combined with and further improve many state-of-the-art approaches.

The paper proceeds as follows. In Section 2, We compare ElRep and related work for group robustness. In Section 3, we formally introduce the proposed method. In Section 4, we use synthetic and real data to showcase the outstanding performance and favorable properties of ElRep. Section 5 theoretically underpins ElRep, and Section 6 concludes the paper.

2 RELATED LITERATURE

Extensive efforts have been made to mitigate spurious correlations. Two of the common practices are optimization-based methods addressing group imbalance and via improved learning of invariant features. Our ElRep framework can be combined to improve an optimization-based method. It also supplements the representation learning literature with a much simpler procedure based on the finding that ERM already learns invariant features. We review the literature in these two streams and refer readers to (Ye et al., 2024) for a comprehensive taxonomy of extant popular approaches.

Neural networks relying on spurious correlations often suffer from degradation of performance consistency across different groups or subpopulations. A typical reason is selection biases on datasets (Ye et al., 2024), where groups are not equally represented. Imbalanced groups can lead neural networks to prioritize the majority and learn their spurious correlations that may not hold for the minority. A considerable amount of work addresses group imbalance by utilizing the group information for distributionally robust optimization (DRO) to improve performance in worst cases. For example, groupDRO (Sagawa et al., 2019) minimizes the worst-group loss instead of the average loss, and there emerges subsequent work also emphasizing minority groups in training (e.g., Goel et al., 2020; Levy et al., 2020; Sagawa et al., 2020; Zhang et al.,

2021; Deng et al., 2023). Notably, Idrissi et al. (2022) show that simple group balancing by subsampling or reweighting achieves state-of-the-art accuracy, highlighting the importance of group information.

Though these approaches have improved worst-case accuracy, they rely on group annotations that are often impractical in real-world applications. Methods that automatically identify minority groups are developed. For example, one can use a biased model to find hard-to-classify data, treat them as a minority group, and then use a downstream model to improve the accuracy on the "minority" for group robustness (Nam et al., 2020; Liu et al., 2021; Yenamandra et al., 2023). These approaches train the models twice and may be computationally expensive. To improve the efficiency, Du et al. (2023), Moayeri et al. (2023), and Yang et al. (2024) find data points or groups with high degrees of spuriosity in an early stage of training and then mitigate the model's reliance on them. Overall, the group information, either manually annotated or automatically identified, plays a crucial role in this stream of research that tries to address group imbalance. In stark comparison, ElRep does not require group information and is readily integrated into many of these optimization-based methods to further improve the performance.

Research in representation learning tries to better understand the underlying relationships between variables, capture improved features, and make models more resilient to spurious correlations (e.g., Sun et al., 2021; Veitch et al., 2021; Eastwood et al., 2023). Recent studies (Kirichenko et al., 2023; Izmailov et al., 2022; Rosenfeld et al., 2022; Chen et al., 2023; Zhong et al., 2024) potentially make representation learning easier by showing that ordinary ERM can learn both spurious and invariant feature representation. This implies that one can efficiently improve group robustness by downplaying spurious features and highlighting invariant features, without the need to explore causal relationships, making the process conceptually and computationally much simpler.

Based on this finding, Kirichenko et al. (2023) and Izmailov et al. (2022) retrain the last layer of a neural network on a small held-out dataset where the spurious correlation breaks. However, this method requires the group information. To avoid group annotations, one can combine the idea of automatic identification of "minority groups" and the last-layer fine-tuning. For example, Chen et al. (2023) alternately use easy- and hard-to-classify data to enforce the learning of richer features in the last layer. Similarly, LaBonte et al. (2023) propose using disagreements between the ERM and early-stopped models to balance the classes in the last-layer fine-tuning.

Since ERM can well learn both spurious and invariant features, a natural way for group robustness is to mitigate spurious correlations through regularization. However, this approach has not been thoroughly explored. We fill this research gap by imposing nuclear- and Frobenius-norm penalties to achieve a good balance between pruning spurious features and keeping invariant features. A closely related study (Shi et al., 2024) uses a nuclear-norm regularization for parsimonious representation. However, as illustrated in Figure 1, it may suffer from over-regularization and losing invariant features. ElRep introduces a Frobenius norm to alleviate this problem. Theoretically, this will maintain the in-distribution (ID) performance while making the invariant feature less sparse. Empirically, it outperforms using a nuclear norm alone and further improves state-of-the-art approaches when combined with them.

3 METHODOLOGY

3.1 Preliminaries and Notations

We consider the setting where the domains of training and testing are different. We have $(\boldsymbol{x}, y) \sim \mathcal{D}_{\text{id}}$ for training data and $(\boldsymbol{x}, y) \sim \mathcal{D}_{\text{ood}}$ for test data. The model we consider is $f(\boldsymbol{x}) = W^\top \Phi(\boldsymbol{x})$, where Φ is a latent representation function. Our goal is to train the model with data from $\mathcal{D}_{\text{id}}$ and reduce the risk $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{ood}}}[\ell(f(\boldsymbol{x}), y)]$ on the test domain, where ℓ is a loss function. To achieve this goal, the representation Φ is trained to extract features of the input data. The features that generate data $\boldsymbol{x}$ include invariant and spurious features, with the former only related to the label y and the latter also related to the environment. Since the environment domains are different between the training and testing distributions, a good Φ should preserve invariant features and remove spurious features. We use $\mathcal{L}(W, \Phi)$ to represent some risk function on the training domain with respect to a weight matrix W and representation Φ, where we omit the loss function ℓ. We use $\|\cdot\|_*$ to denote the nuclear norm of a matrix and $\|\cdot\|_{\text{F}}$ the Frobenius norm. Specifically, $\|A\|_* = \text{Tr}\left((A^\top A)^{1/2}\right)$ and $\|A\|_{\text{F}} = \left(\text{Tr}(A^\top A)\right)^{1/2}$. For vectors, $\|\cdot\|_2$ denotes its ℓ_2 norm.

3.2 Elastic Representation

In classification and regression tasks, models learn features from labeled data. In order to make better predictions for OOD data, the model should learn the features that highly correlate to the label. Invariant features should have a higher correlation than spurious features since the former preserves in both ID and OOD data but the latter only appears in ID data. A

latent representation Φ contains both kinds of features. Our goal is to highlight the invariant and eliminate the spurious.

We consider the model $f(\boldsymbol{x}) = W^\top \Phi(\boldsymbol{x})$ with a latent representation $\Phi(\boldsymbol{x})$. By minimizing $\mathcal{L}(W, \Phi)$, we can obtain label-related features. However, the spurious features are also preserved, which cannot help OOD prediction. According to Shi et al. (2024), by adding the nuclear norm of the representation Φ, the information contained in $\Phi(\boldsymbol{x})$ is reduced. This regularization eliminates spurious features but meanwhile, could also rule out part of invariant features. By Elastic Representation (ElRep) that includes an extra Frobenius-norm regularization, we expect to capture more invariant features. The objective function is

$$\min_{W,\Phi} \mathcal{L}(W, \Phi) + \lambda_1 \|\Phi(\boldsymbol{x})\|_* + \lambda_2 \|\Phi(\boldsymbol{x})\|_{\mathrm{F}}^2, \quad (1)$$

where λ_1 and λ_2 are hyper-parameters that control the intensity of the respective penalty. Note that this regularization can be added to a wide range of risk functions, for example ERM and GroupDRO (Sagawa et al., 2019). For ERM, the risk function $\mathcal{L}(W, \Phi) := \mathbb{E}_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{in}}} [\ell(f(\Phi(\boldsymbol{x})), y)]$.

3.3 Thought Experiment

Figure 2: Connections between ElRep and elastic net.

To demonstrate the intuition behind the benefit of ElRep, we present a simple statistical thought experiment. First, regularizing on the representation $\Phi(x)$ is a dual problem to regularizing the weight W (See Figure 2): Lasso or ElasticNet selects features by learning sparse model weight and thus zero-ing out the effect of the features corresponding to the zero weights. Meanwhile, nuclear norm or ElRep directly enforces learning low rank $\Phi(X)$ ((fewer number of features). We illustrate the benefit of Elastic Net first. Consider two features $\Phi(x)_1$ and $\Phi(x)_2$ with a strong spurious correlation ρ close to 1, but both features are equally important to predict y. If $\Phi(x)_1$ has a smaller magnitude, ℓ_1 regularization will assign its associated weight w_1 to 0, while elastic net ($\ell_1 + \ell_2$) tend to allocate non-zero elements in both w_1 and w_2 (since $\|[0.5, 0.5]\|_2 < \|[0, 1]\|_2$.) If for a target distribution the correlation between features changes, then ℓ_1 regularization fails to utilize the information from $\Phi(x)_1$ to predict y. We defer a more precise analysis to Section 5.2. Similarly, ElRep will also learn diverse features even if they might have some strong spurious correlation. Despite the connection between elastic net and ElRep, the latter is much better, since the success of elastic net depends on the quality of a pre-existing set of features to select from, and features learned through ERM may still have *non-linear spurious correlations* or lack diversity. ElRep addresses these issues by directly learning more robust features.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of ElRep on both synthetic and real data. For synthetic data, we design a setting where our method demonstrates advantages in terms of loss minimization and sparsity. For real data, we consider three popular benchmark datasets in the presence of spurious features: CelebA (Liu et al., 2015), Waterbirds (Sagawa et al., 2019), and CivilComments-WILDS (Koh et al., 2021). We present the worst-group accuracy, which assesses the minimum accuracy across all groups and is commonly used to evaluate the model's robustness against spurious correlations. Overall prediction accuracy is also reported to demonstrate minimum impacts of our approach on ID predictions.

4.1 Synthetic Data

Data generating process. We design $T = 3$ domains for training and one unseen domain for testing. We follow a similar data-generating procedure demonstrated in (Lu et al., 2021): we have a common label-related parameter C to generate invariant features for data in all four domains. For each domain, there is a domain-specific environment E_i, $i = 1, 2, 3, 4$. For each data point $\boldsymbol{x}$, we assume there are three non-trainable functions extracting three different types of features, respectively. The first type is invariant feature $\boldsymbol{z}_1(\boldsymbol{x}) \in \mathbb{R}^d$, which only depends on C. The second $\boldsymbol{z}_2(\boldsymbol{x}) \in \mathbb{R}^d$ is named nuanced features generated by both C and E_i so it has a weak correlation to the label. The third feature $\boldsymbol{z}_3(\boldsymbol{x}) \in \mathbb{R}^{k\times d}$ is spurious and generated by E_i only. Here, k is a hyper-parameter that controls the dimension of spurious feature and we choose $k = 3$ in the experiment. Consequently, the representation has dimension $(k + 2) \times d$.

Model and objectives. For simplicity, we set $W = [1, 1, \ldots, 1]$ that is not trainable and a linear representation Φ. Specifically, we define

$$\Phi(\boldsymbol{x}) = [\boldsymbol{a}_1^\top \odot \boldsymbol{z}_1(\boldsymbol{x})^\top, \boldsymbol{a}_2^\top \odot \boldsymbol{z}_2(\boldsymbol{x})^\top, \boldsymbol{a}_3^\top \odot \boldsymbol{z}_3(\boldsymbol{x})^\top],$$

where $\odot$ is the element-wise product. Denote $\boldsymbol{a} = [\boldsymbol{a}_1^\top, \boldsymbol{a}_2^\top, \boldsymbol{a}_3^\top]^\top$. The ground true label is generated by a representation $\Phi^*(\boldsymbol{x}) = \boldsymbol{a}^* \odot \boldsymbol{z}(\boldsymbol{x})$ plus a random noise, where $\boldsymbol{a}_3^* = 0$. We provide the data generating process

Table 1: The MSE (mean ± standard error) for different objectives on training data, ID test data, and OOD data. The best in OOD generalization is highlighted in bold. The results are averaged over 50 trials.

	Training	ID test	OOD
ERM	$0.0009_{\pm 0.0005}$	$29.30_{\pm 10.56}$	$63.90_{\pm 23.64}$
ℓ_1 regularization	$0.22_{\pm 0.03}$	$3.29_{\pm 0.69}$	$12.82_{\pm 4.60}$
ℓ_2 regularization	$0.10_{\pm 0.01}$	$3.59_{\pm 0.79}$	$13.62_{\pm 4.29}$
$\ell_1 + \ell_2$	$0.17_{\pm 0.02}$	$\mathbf{3.16_{\pm 0.67}}$	$\mathbf{11.77_{\pm 3.83}}$

in the appendix. The nuclear- and Frobenius-norms are reduced to ℓ_1- and ℓ_2-norms of $\boldsymbol{a}$, respectively. The objective function for training is

$$\min_{\boldsymbol{a}} \frac{1}{2nT} \sum_{t=1}^{T} \sum_{i=1}^{n} (y_{ti} - f(\boldsymbol{x}_{ti}))^2 + R(\boldsymbol{a}).$$

Our goal is small mean squared errors (MSEs) in the unseen domain. In the experiment, we consider three different forms of the regularizer $R(\boldsymbol{a})$: $\lambda\|\boldsymbol{a}\|_1$, $\lambda\|\boldsymbol{a}\|_2^2$, and $\lambda_1\|\boldsymbol{a}\|_1 + \lambda_2\|\boldsymbol{a}\|_2^2$. We expect that a Φ with more non-zero elements in the representation of invariant features and more zero elements for spurious features leads to a better performance on OOD predictions.

Results. We optimize the loss with the three different forms of $R(\boldsymbol{a})$ and without $R(\boldsymbol{a})$ (i.e., ERM), respectively. We run the simulation 50 times independently and compare the MSE of the training set, ID testing set, and OOD set. The result is shown in Table 1. Unsurprisingly, ERM has the lowest training MSE but the test error is significantly larger than using the regularized objectives for both ID and OOD tests. Notebaly, using both ℓ_1 and ℓ_2 penalties achieves the smallest ID and OOD test errors, and performance is consistent as reflected by the smallest standard errors.

We also examined $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ learned by different objectives. In particular, we compare the proportion of zero elements for each type of features between using ℓ_1 regularization alone and using the ℓ_1 and ℓ_2. The result is presented in Table 2. The average number of zero elements from ℓ_1 regularized loss is larger for all the three types of features. Using both ℓ_1 and ℓ_2 helps extract more information from invariant and nuanced features but more spurious features are also captured, implying a trade-off between preserving label-related features and eliminating environmental features. One can address this issue by mannually adjust λ's, and we will show their impacts, shortly.

4.2 Real Data

Datasets. (1) **CelebA** is comprised of 202,599 face images. We use it for hair-color classification with gender as the spurious feature. The smallest group is blond-hair men, which make up only 1% of total data, and over 93% of blond-hair examples are women. (2) **Waterbirds** is crafted by placing birds (Wah et al., 2011) on land or water backgrounds (Zhou et al., 2018). The goal is to classify birds as landbirds or waterbirds, and the spurious feature is the background. The smallest group is waterbirds on land. (3) **CivilComments-WILDS** is used to classify whether an online comment is toxic or not, and the label is spuriously correlated with mentions of eight demographic identities (DI), i.e. (male, female, White, Black, LGBTQ, Muslim, Christian, other religions). There are 16 group combinations, i.e., (DI, toxic) and (DI, non-toxic).

Table 2: The average proportion of zero elements for different types of features among 50 trials. The optimized features from ℓ_1 regularization is sparser than $\ell_1 + \ell_2$ for all kinds of features.

Feature	Invariant	Nuanced	Spurious
ℓ_1 regularization	$0.493_{\pm 0.044}$	$0.259_{\pm 0.044}$	$0.676_{\pm 0.023}$
$\ell_1 + \ell_2$	$0.348_{\pm 0.043}$	$0.168_{\pm 0.036}$	$0.560_{\pm 0.023}$

Baseline Models. Extant group robustness methods can be categorized into train-once and train-twice, as discussed in Section 2. The former often relies on the results from a single run. The latter, such as (Liu et al., 2021), requires running the training procedure in two stages to achieve ideal performance. In this paper, we compare the proposed ElRep against several state-of-the-art train-once methods, but ours is also readily combined with the train-twice approaches. Apart from standard ERM, we integrate the ElRep into several state-of-the-art methods, including Upweighting (UW) that inversely reweights group losses by group sizes, GroupDRO (Sagawa et al., 2019) that directly optimizes the worst group loss, the more recent PDE (Deng et al., 2023) that trains on a balanced subset of data then progressively expands the training set, and Subsample (Deng et al., 2023), a simplified version of PDE without the expansion stage. We compare the performance of these methods with and without ElRep.

Experiment Setup. We strictly follow the training and evaluation protocols used for the three datasets in previous studies (Piratla et al., 2022; Deng et al., 2023). The experiments are implemented based on the WILDS package (Koh et al., 2021) which uses pretrained ResNet-50 model (He et al., 2015) from Torchvision for CelebA and Waterbirds, and pretrained Bert model (Devlin et al., 2019) from HuggingFace for CivilComments-WILDS. All experiments were conducted on a single NVIDIA RTX 8000 GPU

Table 3: The worst-group and average accuracy (%) of ElRep compared with state-of-the-art methods. The best worst-group accuracy is highlighted in **bold**. The best average accuracy is also highlighted in bold if the worst-group accuracy is the same for multiple methods. Performance is evaluated on the test set with models early stopped at the highest worst-group accuracy on the validation set. N/A means no result is reported for UW on CivilComments, therefore we do not test our approach for this particular setting.

Method	Waterbirds		CelebA		CivilComments	
	Worst	Average	Worst	Average	Worst	Average
ERM	$70.0_{\pm 2.3}$	$97.1_{\pm 0.1}$	$45.0_{\pm 1.5}$	$94.8_{\pm 0.2}$	$58.2_{\pm 2.8}$	$92.2_{\pm 0.1}$
UW	$88.0_{\pm 1.3}$	$95.1_{\pm 0.3}$	$83.3_{\pm 2.8}$	$92.9_{\pm 0.2}$	N/A	N/A
Subsample	$86.9_{\pm 2.3}$	$89.2_{\pm 1.2}$	$86.1_{\pm 1.9}$	$91.3_{\pm 0.2}$	$64.7_{\pm 7.8}$	$83.7_{\pm 3.4}$
GroupDRO	$86.7_{\pm 0.6}$	$93.2_{\pm 0.5}$	$86.3_{\pm 1.1}$	$90.5_{\pm 0.3}$	$69.4_{\pm 0.9}$	$89.6_{\pm 0.5}$
PDE	$90.3_{\pm 0.3}$	$92.4_{\pm 0.8}$	$91.0_{\pm 0.4}$	$92.0_{\pm 0.5}$	$71.5_{\pm 0.5}$	$86.3_{\pm 1.7}$
ERM+ElRep	$79.8_{\pm 0.7}$	$89.5_{\pm 0.7}$	$52.6_{\pm 1.4}$	$95.5_{\pm 0.2}$	$60.5_{\pm 1.6}$	$91.5_{\pm 0.2}$
UW+ElRep	$89.1_{\pm 0.5}$	$92.5_{\pm 0.3}$	$90.2_{\pm 0.7}$	$92.4_{\pm 0.3}$	N/A	N/A
Subsample+ElRep	$88.7_{\pm 0.3}$	$90.8_{\pm 0.7}$	$89.6_{\pm 0.3}$	$91.1_{\pm 0.5}$	$70.8_{\pm 0.5}$	$82.1_{\pm 0.5}$
GroupDRO+ElRep	$88.8_{\pm 0.7}$	$92.9_{\pm 0.7}$	$\mathbf{91.4}_{\pm 1.0}$	$\mathbf{92.8}_{\pm 0.2}$	$70.5_{\pm 0.5}$	$79.0_{\pm 0.7}$
PDE+ElRep	$\mathbf{90.4}_{\pm 0.2}$	$91.6_{\pm 0.7}$	$\mathbf{91.4}_{\pm 0.5}$	$92.4_{\pm 0.3}$	$\mathbf{71.7}_{\pm 0.2}$	$80.7_{\pm 0.9}$

with 48GB memory. Our code is available at `https://github.com/TaoWen0309/ElRep`.

We follow previous work and run all experiments with three different random seeds and report the mean and standard deviation of worst-group and average accuracy. For a fair comparison, the baseline performance is the original results from recent studies (Wu et al., 2023; Deng et al., 2023; Phan et al., 2024). We have not modified their published code or hyper-parameters except for adding the regularization. Also, we do not report the performance of UW on the CivilComments dataset since it has not been benchmarked by extant work. We select the hyper-parameters for the nuclear and Frobenius norms by cross-validation with candidate λ_1 between 10^{-4} and 10^{-3} and candidate λ_2 between 10^{-5} and 10^{-4}.

4.2.1 Average and Worst-Group Accuracy

We compare the performance of ERM, UW, Subsample, GroupDRO, and PDE with and without ElRep and report in Table 3 their average and worst-group prediction accuracy. As a result, the proposed ElRep improves all the state-of-the-art methods compared in worst-group accuracy (the top half versus the bottom half of the table), demonstrating its effectiveness in group robustness. The best worst-group accuracy is achieved by GroupDRO or PDE together with ElRep. The improvement is more pronounced if ElRep is combined with a more naive model. For example, ERM has been improved by 6.6 percentage on average. We show how much these extant methods are improved by ElRep in the left panel of Figure 3.

Furthermore, ElRep helps reduce performance fluctuation. Specifically, the standard deviation of the worst-group accuracy is typically smaller when a method is combined with ElRep, suggesting its consistently effective learning of invariant features, which may be indispensable for domain generalization. Although enhanced group robustness is often achieved at the cost of reduced overall accuracy, we observe that ElRep simultaneously improves both average and worst-group accuracy for several baselines on the waterbirds and CelebA datasets, which is shown in the right panel of Figure 3. This is attributed to the theoretical underpinning that ElRep does not undermine ID prediction, as shown in Section 5, shortly.

4.2.2 Ablation Study of the Regularization

Regularization by either nuclear- or Frobenius-norm. The advantage of ElRep comes from the combination of a nuclear norm and a Frobenius norm. We consider only using either of them and compare the performance. As reported in Table 4, in most cases, our approach is the best. Removing either norm would lead to a degradation of worst-group accuracy, and sometimes, it even underperforms the method without regularization, like ERM on CelebA. In addition, our results show that using one norm does not consistently outperform using the other.

Regularization via Weight Decay. Though intuitively similar to the elastic net, we regularize the representation rather than the weights. We compare the proposed method with weight decay (WD), which imposes an ℓ_2 penalty on the weights of the linear classification layer of a neural network. We leave CivilComments out for a fair comparison because the Bert model

Table 4: The worst-group and average accuracy (%) of our approach compared with nuclear Norm (NN) or Frobenius Norm (FN) alone. The experiment settings are the same as in Table 3. ElRep achieves the best worst-group performance in almost all settings.

Method	Waterbirds		CelebA		CivilComments	
	Worst	Average	Worst	Average	Worst	Average
ERM (FN)	$78.0_{\pm 0.3}$	$89.0_{\pm 0.2}$	$43.9_{\pm 4.0}$	$95.5_{\pm 0.1}$	$58.9_{\pm 1.0}$	$91.6_{\pm 0.1}$
ERM (NN)	$78.8_{\pm 0.3}$	$89.6_{\pm 0.5}$	$44.1_{\pm 4.7}$	$95.5_{\pm 0.1}$	$59.3_{\pm 0.2}$	$91.9_{\pm 0.2}$
ERM (Ours)	$\mathbf{79.8}_{\pm 0.4}$	$89.5_{\pm 0.4}$	$\mathbf{52.6}_{\pm 0.8}$	$95.5_{\pm 0.1}$	$\mathbf{60.5}_{\pm 0.9}$	$91.5_{\pm 0.1}$
UW (FN)	$88.2_{\pm 0.6}$	$92.1_{\pm 0.1}$	$89.4_{\pm 0.5}$	$92.5_{\pm 0.2}$	N/A	
UW (NN)	$88.4_{\pm 0.6}$	$92.0_{\pm 0.1}$	$89.7_{\pm 0.3}$	$92.2_{\pm 0.3}$		
UW (Ours)	$\mathbf{89.1}_{\pm 0.3}$	$92.5_{\pm 0.2}$	$\mathbf{90.2}_{\pm 0.4}$	$92.4_{\pm 0.2}$		
Subsample (FN)	$\mathbf{89.1}_{\pm 0.3}$	$90.9_{\pm 0.4}$	$87.8_{\pm 0.5}$	$91.9_{\pm 0.2}$	$70.3_{\pm 0.4}$	$81.2_{\pm 0.4}$
Subsample (NN)	$88.7_{\pm 0.1}$	$91.0_{\pm 0.3}$	$88.9_{\pm 0.5}$	$91.3_{\pm 0.1}$	$70.5_{\pm 0.3}$	$80.5_{\pm 0.3}$
Subsample (Ours)	$88.7_{\pm 0.2}$	$90.8_{\pm 0.4}$	$\mathbf{89.6}_{\pm 0.2}$	$91.1_{\pm 0.3}$	$\mathbf{70.8}_{\pm 0.3}$	$82.1_{\pm 0.3}$
GroupDRO (FN)	$88.7_{\pm 0.5}$	$92.5_{\pm 0.3}$	$90.8_{\pm 0.2}$	$93.1_{\pm 0.1}$	$69.9_{\pm 0.5}$	$78.2_{\pm 0.5}$
GroupDRO (NN)	$86.8_{\pm 0.9}$	$92.4_{\pm 0.4}$	$90.8_{\pm 1.0}$	$92.8_{\pm 0.3}$	$\mathbf{70.5}_{\pm 0.5}$	$\mathbf{79.2}_{\pm 0.7}$
GroupDRO (Ours)	$\mathbf{88.8}_{\pm 0.4}$	$92.9_{\pm 0.4}$	$\mathbf{91.4}_{\pm 0.6}$	$92.8_{\pm 0.1}$	$\mathbf{70.5}_{\pm 0.3}$	$79.0_{\pm 0.4}$
PDE (FN)	$89.8_{\pm 0.1}$	$91.4_{\pm 0.1}$	$90.2_{\pm 0.4}$	$91.7_{\pm 0.2}$	$70.2_{\pm 0.1}$	$80.8_{\pm 0.7}$
PDE (NN)	$89.8_{\pm 0.2}$	$91.2_{\pm 0.3}$	$\mathbf{91.4}_{\pm 0.3}$	$91.9_{\pm 0.3}$	$71.0_{\pm 0.3}$	$82.2_{\pm 0.5}$
PDE (Ours)	$\mathbf{90.4}_{\pm 0.1}$	$91.6_{\pm 0.4}$	$\mathbf{91.4}_{\pm 0.3}$	$\mathbf{92.4}_{\pm 0.2}$	$\mathbf{71.7}_{\pm 0.1}$	$80.7_{\pm 0.5}$

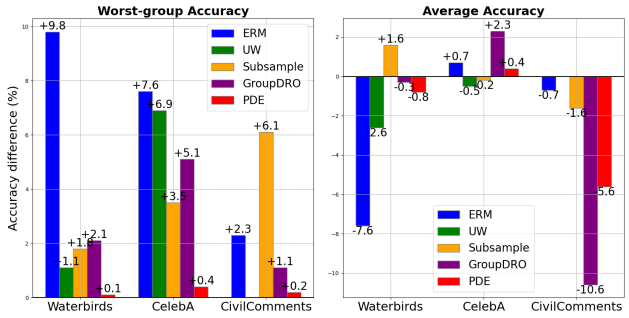

Figure 3: Left: The difference in the worst-group accuracy between the baseline methods with and without ElRep. The improvement is ubiquitous among all the methods compared on all the three datasets. Right: The difference in the average accuracy between the baseline methods with and without ElRep. Usually, an increase in worst-group accuracy comes with a decrease in average accuracy. Our approach can also improve the average accuracy for some baselines on the image datasets.

Table 5: The accuracy (%) of ERM with weight decay (WD) and ElRep. ElRep significantly outperforms WD in worst-group performance with minimal sacrifice of average accuracy.

Method	Waterbirds		CelebA	
	Worst	Average	Worst	Average
ERM+WD	$78.9_{\pm 0.6}$	$89.7_{\pm 0.6}$	$44.8_{\pm 3.4}$	$95.8_{\pm 0.1}$
ERM+ElRep	$79.8_{\pm 0.4}$	$89.5_{\pm 0.4}$	$52.6_{\pm 0.8}$	$95.5_{\pm 0.1}$

uses its own learning schedule, and magnified weight decay can undermine its performance. The results in Table 5 indicate that ours is better than regularization on the weights in group robustness at a minimum cost of average accuracy.

4.2.3 Regularization Intensity

We study the influence of the regularization intensities. Specifically, λ_1 and λ_2 control the nuclear-norm and Frobenius-norm penalties, respectively, and their values affect the model performance. Too small values cannot effectively regularize spurious correlations, while too large values make the penalties overwhelm the classification loss. In Figure 4, we try various values of λ within a reasonable range on CelebA, and show the accuracy on each group and the average accuracy. An obvious trend can be observed that the minority-group (blonde hair) accuracy gradually increases with the value of λ_1 or λ_2. If λ is sufficiently large the minority group accuracy would eventually surpass the average accuracy. The opposite trend can be observed for the majority groups (non-blonde females and males).

To further validate this observation, we randomly downsample the original majority groups, i.e. non-blonde-hair females and males to approximately 1%. By Figure 5, we can observe the same trend although the roles of majority and minority groups are now switched compared to Figure 4. This observation is

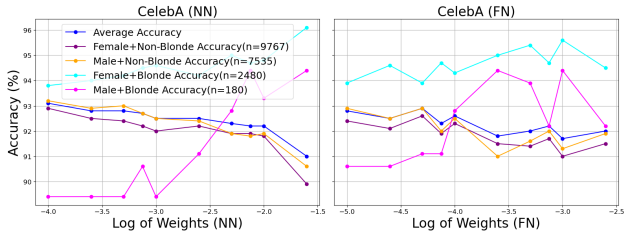

Figure 4: Accuracy per group and average accuracy against the log of λ_1 (left) and λ_2 (right). As their value increases, the accuracy of the two minority groups will gradually increase and eventually surpass the average accuracy. The trend is reversed for the two majority groups.

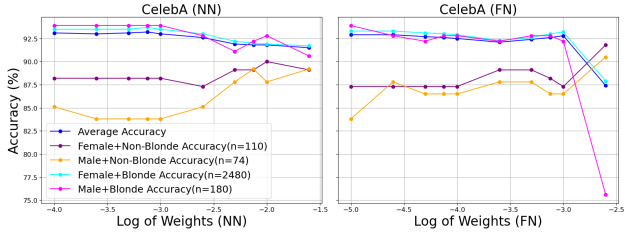

Figure 5: The two majority groups downsampled to about 1%. Reversed trends are observed.

useful in cases where we only care about small group accuracy since we can set arbitrarily large values for λ_1 and λ_2, as long as the regularization term does not overwhelm the classification loss.

5 THEORETICAL ANALYSIS

In this section, we provide some theoretical analysis to ElRep, showing that 1) the regularization term will not hurt ID prediction and 2) adding a Frobenius norm term towards nuclear norm penalty can effectively capture more invariant features.

5.1 ID Prediction

When training deep learning models, regularization is used to prevent overfits. Previous sections illustrated that ElRep makes OOD prediction more accurate by introducing nuclear- and Frobenius-norm penalties, mitigating an over-regularization of invariant features. However, regularization may hurt ID performance. In this section, we show that the regularization of ElRep does not hurt ID prediction.

Settings. In our analysis, we consider a regression problem on space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. We set the model be a linear regression problem $f(\boldsymbol{x}) = \theta^\top \boldsymbol{x}$ for simplicity. In multi domain learning, there are T different training domains. For each domain, every sample in $X_t \in \mathbb{R}^{n\times d}$ is generated from a distribution p_t supported on $\mathcal{X}$. We assume that $\mathbb{E}_{\boldsymbol{x}\sim p_t}\boldsymbol{x} = 0$ and $\mathbb{E}_{\boldsymbol{x}\sim p_t}\boldsymbol{x}\boldsymbol{x}^\top = \Sigma_t$. Then for $\bar{\boldsymbol{x}} = \Sigma^{-1/2}\boldsymbol{x}$ generated from $\bar{p}_t$, $\mathbb{E}_{\bar{\boldsymbol{x}}\sim\bar{p}_t}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top = I$. The labels $Y_t \in \mathbb{R}^n$ is generated by $Y_t = X_t\theta^* + \epsilon$, where Θ^* is the ground truth parameter and $\epsilon \sim \mathcal{N}(0, \sigma I_n)$.

Assumption 5.1. *There exists a positive semi-definite matrix Σ such that $\Sigma_t \preceq \Sigma$ for any t.*

Assumption 5.2. *There exists $\rho > 0$ such that the random vector $\bar{\boldsymbol{x}} \sim \bar{p}_t$ is ρ^2-subgaussian for any t.*

Objective. In the multi-task regression with ElRep, we minimize the following objective

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{2nT}\|\mathcal{X}(\theta) - Y\|_F^2 + \lambda_1\|\theta\|_1 + \lambda_2\|\theta\|_2, \qquad (2)$$

where $\mathcal{X}(\theta) = [X_1\theta, \cdots, X_T\theta] \in \mathbb{R}^{n\times T}$. Note that we penalize both l_1 and l_2 norm of the regression weight θ, which has a similar effect of penalizing the representation in representation learning setting.

Theoretical results. If we denote the solution of (2) by $\hat{\theta}$, we are interested in the population excess risk, i.e. $\frac{1}{2T}\sum_{t=1}^T \mathbb{E}_{p_t}\|X\Delta\|_{\mathrm{F}}^2$, where $\Delta = \hat{\theta} - \theta^*$. The following theorem gives an upper bound.

Theorem 5.1. *Under Assumption 5.1 and 5.2, we fix a failure probability δ and choose proper $\lambda_1, \lambda_2, \lambda_3$. Then with probability at least $1-\delta$ over training samples, the prediction difference between our approach and the ground truth satisfies:*

$$\frac{1}{2T}\sum_{t=1}^T \mathbb{E}_{p_t}\|X\Delta\|_{\mathrm{F}}^2 \le \tilde{O}\left(\frac{\sigma R\sqrt{\mathrm{Tr}(\Sigma)}}{\sqrt{nT}}\right) + \tilde{O}\left(\frac{\rho^4 R^2 \mathrm{Tr}(\Sigma)}{nT}\right), \qquad (3)$$

where $R = \|\theta^*\|_1$ and we omit logarithmic factors.

The proof of Theorem 5.1 is deferred to the appendix. This upper bound shows that prediction using ElRep is close to the ground truth if the number of samples n is large, implying ElRep does not hurt ID performance. Note that for nuclear norm regularization, the bound only differs in constant coefficients according to Du et al. (2021). The analysis of OOD performance is not included because more assumptions of the testing domain are needed, and we defer it to future work.

5.2 Feature Selection

Nuclear norm regularization improves the OOD prediction by eliminating spurious features. However, the experiments in Section 4 show that ElRep performs

better than the nuclear norm penalty in worst group prediction. One reason is that nuclear norm regularization rules out some invariant features highly correlated with others. In OOD tasks, the correlation may be changed and the eliminated features can be vital in prediction. In this section, we show that ElRep is more likely to keep correlated features than the nuclear norm penalty.

Settings. For simplicity, we consider a linear regression problem $f(\boldsymbol{x}) = \theta^\top \boldsymbol{x}$. The training sample $X \in \mathbb{R}^{n\times d}$ has zero mean and satisfies that empirical variance $\frac{1}{n}X^\top X = I_d + \rho(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top)$, where $\boldsymbol{e}_i$ is the i-th standard basis vector and $0 < \rho < 1$. Note that there is a positive correlation ρ between the i-th and the j-th entry of the data, which is a simplified setting of correlated features. With the ground truth parameter θ^* and noise $\epsilon \sim \mathcal{N}(0, \sigma I_n)$, the label is generated by $Y = X^\top\theta^* + \epsilon$. We introduce the unregularized least square solution $\hat{\theta} := \arg\min \|X\theta - Y\|^2$ satisfying $X^\top X\hat{\theta} = X^\top Y$ for the ease of presentation and assume $0 < \hat{\theta}_i < \hat{\theta}_j$ without loss of generality.

Theoretical results. If we denote the least square solution with ℓ_1 norm regularization by

$$\theta^1 = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|X\theta - Y\|_2^2 + \lambda_1\|\theta\|_1$$

and the least square solution with $\ell_1 + \ell_2$ regularizers by

$$\theta^{\mathrm{El}} = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|X\theta - Y\|_2^2 + \lambda_1\|\theta\|_1 + \frac{\lambda_2}{2}\|\theta\|_2^2,$$

we have the following result.

Proposition 5.2. *Under the following conditions on regularizers λ_1, λ_2 and unregularized least square solution $\hat{\theta}$, the regularized least square solutions θ^1 and θ^{El} satisfy:*

θ stands for:	ℓ_1 regularization (θ^1)	ElRep (θ^{El})
$\theta_i, \theta_j > 0$	$\lambda_1 < (1+\rho)\hat{\theta}_i$	$\lambda_1 < c$
$\theta_i = 0, \theta_j > 0$	$(1+\rho)\hat{\theta}_i \le \lambda_1 < \hat{\theta}_j + \rho\hat{\theta}_i$	$c \le \lambda_1 < \hat{\theta}_j + \rho\hat{\theta}_i$
$\theta_i = \theta_j = 0$	$\lambda_1 \ge \hat{\theta}_j + \rho\hat{\theta}_i$	$\lambda_1 \ge \hat{\theta}_j + \rho\hat{\theta}_i$

where $c = \frac{(1+\lambda_2-\rho^2)\hat{\theta}_i + \lambda_2\rho\hat{\theta}_j}{1+\lambda_2-\rho}$.

See the appendix for the proof of Proposition 5.2. We note that $c > (1+\rho)\hat{\theta}_i$ always holds as we assumed $\hat{\theta}_i < \hat{\theta}_j$ WLOG. Therefore the proposition indicates that ElRep always keeps the features when they are both selected by Lasso: as long as $\theta_i^1, \theta_j^1 > 0$, we always have $\theta_i^{\mathrm{El}}, \theta_j^{\mathrm{El}} > 0$. In contrast, there exists cases when $\theta_i^{\mathrm{El}}, \theta_j^{\mathrm{El}} > 0$ while $\theta_i^1 = 0$. This result indicates that ElRep is more likely to capture correlated features simultaneously. Moreover, the gap between the left threshold c and $(1+\rho)\hat{\theta}_i$ is larger when ρ and λ_2 (for λ_1 are larger; this means the distinction in feature selection between Lasso and ElRep is more significant with higher correlated features and more intense Frobenius norm regularization.

6 CONCLUSION

In conclusion, we propose to address spurious correlations by Elastic Representation. It enables neural networks to learn more invariant features by imposing the nuclear norm and Frobenius norm of the feature representations and can be readily integrated into a wide range of extant approaches. Theoretically, we show that adding the regularization will not hurt the in-distribution performance. Empirically, extensive experiments validate the proposed method.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science Energy Earthshot Initiative as part of the project "Learning reduced models under extreme data conditions for design and rapid decision-making in complex systems" under Award #DE-SC0024721.

References

Chen, Y., Huang, W., Zhou, K., Bian, Y., Han, B., and Cheng, J. (2023). Understanding and improving feature learning for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36.

Deng, Y., Yang, Y., Mirzasoleiman, B., and Gu, Q. (2023). Robust learning with progressive data expansion against spurious correlation. *Advances in neural information processing systems*, 36.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. (2021). Few-shot learning via learning the representation, provably. In *9th International Conference on Learning Representations, ICLR 2021*.

Du, Y., Yan, J., Chen, Y., Liu, J., Zhao, S., She, Q., Wu, H., Wang, H., and Qin, B. (2023). Less learn shortcut: Analyzing and mitigating learning of spurious feature-label correlation. *IJCAI*.

Eastwood, C., Singh, S., Nicolicioiu, A. L., Vlastelica Pogančić, M., von Kügelgen, J., and Schölkopf,

B. (2023). Spuriosity didn't kill the classifier: Using invariant predictions to harness spurious features. *Advances in Neural Information Processing Systems*, 36.

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.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*.

Goel, K., Gu, A., Li, Y., and Ré, C. (2020). Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*.

Haghtalab, N., Jordan, M., and Zhao, E. (2022). On-demand sampling: Learning optimally from multiple distributions. *Advances in Neural Information Processing Systems*, 35:406–419.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Huang, S.-C., Chaudhari, A. S., Langlotz, C. P., Shah, N., Yeung, S., and Lungren, M. P. (2022). Developing medical imaging ai for emerging infectious diseases. *nature communications*, 13(1):7060.

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.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. (2022). On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532.

Kirichenko, P., Izmailov, P., and Wilson, A. G. (2023). Last layer re-training is sufficient for robustness to spurious correlations. *International Conference on Learning Representations*.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. (2021). Wilds: A benchmark of in-the-wild distribution shifts.

LaBonte, T., Muthukumar, V., and Kumar, A. (2023). Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, 36.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. (2020). Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860.

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. (2015). Deep learning face attributes in the wild.

Lu, C., Wu, Y., Hernández-Lobato, J. M., and Schölkopf, B. (2021). Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*.

Moayeri, M., Wang, W., Singla, S., and Feizi, S. (2023). Spuriosity rankings: sorting data to measure and mitigate biases. *Advances in Neural Information Processing Systems*, 36:41572–41600.

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.

Phan, H., Wilson, A. G., and Lei, Q. (2024). Controllable prompt tuning for balancing group distributional robustness.

Piratla, V., Netrapalli, P., and Sarawagi, S. (2022). Focus on the common good: Group distributional robustness follows.

Rosenfeld, E., Ravikumar, P., and Risteski, A. (2022). Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*.

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. *International Conference on Learning Representations*.

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.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

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, Z., Ming, Y., Fan, Y., Sala, F., and Liang, Y. (2024). Domain generalization via nuclear norm regularization. In *Conference on Parsimony and Learning*, pages 179–201. PMLR.

Sun, X., Wu, B., Zheng, X., Liu, C., Chen, W., Qin, T., and Liu, T.-Y. (2021). Recovering latent causal factor for generalization to distributional shifts. *Advances in Neural Information Processing Systems*, 34:16846–16859.

Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. (2021). Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 Dataset*.

Wu, S., Yuksekgonul, M., Zhang, L., and Zou, J. (2023). Discover and cure: concept-aware mitigation of spurious correlation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Xiao, K., Engstrom, L., Ilyas, A., and Madry, A. (2021). Noise or signal: The role of image backgrounds in object recognition. *International Conference on Learning Representations*.

Yang, Y., Gan, E., Dziugaite, G. K., and Mirzasoleiman, B. (2024). Identifying spurious biases early in training through the lens of simplicity bias. In *International Conference on Artificial Intelligence and Statistics*, pages 2953–2961. PMLR.

Ye, W., Zheng, G., Cao, X., Ma, Y., Hu, X., and Zhang, A. (2024). Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

Yenamandra, S., Ramesh, P., Prabhu, V., and Hoffman, J. (2023). Facts: First amplify correlations and then slice to discover bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4794–4804.

Zhang, J., Menon, A., Veit, A., Bhojanapalli, S., Kumar, S., and Sra, S. (2021). Coping with label shift via distributionally robust optimisation. *International Conference on Learning Representations*.

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. In *International Conference on Machine Learning*. PMLR.

Zhong, Z. S., Pan, X., and Lei, Q. (2024). Bridging domains with approximately shared features. *arXiv preprint arXiv:2403.06424*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

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/No/Not Applicable] Yes. Our proposed approach is straightforward and discussed in detail throughout the paper.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Not Applicable. Our proposed regularization does not incur extra computational costs and its complexity depends on the baselines.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] Yes. Our code is included.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Yes. We state all assumptions on data and model in Section 5.
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Yes. The full proofs are contained in Appendix.
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Yes. We explain all the assumptionss in Section 5.

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/No/Not Applicable] Yes. Our code is included.
   (b) All the training details (e.g., data splits, hyper-parameters, how they were chosen). [Yes/No/Not Applicable] Yes. We introduced our experiment setup details in Section 4.2.
   (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/No/Not Applicable] Yes. We introduced the accuracy measure in Section 4.2.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] Yes. It is also given in Section 4.2.

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/No/Not Applicable] Yes. We cited them at the beginning of Section 4 and in 4.2.
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] Not Applicable.
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Not Applicable. All data is public and free for research purposes.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] 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. [Yes/No/Not Applicable] Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] Not Applicable.

A Details of synthetic data experiment

In the synthetic data experiment, we generated 3 training domains and 1 testing domain. In the data generating process, we consider a label-related parameter $C \in \mathbb{R}^d$, and for each domain, there is an environmental parameter $E_i \in \mathbb{R}^d$. The features $\boldsymbol{z}$ are generated from those parameters. Specifically, the invariant feature $\boldsymbol{z}_1 = c_1 C + \epsilon_1 \in \mathbb{R}^d$. The nuanced feature $\boldsymbol{z}_2 = c_2(\rho C + \sqrt{1-\rho^2} E_i) + \epsilon_2 \in \mathbb{R}^d$, where ρ is a hyperparameter controlling the ratio of two types of parameters. The spurious feature $\boldsymbol{z}_3 = E\boldsymbol{c}_3 + \epsilon_3$. Here $c_1, c_2 \in \mathbb{R}$, $\boldsymbol{c}_3 \in \mathbb{R}^{1\times k}$ are random coefficients and $\epsilon_1, \epsilon_2 \in \mathbb{R}^d$, $\epsilon_3 \in \mathbb{R}^{d\times k}$ are random noise. As we mentioned in Section 4.1, we choose the dimension of spurious feature $k = 3$. Moreover, we set $\rho = 0.5$, $d = 100$ and $n = 120$.

B Proof of Theorem 5.1

In order to prove the in-distribution generalization result in Theorem 5.1, we first give some lemmas showing the bound for training error $\mathcal{X}(\Delta)$, where $\mathcal{X}(\theta) = [X_1\theta, \cdots, X_T\theta] \in \mathbb{R}^{n\times T}$. We denote the total noise by $Z := \mathcal{X}(\theta^*) - Y$, where each column $Z_t \sim \mathcal{N}(0, \sigma I_n)$.

Lemma B.1. *If Assumption 5.1 holds, then with probability at least* $1-\delta$

$$\frac{1}{nT}\left\|\mathcal{X}^*(Z)\right\|_2 \le \tilde{O}\left(\frac{\sigma\sqrt{\mathrm{Tr}(\Sigma)}}{\sqrt{nT}}\right),$$

where $\mathcal{X}^*(Z) = \sum_{t=1}^T X_t^\top Z_t$ *and the log terms are omitted.*

Proof. Let

$$A = \frac{1}{\sqrt{n}}\mathcal{X}^*(Z) = \frac{1}{\sqrt{n}}\sum_{t=1}^T X_T^\top Z_t =: \sum_{t=1}^T S_t.$$

Then we have

$$\begin{aligned}\mathbb{E}\left[AA^\top\right] &= \mathbb{E}_X\left[\sum_{t=1}^T \frac{1}{n} X_t^\top \mathbb{E}\left[Z_t Z_t^\top\right] X_t\right] \\ &= \sigma^2 \sum_{t=1}^T \Sigma_t\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[A^\top A\right] &= \frac{1}{n}\sum_{t=1}^T \mathbb{E}_Z\left[Z_t^\top \mathbb{E}_X\left[X_t X_t^\top\right] Z_t\right] \\ &= \sigma^2 \sum_{t=1}^T \mathrm{Tr}(\Sigma_t).\end{aligned}$$

Then

$$\nu(A) := \max\left\{\mathbb{E}\left[AA^\top\right], \mathbb{E}\left[A^\top A\right]\right\} = \sigma^2 \sum_{t=1}^T \mathrm{Tr}(\Sigma_t).$$

Let $V(A) := \mathrm{diag}\left(\mathbb{E}\left[AA^\top\right], \mathbb{E}\left[A^\top A\right]\right)$. Then

$$V(A) = \sigma^2 \mathrm{diag}\left(\sum_{t=1}^T \Sigma_t, \sum_{t=1}^T \mathrm{Tr}(\Sigma_t)\right)$$

and we define $d(A) := \mathrm{Tr}(V(A))/\|V(A)\|_2 = 2$. Besides, by Hanson-Wright inequality, we have

$$\|S_t\|_2^2 \le \sigma^2 \mathrm{Tr}(\Sigma_t) + \sigma^2 \|\Sigma_t\| \log\frac{2}{\delta} + \sigma^2 \|\Sigma_t\|_{\mathrm{F}} \sqrt{\log\frac{2}{\delta}}$$

with probability $1-\delta/2$. Since $\|\Sigma_t\|_{\mathrm{F}} \le \mathrm{Tr}(\Sigma_t)$ and $\Sigma_t \preceq \Sigma$, we have

$$\|S_t\|_2 \le \sigma\sqrt{\left(1+\sqrt{\log\frac{2}{\delta}}\right)\mathrm{Tr}(\Sigma) + \|\Sigma\|\log\frac{2}{\delta}} =: L.$$

Then by Theorem 7.3.1 in (Tropp et al., 2015), with probability $1-\delta/2$,

$$\begin{aligned}\|A\|_2 &\lesssim \sigma\sqrt{\log\frac{2}{\delta}\nu(A)\log(d(A))} + \log\frac{2}{\delta}\sigma L\log(d(A)) \\ &\lesssim \sigma\left(\log\frac{2}{\delta}\right)^{3/2}\sqrt{T\mathrm{Tr}(\Sigma)}.\end{aligned}$$

Thus, with probability at list $1-\delta$,

$$\frac{1}{nT}\|\mathcal{X}^*(Z)\|_2 \le \tilde{O}\left(\frac{\sigma\sqrt{\mathrm{Tr}(\Sigma)}}{\sqrt{nT}}\right).$$

□

Lemma B.2. *If Assumption 5.1 holds and choose proper λ_1, λ_2 and λ_3, then with probability at least $1-\delta$,*

$$\frac{1}{2nT}\|\mathcal{X}(\Delta)\|_{\mathrm{F}}^2 \le \tilde{O}\left(\frac{\sigma R\sqrt{\mathrm{Tr}(\Sigma)}}{nT}\right)$$

and the optimal solution $\hat{\theta}$ satisfies

$$\|\hat{\theta}\|_2 \lesssim R,$$

where $R = \|\Theta^*\|_1$ and the log terms are omitted.

Proof. By the definition of $\hat{\theta}$, we have the following inequality:

$$\frac{1}{2nT}\|\mathcal{X}(\Delta)+Z\|_{\mathrm{F}}^2 + \lambda_1\|\hat{\theta}\|_1 + \lambda_2\|\hat{\theta}\|_2 \le \frac{1}{2nT}\|Z\|_{\mathrm{F}}^2 + \lambda_1\|\theta^*\|_1 + \lambda_2\|\theta^*\|_2.$$

Then

$$\frac{1}{2nT}\|\mathcal{X}(\Delta)\|_{\mathrm{F}}^2 + \frac{1}{nT}\langle\mathcal{X}(\Delta), Z\rangle + R(\hat{\theta}) \le R(\theta^*),$$

where $R(\theta) = \lambda_1\|\theta\|_1 + \lambda_2\|\theta\|_2$. By reordering the inequality

$$\begin{aligned}\frac{1}{2nT}\|\mathcal{X}(\Delta)\|_{\mathrm{F}}^2 &\le -\frac{1}{nT}\langle\mathcal{X}(\Delta), Z\rangle + R(\theta^*) - R(\hat{\theta}) \\ &\le \frac{1}{nT}\left(\|\hat{\theta}\|_2 + \|\theta^*\|_2\right)\|\mathcal{X}^*(Z)\|_2 + R(\theta^*) - R(\hat{\theta}).\end{aligned}$$

If we choose $\lambda_1 = \frac{\|\mathcal{X}^*(Z)\|_2}{nT}$ and $\lambda_2 = \frac{2\|\mathcal{X}^*(Z)\|_2}{nT}$, then

$$\frac{1}{2nT}\|\mathcal{X}(\Delta)\|_{\mathrm{F}}^2 + \lambda_1\|\hat{\theta}\|_1 + \frac{\lambda_2}{2}\|\hat{\theta}\|_2 \le \frac{1}{nT}\|\theta^*\|_2\|\mathcal{X}^*(Z)\|_2 + R(\theta^*).$$

The right hand side

$$\begin{aligned}\frac{1}{nT}\|\theta^*\|_2\|\mathcal{X}^*(Z)\|_2 + R(\theta^*) &= \frac{1}{nT}\left(\|\theta^*\|_2 + \|\theta^*\|_1 + 2\|\theta^*\|_2\right)\|\mathcal{X}^*(Z)\|_2 \\ &= \frac{1}{nT}\left(4\|\theta^*\|_1\right)\|\mathcal{X}^*(Z)\|_2 \\ &\le \tilde{O}\left(\frac{\sigma R\sqrt{\mathrm{Tr}(\Sigma)}}{\sqrt{nT}}\right),\end{aligned} \tag{4}$$

where the last equation applies Lemma B.1. Therefore

$$\frac{1}{2nT}\|\mathcal{X}(\Delta)\|_{\mathrm{F}}^2 \leq \tilde{O}\left(\frac{\sigma R\sqrt{\mathrm{Tr}(\Sigma)}}{\sqrt{nT}}\right),$$

and

$$\frac{\|\mathcal{X}^*(Z)\|_2}{nT}\|\hat{\theta}\|_2 \leq \frac{4R}{nT}\|\mathcal{X}^*(Z)\|_2,$$

implying $\|\hat{\theta}\|_2 \lesssim R$. □

With the result of above, we can now proof Theorem 5.1.

Proof of Theorem 5.1. By Lemma C.10 in (Du et al., 2021), if Assumption 5.2 holds,

$$\left\|\Sigma_t^{1/2}\Delta\right\|_2 \leq \frac{1}{\sqrt{n}}\|X_t\Delta\|_2 + \frac{C\rho}{\sqrt{n}}\left(\sqrt{\mathrm{Tr}(\Sigma_t)} + \sqrt{\log\frac{2}{\delta}\|\Sigma_t\|}\right)\|\Delta\|_2.$$

Then

$$\begin{aligned}
\mathbb{E}_{p_t}\|X\Delta\|_2^2 = \left\|\Sigma_t^{1/2}\Delta\right\|_2^2 &\lesssim \frac{1}{n}\|X_t\Delta\|_2^2 + \frac{C\rho^3}{n}\left(\mathrm{Tr}(\Sigma_t) + \log\frac{2}{\delta}\|\Sigma_t\|\right)\|\Delta\|_2^2 \\
&\lesssim \frac{1}{n}\|X_t\Delta\|_2^2 + \frac{C\rho^4\log\frac{2}{\delta}\mathrm{Tr}(\Sigma_t)}{n}\left(\|\hat{\theta}\|_2^2 + \|\theta^*\|_2^2\right) \\
&\leq \frac{1}{n}\|X_t\Delta\|_2^2 + \frac{C\rho^4\log\frac{2}{\delta}\mathrm{Tr}(\Sigma_t)}{n}R^2,
\end{aligned}$$

where the last inequality is from the second part of Lemma B.2. We sum the above inequality up for all $t = 1, \ldots, T$,

$$\begin{aligned}
\frac{1}{2T}\sum_{t=1}^{T}\mathbb{E}_{p_t}\|X\Delta\|_2^2 &\lesssim \frac{1}{2T}\sum_{t=1}^{T}\left(\frac{1}{n}\|X_t\Delta\|_2^2 + \frac{C\rho^4\log\frac{2}{\delta}\mathrm{Tr}(\Sigma_t)}{n}R^2\right) \\
&= \frac{1}{2nT}\|\mathcal{X}\Delta\|_{\mathrm{F}}^2 + \frac{1}{2T}\sum_{t=1}^{T}\frac{C\rho^4\log\frac{2}{\delta}\mathrm{Tr}(\Sigma_t)}{n}R^2 \\
&\leq \tilde{O}\left(\frac{\sigma R\sqrt{\mathrm{Tr}(\Sigma)}}{\sqrt{nT}}\right) + \tilde{O}\left(\frac{\rho^4 R^2\mathrm{Tr}(\Sigma)}{nT}\right),
\end{aligned}$$

where the last inequality is given by the first part of Lemma B.2. □