Domain Generalization with Nuclear Norm Regularization

Zhenmei Shi∗**, Yifei Ming**∗**, Ying Fan**∗**, Frederic Sala, Yingyu Liang**
University of Wisconsin-Madison
zhmeishi, alvinming, yingfan, fredsala, yliang@cs.wisc.edu

Abstract

The ability to generalize to unseen domains is crucial for machine learning systems, especially when we only have data from limited training domains and must deploy the resulting models in the real world. In this paper, we study domain generalization via the classic empirical risk minimization (ERM) approach with a simple regularizer based on the nuclear norm of the learned features from the training set. Theoretically, we provide intuitions on why nuclear norm regularization works better than ERM and ERM with L2 weight decay in linear settings. Empirically, we show that nuclear norm regularization achieves state-of-the-art average accuracy compared to existing methods in a wide range of domain generalization tasks (e.g. 1.7% test accuracy improvements over the second-best baseline on DomainNet).

1 Introduction

Making machine learning models reliable under distributional shift is crucial for real-world applications such as autonomous driving, health risk prediction, and medical imaging. In this work, we consider the task of domain generalization, which aims to obtain models that generalize to unseen domains by learning from a limited set of training domains.

To improve model robustness under domain shifts, a plethora of algorithms have been recently proposed [48, 2, 43, 28, 46]. In particular, methods that learn invariant feature representations or invariant predictors across domains demonstrate promising performance both empirically and theoretically [12, 27, 39, 25]. However, it remains challenging to improve on empirical risk minimization (ERM) when evaluating on a broad range of real-world datasets [16, 20]. Notice that ERM is a fairly reasonable baseline method since it is necessary for ERM to utilize invariant features to achieve optimal in-distribution performance. It has been empirically shown [36] that ERM already learns invariant features sufficient for domain generalization. The main issue ERM faces in domain generalization is that the invariant features learned via ERM can be arbitrarily mixed: environmental features are hard to disentangle from invariant features. Although various regularization techniques that control empirical risks across domains have been proposed [1, 22, 35], few directly regularize ERM. Thus, a natural idea is to identify the subset of solutions from ERM with minimal information retrieved from training domains.

In this work, we propose a simple yet effective algorithm, `ERM-NU` (Empirical Risk Minimization with Nuclear Norm Regularization), for improving domain generalization without acquiring domain annotations. Our method is inspired by works in low-rank matrix completion and recovery with nuclear norm minimization [9, 8, 10, 7, 21, 15]. Given latent feature representations from pre-trained models via ERM, ERM-NU aims to extract class-related (domain-invariant) features by adding a linear projection layer and fine-tuning the network with nuclear norm regularization. Specifically, we propose to minimize the nuclear norm of the projected features, which is a convex envelope to

[∗] Equal Contribution

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the rank of the feature matrix [34]. Empirically, we evaluate the performance of ERM-NU on seven benchmark datasets. Despite its simplicity, ERM-NU demonstrates competitive performance and improves on existing methods on some large-scale datasets such as TerraInc and DomainNet. In particular, ERM-NU achieves the state-of-the-art average accuracy in DomainBed. Theoretically, we show that even training with infinite data from in-domain tasks, ERM with weight decay may perform worse than random guessing on out-of-domain tasks, while ERM with bounded rank (corresponding to ERM-NU) can guarantee 100% test accuracy on the out-of-domain task. Moreover, our method is computationally efficient as it does not require training domain annotations. As a regularization, our method is also potentially orthogonal to other baseline methods besides ERM.

2 Method

Preliminaries. We use $\mathcal{X}$ and $\mathcal{Y}$ to denote the input and label space, respectively. Following [20, 46, 35], we consider data distributions consisting of environments (domains) $\mathcal{E} = \{1, \dots, E\}$. For a given environment $e \in \mathcal{E}$ and label $y \in \mathcal{Y}$, the data generation process is the following: latent **environmental** features $\mathbf{z}_e$ and **invariant** features $\mathbf{z}_c$ are sampled where invariant features only depend on y, while environmental features depend on e and y (i.e., environmental features and the label may have correlations). The input data is generated from the latent features $\mathbf{x} = g(\mathbf{z}_c, \mathbf{z}_e)$ by some injective function g. See illustration in Fig. 1. We assume that the training data is drawn from a mixture of $E^{tr} \subset \mathcal{E}$ domains and test data is drawn from some unseen domain in $E^{ts} \subset \mathcal{E}$. In the domain shift setup, training domains are disjoint from test domains: $\mathcal{E}^{tr} \cap \mathcal{E}^{ts} = \emptyset$. In this work, as we do not require domain annotations for training data, we remove notation involving $\mathcal{E}$ for simplicity and denote the training data distribution as $\mathcal{D}_{\text{id}}$ and the unseen domain test data distribution as $\mathcal{D}_{\text{ood}}$.

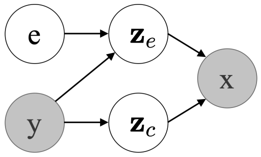

Figure 1: Data distribution inspired by [35]. Shading indicates the variable is observed.

We assume we have access to infinite samples. Our objective is to learn a feature embedder Φ that maps inputs to a d-dimensional feature embedding (usually fine-tuned from a pre-trained backbone, e.g. ResNet [17] pre-trained on ImageNet) and a classifier $\hat{f}$ to minimize the risk on **unseen** environments,

$$\mathcal{L}(\hat{f}, \Phi) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ood}}} \left[\ell(\hat{f}(\Phi(\mathbf{x})), y)\right], \tag{1}$$

where the function ℓ can be any loss appropriate to classification.

Method Overview. Intuitively, to guarantee low risk on $\mathcal{D}_{\text{ood}}$, Φ needs to rely only on invariant features for prediction. It must not use environmental features in order to avoid spurious correlations to ensure domain generalization. If we assume that environmental features do not have a perfect correlation with the label, we can eliminate environmental features by constraining the rank of the learned representations from the training data.

We consider fine-tuning the backbone (feature extractor) Φ with a linear prediction head. Denote the linear head as $\mathbf{a} \in \mathbb{R}^{d\times m}$, where m is the class number. The goal of ERM is to minimize the expected risk

$$\mathcal{L}(\mathbf{a}, \Phi) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{in}}} \left[\ell(\mathbf{a}^\top \Phi(\mathbf{x}), y)\right].$$

The intuition behind our approach is the following. Consider the latent vector $\Phi(\mathbf{x}) \in \mathbb{R}^d$. This vector may contain both environment-related and class-related features. In order to obtain just the class-related features, we would like for Φ to extract as little information as possible while simultaneously optimizing the ERM loss. Note that, we assume that the correlation between environmental features and labels is lower than the correlation between invariant features and labels. Let $\mathbf{X}$ be a batch of training data points (batch size $> d$). To minimize information and so rule out environmental features, we minimize the rank of $\Phi(\mathbf{X})$. Our objective is

$$\min_{\mathbf{a},\Phi} \mathcal{L}(\mathbf{a}, \Phi) + \lambda \text{rank}(\Phi(\mathbf{X})). \tag{2}$$

As the nuclear norm is a convex envelope to the rank of a matrix, our convex relaxation objective is

$$\min_{\mathbf{a},\Phi} \mathcal{L}(\mathbf{a}, \Phi) + \lambda \|\Phi(\mathbf{X})\|_*, \tag{3}$$

where λ is the regularization weight and $\|\|_*$ indicates the nuclear norm.

3 Theoretical Analysis

Next, we present a simple theoretical result showing that, for a particular setting, the ERM-rank solution to (2) is much more robust than the ERM solution.

Data Distribution. Consider binary classification. Let $\mathcal{X}$ be the input space, and $\mathcal{Y} = \{\pm 1\}$ be the label space. Let $\tilde{\mathbf{z}}(\mathbf{x}) \in \mathbb{R}^d$ be a feature pattern encoder. For any $j \in [d]$, assume $\Phi(\mathbf{x})_j = \mathbf{w}_j \tilde{\mathbf{z}}_j(\mathbf{x})$ where $\mathbf{w}_i$ is a scalar and $\tilde{\mathbf{z}}_j$ is a specific feature pattern encoder. Assume $\tilde{\mathbf{z}}(\mathbf{x})$ are drawn from some distribution condition on label y. We denote $\mathbf{z} = \tilde{\mathbf{z}}(\mathbf{x})y$ for simplicity. Let $R \subseteq [d]$ be a subset of size r corresponding to the class-relevant patterns (invariant features $\mathbf{z}_c$ in Figure 1) and $U = [d] \setminus R$ be a subset of size $d - r$ corresponding to the spurious patterns (environmental features $\mathbf{z}_e$ in Figure 1). For simplicity, we assume, for any $j, j' \in [d]$, $\mathbf{z}_j, \mathbf{z}_{j'}$ are independent when $j \neq j'$.

For invariant features, we assume, for any $j \in R$, $\mathbf{z}_j \sim [0, 1]$ uniformly, so $\mathbb{E}[\mathbf{z}_j] = \frac{1}{2}$. Next, we define in-domain (ID) tasks and out-of-domain (OOD) tasks. We first define $\mathcal{D}_\gamma$ where $\gamma \in (-\frac{1}{2}, \frac{1}{2})$. A random variable $z \sim \mathcal{D}_\gamma$ means, $z \sim [0, 1]$ uniformly with probability $\frac{1}{2} + \gamma$ and $z \sim [-1, 0]$ uniformly with probability $\frac{1}{2} - \gamma$, so $\mathbb{E}[z] = \gamma$. In ID tasks, for environmental features, for any $j \in U$, we assume that $\mathbf{z}_j \sim \mathcal{D}_\gamma = \mathcal{D}_{\text{id}}$ where $\gamma \in \left(\frac{3}{\sqrt{r}}, \frac{1}{2}\right)$. In OOD tasks, for any $j \in U$, we assume that $\mathbf{z}_j \sim \mathcal{D}_{-\gamma} = \mathcal{D}_{\text{ood}}$.

Objectives. We simplify the fine-tuning process, setting $\mathbf{a} = [1, 1, \dots, 1]^\top$ and the trainable parameter to be $\mathbf{w}$ (varying the impact of each feature). Thus, the network output is $f_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{d} \mathbf{w}_j \tilde{\mathbf{z}}_j(\mathbf{x})$. We consider two objective functions. The first is traditional ERM with weight decay (ℓ_2 norm regularization). The ERM-ℓ_2 objective function is

$$\min_{\mathbf{w}} \mathcal{L}^\lambda(\mathbf{w}) := \mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \tag{4}$$

where $\mathcal{L}_{(\mathbf{x},y)}(\mathbf{w}) = \ell(y f_{\mathbf{w}}(\mathbf{x}))$ is the loss on an example $(\mathbf{x}, y)$ and $\ell(z)$ is the logistic loss $\ell(z) = \ln(1 + \exp(-z))$.

The second objective we consider is ERM with bounded rank. For a batch $\mathbf{X}$ with batch size $> d$, it is full rank with probability 1. Thus, we say the total feature rank is $\|\mathbf{w}\|_0 \leq d$ ($\|\mathbf{w}\|_0$ indicates the number of nonzero elements in $\mathbf{w}$). Thus, the ERM-rank objective function is

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \quad \text{subject to} \quad \|\mathbf{w}\|_0 \leq B_{\text{rank}}. \tag{5}$$

The ERM-rank objective function is equivalent to Eq. (2).

Proposition 1. *Assume* $1 \leq B_{\text{rank}} \leq r, \lambda > \Omega\left(\frac{\sqrt{r}}{\exp\left(\frac{\sqrt{r}}{5}\right)}\right), d > \frac{r}{\gamma^2} + r, r > C$*, where* C *is some constant* < 20*. The optimal solution for the ERM-rank objective function on the ID tasks has 100% OOD test accuracy, while the optimal solution for the ERM-*ℓ_2 *objective function on the ID tasks has OOD test accuracy at most* $\exp\left(-\frac{r}{10}\right) \times 100\%$ *(much worse than random guessing).*

Discussion. The assumption of λ and d means that the regularization strength cannot be too small and the environmental features signal level should be compatible with invariant features signal level. Then, Proposition 1 shows that even with infinite data, the optimal solution for ERM-ℓ_2 on the ID tasks cannot produce better performance than random guessing on the OOD task. However, the optimal solution for ERM-rank on the ID tasks can still produce 100% test accuracy. The proof idea is that the ERM-ℓ_2 objective will encode all features correlated with labels, even when the correlation between spurious feature and label is weak (e.g. $\gamma \in O\left(1/\sqrt{r}\right)$). Thus, when the OOD tasks have a different spurious feature distribution, the optimal solution of ERM-ℓ_2 objective may thoroughly fail i.e. much worse than random guessing. However, the optimal solution of the ERM-rank objective will only encode the features which have a strong correlation with labels (intrinsic features). Thus, it can guarantee 100% test accuracy on OOD tasks. See the full proof in Appendix A.

4 Experiments

Experimental Setup. For a fair comparison with baseline methods, we evaluate our algorithm on the DomainBed testbed [16], an open-source benchmark that aims to rigorously compare different algorithms for domain generalization. The testbed consists of a wide range of datasets for multi-domain image classification tasks, including Colored MNIST [1], Rotated MNIST [14], PACS [24], VLCS [11], Office-Home [42], Terra Incognita [5], and DomainNet [31]. For the model selection criterion, we use the "training-domain validation set" strategy, which refers to choosing the model maximizing the accuracy on the overall validation set pooled from each training domain. For each dataset and model, we report test domains accuracy of the best-selected model (average over three independent runs). We use ResNet-50 [17] as the feature backbone and fine-tune the whole model. For the weight scale λ, we set the default value as 0.01 and distributions for random search as $10^{\text{Uniform }(-2.5,-1.5)}$. The default batch size is 32 and the distribution for random search is $2^{\text{Uniform }(5,6)}$. During training, similar to [1] which uses batch-wise statistics for invariant risk minimization, we perform batch-wise nuclear norm regularization.

Algorithm	**clip**	**info**	**paint**	**quick**	**real**	**sketch**	**Average**
ERM [41]	58.1 ± 0.3	18.8 ± 0.3	46.7 ± 0.3	12.2 ± 0.4	59.6 ± 0.1	49.8 ± 0.4	40.9
IRM [1]	48.5 ± 2.8	15.0 ± 1.5	38.3 ± 4.3	10.9 ± 0.5	48.2 ± 5.2	42.3 ± 3.1	33.9
GroupDRO [37]	47.2 ± 0.5	17.5 ± 0.4	33.8 ± 0.5	9.3 ± 0.3	51.6 ± 0.4	40.1 ± 0.6	33.3
Mixup [45]	55.7 ± 0.3	18.5 ± 0.5	44.3 ± 0.5	12.5 ± 0.4	55.8 ± 0.3	48.2 ± 0.5	39.2
MLDG [23]	59.1 ± 0.2	19.1 ± 0.3	45.8 ± 0.7	$\underline{13.4}$ ± 0.3	59.6 ± 0.2	50.2 ± 0.4	41.2
CORAL [39]	$\underline{59.2}$ ± 0.1	19.7 ± 0.2	46.6 ± 0.3	$\underline{13.4}$ ± 0.4	59.8 ± 0.2	50.1 ± 0.6	41.5
MMD [25]	32.1 ± 13.3	11.0 ± 4.6	26.8 ± 11.3	8.7 ± 2.1	32.7 ± 13.8	28.9 ± 11.9	23.4
DANN [12]	53.1 ± 0.2	18.3 ± 0.1	44.2 ± 0.7	11.8 ± 0.1	55.5 ± 0.4	46.8 ± 0.6	38.3
CDANN [27]	54.6 ± 0.4	17.3 ± 0.1	43.7 ± 0.9	12.1 ± 0.7	56.2 ± 0.4	45.9 ± 0.5	38.3
MTL [6]	57.9 ± 0.5	18.5 ± 0.4	46.0 ± 0.1	12.5 ± 0.1	59.5 ± 0.3	49.2 ± 0.1	40.6
SagNet [29]	57.7 ± 0.3	19.0 ± 0.2	45.3 ± 0.3	12.7 ± 0.5	58.1 ± 0.5	48.8 ± 0.2	40.3
ARM [49]	49.7 ± 0.3	16.3 ± 0.5	40.9 ± 1.1	9.4 ± 0.1	53.4 ± 0.4	43.5 ± 0.4	35.5
VREx [22]	47.3 ± 3.5	16.0 ± 1.5	35.8 ± 4.6	10.9 ± 0.3	49.6 ± 4.9	42.0 ± 3.0	33.6
RSC [18]	55.0 ± 1.2	18.3 ± 0.5	44.4 ± 0.6	12.2 ± 0.2	55.7 ± 0.7	47.8 ± 0.9	38.9
AND-mask [30]	52.3 ± 0.8	16.6 ± 0.3	41.6 ± 1.1	11.3 ± 0.1	55.8 ± 0.4	45.4 ± 0.9	37.2
SelfReg [19]	58.5 ± 0.1	$\underline{20.7}$ ± 0.1	47.3 ± 0.3	13.1 ± 0.3	58.2 ± 0.2	$\underline{51.1}$ ± 0.3	41.5
Fishr [33]	58.2 ± 0.5	20.2 ± 0.2	$\underline{47.7}$ ± 0.3	12.7 ± 0.2	$\underline{60.3}$ ± 0.2	50.8 ± 0.1	$\underline{41.7}$
ERM-NU (ours)	**60.9** ± 0.0	**21.1** ± 0.2	**49.9** ± 0.3	**13.7** ± 0.2	**62.5** ± 0.2	**52.5** ± 0.4	**43.4**

Table 1: Results on DomainNet. For each column, bold indicates the best performance, and underline indicates the second-best performance.

Algorithm	**L100**	**L38**	**L43**	**L46**	**Avg**
ERM	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
IRM	$\underline{54.6}$ ± 1.3	39.8 ± 1.9	56.2 ± 1.8	39.6 ± 0.8	47.6
GroupDRO	41.2 ± 0.7	38.6 ± 2.1	56.7 ± 0.9	36.4 ± 2.1	43.2
Mixup	**59.6** ± 2.0	42.2 ± 1.4	55.9 ± 0.8	33.9 ± 1.4	47.9
MLDG	54.2 ± 3.0	$\underline{44.3}$ ± 1.1	55.6 ± 0.3	36.9 ± 2.2	47.7
CORAL	51.6 ± 2.4	42.2 ± 1.0	57.0 ± 1.0	39.8 ± 2.9	47.6
MMD	41.9 ± 3.0	34.8 ± 1.0	57.0 ± 1.9	35.2 ± 1.8	42.2
DANN	51.1 ± 3.5	40.6 ± 0.6	57.4 ± 0.5	37.7 ± 1.8	46.7
CDANN	47.0 ± 1.9	41.3 ± 4.8	54.9 ± 1.7	39.8 ± 2.3	45.8
MTL	49.3 ± 1.2	39.6 ± 6.3	55.6 ± 1.1	37.8 ± 0.8	45.6
SagNet	53.0 ± 2.9	43.0 ± 2.5	$\underline{57.9}$ ± 0.6	40.4 ± 1.3	$\underline{48.6}$
ARM	49.3 ± 0.7	38.3 ± 2.4	55.8 ± 0.8	38.7 ± 1.3	45.5
VREx	48.2 ± 4.3	41.7 ± 1.3	56.8 ± 0.8	38.7 ± 3.1	46.4
RSC	50.2 ± 2.2	39.2 ± 1.4	56.3 ± 1.4	**40.8** ± 0.6	46.6
AND-mask	50.0 ± 2.9	40.2 ± 0.8	53.3 ± 0.7	34.8 ± 1.9	44.6
SelfReg	48.8 ± 0.9	41.3 ± 1.8	57.3 ± 0.7	40.6 ± 0.9	47.0
Fishr	50.2 ± 3.9	43.9 ± 0.8	55.7 ± 2.2	39.8 ± 1.0	47.4
Ours	52.5 ± 1.2	**45.0** ± 0.5	**60.2** ± 0.2	$\underline{40.7}$ ± 1.0	**49.6**

Table 2: Results on Terra Incognita.

ERM-NU achieves state-of-the-art on DomainNet & Terra Incognita. The results for each domain in DomainNet and Terra Incognita are shown in Table 1 and Table 2 respectively. For the DomainNet, our method achieves state-of-the-art performance, beating baselines on all domains, including prior invariance-learning methods such as IRM, VREx, and DANN. For example, compared to IRM, ERM-NU significantly improves the average accuracy by 9.5%. For the Terra Incognita, we can see a similar phenomenon. This validates the notion that nuclear norm-based regularization effectively promotes learning invariant features. An additional benefit, compared to IRM, is that ERM-NU does not require any domain annotations.

ERM-NU remains competitive across a wide range of datasets. The results for all datasets in DomainBed are shown in Table 3. We observe that ERM-NU remains very competitive across most

Algorithm	**CMNIST**	**RMNIST**	**VLCS**	**PACS**	**OfficeHome**	**TerraInc**	**DomainNet**	**Average**
ERM	51.5 ± 0.1	$\underline{98.0}$ ± 0.0	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.1	66.6
IRM	52.0 ± 0.1	97.7 ± 0.1	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	47.6 ± 0.8	33.9 ± 2.8	65.4
GroupDRO	$\underline{52.1}$ ± 0.0	$\underline{98.0}$ ± 0.0	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	43.2 ± 1.1	33.3 ± 0.2	64.8
Mixup	$\underline{52.1}$ ± 0.2	$\underline{98.0}$ ± 0.1	77.4 ± 0.6	84.6 ± 0.6	$\underline{68.1}$ ± 0.3	47.9 ± 0.8	39.2 ± 0.1	66.7
MLDG	51.5 ± 0.1	97.9 ± 0.0	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	47.7 ± 0.9	41.2 ± 0.1	66.7
CORAL	51.5 ± 0.1	$\underline{98.0}$ ± 0.1	**78.8** ± 0.6	$\underline{86.2}$ ± 0.3	**68.7** ± 0.3	47.6 ± 1.0	41.5 ± 0.1	$\underline{67.5}$
MMD	51.5 ± 0.2	97.9 ± 0.0	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	42.2 ± 1.6	23.4 ± 9.5	63.3
DANN	51.5 ± 0.3	97.8 ± 0.1	$\underline{78.6}$ ± 0.4	83.6 ± 0.4	65.9 ± 0.6	46.7 ± 0.5	38.3 ± 0.1	66.1
CDANN	51.7 ± 0.1	97.9 ± 0.1	77.5 ± 0.1	82.6 ± 0.9	65.8 ± 1.3	45.8 ± 1.6	38.3 ± 0.3	65.6
MTL	51.4 ± 0.1	97.9 ± 0.0	77.2 ± 0.4	84.6 ± 0.5	66.4 ± 0.5	45.6 ± 1.2	40.6 ± 0.1	66.2
SagNet	51.7 ± 0.0	$\underline{98.0}$ ± 0.0	77.8 ± 0.5	**86.3** ± 0.2	$\underline{68.1}$ ± 0.1	$\underline{48.6}$ ± 1.0	40.3 ± 0.1	67.2
ARM	**56.2** ± 0.2	**98.2** ± 0.1	77.6 ± 0.3	85.1 ± 0.4	64.8 ± 0.3	45.5 ± 0.3	35.5 ± 0.2	66.1
VREx	51.8 ± 0.1	97.9 ± 0.1	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	46.4 ± 0.6	33.6 ± 2.9	65.6
RSC	51.7 ± 0.2	97.6 ± 0.1	77.1 ± 0.5	85.2 ± 0.9	65.5 ± 0.9	46.6 ± 1.0	38.9 ± 0.5	66.1
AND-mask	51.3 ± 0.2	97.6 ± 0.1	78.1 ± 0.9	84.4 ± 0.9	65.6 ± 0.4	44.6 ± 0.3	37.2 ± 0.6	65.5
SelfReg	$\underline{52.1}$ ± 0.2	$\underline{98.0}$ ± 0.1	77.8 ± 0.9	85.6 ± 0.4	67.9 ± 0.7	47.0 ± 0.3	41.5 ± 0.2	67.1
Fishr	52.0 ± 0.2	97.8 ± 0.0	77.8 ± 0.1	85.5 ± 0.4	67.8 ± 0.1	47.4 ± 1.6	$\underline{41.7}$ ± 0.0	67.1
ERM-NU (ours)	51.8 ± 0.2	$\underline{98.0}$ ± 0.1	77.8 ± 0.7	85.6 ± 0.1	$\underline{68.1}$ ± 0.1	**49.6** ± 0.6	**43.4** ± 0.1	**67.8**

Table 3: Results on DomainBed benchmark.

datasets. For example, ERM-NU improves over IRM by 2.4% when averaged over seven datasets, still achieving the best performance compared to other competitive baselines.

5 Related Works

Nuclear Norm Minimization. Nuclear norm is commonly used to approximate the matrix rank [34]. Nuclear norm minimization has been widely applied for low-rank matrix completion and recovery [9, 8, 10, 7] with applications such as graph clustering [21], community detection [26] and image denoising [15].

Contextual Bias in Recognition. There has been a rich literature studying the classification performance in the presence of pre-defined contextual bias [40, 5, 4]. The reliance on contextual bias such as image backgrounds, texture, and color for object detection are investigated in [50, 3, 13, 47, 44, 37]. In contrast, our study requires no prior information on the type of contextual bias and is broadly applicable to different categories of bias.

Domain Generalization. The task of domain generalization aims to achieve high classification accuracy on new test environments. A plethora of algorithms are proposed in recent years: learning invariant representation across domains [12, 27, 39, 25], minimizing the weighted combination of risks from training domains [37], using different risk penalty terms to facilitate invariance prediction [1, 22], causal inference approaches [32], and forcing the learned representation different from a set of pre-defined biased representations [2], mixup-based approaches [48, 43, 28, 46], etc. However, it remains challenging to improve over empirical risk minimization (ERM) when evaluated on a broad range of real-world datasets [16, 20].

6 Conclusions and Outlook

In this work, we propose ERM-NU, a simple yet effective regularization method based on empirical risk minimization for improving domain generalization without acquiring domain annotations. Key to our method is minimizing the nuclear norm of the feature embeddings as a convex proxy for rank minimization. Empirically, we show that ERM-NU achieves competitive performance across a wide range of datasets in the DomainBed benchmarks when compared to a large number of baselines. Theoretically, we show that it outperforms ERM and ERM with L2 regularization in the linear setting. We aim to generalize this result to nonlinear cases in the future. We hope our work will inspire effective algorithm design and a better understanding of domain generalization.

7 Acknowledgements

The work is partially supported by Air Force Grant FA9550-18-1-0166, the National Science Foundation (NSF) Grants 2008559-IIS and CCF-2046710.

References

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

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020.

[3] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12):1–43, 12 2018.

[4] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32:9453–9463, 2019.

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[6] Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100, 2021.

[7] Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[8] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[9] Emmanuel J Candes and Justin Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Computational Mathematics*, 6(2):227–254, 2006.

[10] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[11] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[14] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

[15] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014.

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

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020.

[19] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021.

[20] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[21] Ramya Korlakai Vinayak, Samet Oymak, and Babak Hassibi. Graph clustering with missing data: Convex algorithms and analysis. *Advances in Neural Information Processing Systems*, 27, 2014.

[22] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[25] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[26] Xiaodong Li, Yudong Chen, and Jiaming Xu. Convex relaxation methods for community detection. *Statistical Science*, 36(1):2–15, 2021.

[27] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[28] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In *European Conference on Computer Vision*, 2020.

[29] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.

[30] Giambattista Parascandolo, Alexander Neitz, ANTONIO ORVIETO, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *International Conference on Learning Representations*, 2020.

[31] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[32] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[33] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022.

[34] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[35] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9, 2021.

[36] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022.

[37] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations, ICLR*, 2019.

[38] Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha. The trade-off between label efficiency and universality of representations from contrastive learning. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022.

[39] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[40] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003.

[41] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[42] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[43] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020.

[44] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021.

[45] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

[46] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *Proceeding of the Thirty-ninth International Conference on Machine Learning*, 2022.

[47] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11), 2018.

[48] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[49] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 8:9, 2020.

[50] Zhuotun Zhu, Lingxi Xie, and Alan Yuille. Object recognition with and without objects. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3609–3615, 2017.

Appendix

A Proof of Theoretical Analysis

A.1 Auxiliary Lemmas

Lemma 2. *For the logistic loss $\ell(z) = \ln(1 + \exp(-z))$, we have the following statements (1) $\ell(z)$ is strictly decreasing and convex function on $\mathbb{R}$ and $\ell(z) > 0$; (2) $\ell'(z) = \frac{-1}{1+\exp(z)}$, $\ell'(z) \in (-1, 0)$; (3) $\ell'(z)$ is strictly concave on $[0, +\infty)$, (4) for any $c > 0$, $\ell'(z + c) \leq \exp(-c)\ell'(z)$.*

Proof. These can be verified by direct calculation. □

Lemma 3.

$$\frac{\partial \mathcal{L}_{(\mathbf{x},y)}(\mathbf{w})}{\partial \mathbf{w}_j} = \ell'(yf_{\mathbf{w}}(\mathbf{x}))\mathbf{z}_j, \tag{6}$$

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}_j} = \mathbb{E}_{(\mathbf{x},y)}\left[\ell'(yf_{\mathbf{w}}(\mathbf{x}))\mathbf{z}_j\right] \tag{7}$$

$$\frac{\partial \mathcal{L}^{\lambda}(\mathbf{w})}{\partial \mathbf{w}_j} = \mathbb{E}_{(\mathbf{x},y)}\left[\ell'(yf_{\mathbf{w}}(\mathbf{x}))\mathbf{z}_j\right] + \lambda \mathbf{w}_j \tag{8}$$

Proof. These can be verified by direct calculation. □

Lemma 4. *For any $j \in R$, we have probability density function of $\mathbf{z}_j$ with mean $\frac{1}{2}$ and variance $\frac{1}{12}$ following the form*

$$f_{\{\mathbf{z}_j\}}(z) = \begin{cases} 1, & \textit{if} \quad 0 \leq z \leq 1 \\ 0, & \textit{otherwise} \; . \end{cases}$$

For any $j \in U$, we have probability density function of $\mathbf{z}_j$ with mean γ and variance $\frac{1}{3} - \gamma^2$ following the form

$$f_{\{\mathbf{z}_j\}}(z) = \begin{cases} \frac{1}{2} - \gamma, & \textit{if} \quad -1 \leq z < 0 \\ \frac{1}{2} + \gamma, & \textit{if} \quad 0 \leq z \leq 1 \\ 0, & \textit{otherwise} \; . \end{cases}$$

Proof. Then these can be verified by direct calculation from the definition. □

Lemma 5. *We have $\mathbb{P}\left[\sum_{j\in U} \mathbf{z}_j \leq 0\right] \leq \exp\left(-\frac{(d-r)\gamma^2}{2}\right)$, $\mathbb{P}\left[\sum_{j\in R} \mathbf{z}_j \leq \frac{r}{4}\right] \leq \exp\left(-\frac{r}{8}\right)$.*

Proof. By Hoeffding's inequality,

$$\mathbb{P}\left[\sum_{j\in U} \mathbf{z}_j \leq 0\right] = \mathbb{P}\left[\sum_{j\in U} (\mathbf{z}_j - \gamma) \leq -(d-r)\gamma\right] \tag{9}$$

$$\leq \exp\left(-\frac{(d-r)\gamma^2}{2}\right). \tag{10}$$

The others are proven in a similar way. □

A.2 Optimal Solution of ERM-ℓ_2 on ID Task

In this section, we will analyze the property of the optimal solution of ERM-ℓ_2 on the ID task. We will show that the ERM-ℓ_2 objective will encode all features correlated with labels, even when the correlation between spurious feature and label is weak (e.g. $\gamma \in O\left(1/\sqrt{r}\right)$).

Following the idea from Lemma B.1 of [38], we have the Lemma below.

Lemma 6. *Consider the ID setting with the ERM-ℓ_2 objective function. Then the optimal $\mathbf{w}^*$ for the ERM-ℓ_2 objective function following the condition (1) for any $j \in R$, $\mathbf{w}_j^* =: \alpha$ and (2) for any $j \in U$, $\mathbf{w}_j^* := \beta$.*

Proof.

$$\begin{aligned}\mathcal{L}^\lambda(\mathbf{w}^*) =& \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{id}}}\mathcal{L}_{(\mathbf{x},y)}(\mathbf{w}^*) + \frac{\lambda}{2}\|\mathbf{w}^*\|_2^2 \\ =& \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{id}}}\ell(yf_{\mathbf{w}^*}(\mathbf{x})) + \frac{\lambda}{2}\|\mathbf{w}^*\|_2^2 \\ =& \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{id}}}\ell\left(\sum_{j=1}^d \mathbf{w}_j^*\mathbf{z}_j\right) + \frac{\lambda}{2}\|\mathbf{w}^*\|_2^2\end{aligned}$$

By Lemma 2, we have $\mathcal{L}^\lambda(\mathbf{w})$ a is convex function. By symmetry of $\mathbf{z}_j$, for any $l, l' \in R, l \neq l'$,

$$\begin{aligned}&\mathbb{E}\left[\ell\left(\sum_{j=1}^d \mathbf{w}_j^*\mathbf{z}_j\right)\right] + \frac{\lambda}{2}\|\mathbf{w}^*\|_2^2 &(11)\\ =&\frac{1}{2}\left(\mathbb{E}\left[\ell\left(\sum_{j\in[d],j\neq l,j\neq l'}\mathbf{w}_j^*\mathbf{z}_j + \mathbf{w}_l^*\mathbf{z}_l(\mathbf{x},y) + \mathbf{w}_{l'}^*\mathbf{z}_{l'}(\mathbf{x},y)\right)\right] + \frac{\lambda}{2}\|\mathbf{w}^*\|_2^2\right) &(12)\\ +&\frac{1}{2}\left(\mathbb{E}\left[\ell\left(\sum_{j\in[d],j\neq l,j\neq l'}\mathbf{w}_j^*\mathbf{z}_j + \mathbf{w}_l^*\mathbf{z}_{l'}(\mathbf{x},y) + \mathbf{w}_{l'}^*\mathbf{z}_{l}(\mathbf{x},y)\right)\right] + \frac{\lambda}{2}\|\mathbf{w}^*\|_2^2\right) &(13)\\ \geq&\mathbb{E}\left[\ell\left(\sum_{j\in[d],j\neq l,j\neq l'}\mathbf{w}_j^*\mathbf{z}_j + \frac{\mathbf{w}_l^*+\mathbf{w}_{l'}^*}{2}\mathbf{z}_{l'}(\mathbf{x},y) + \frac{\mathbf{w}_l^*+\mathbf{w}_{l'}^*}{2}\mathbf{z}_{l}(\mathbf{x},y)\right)\right] + \frac{\lambda}{2}\|\mathbf{w}^*\|_2^2, &(14)\end{aligned}$$

where the last inequality follows Jensen's inequality. The minimum is achieved when $\mathbf{w}_l^* = \mathbf{w}_{l'}^*$. A similar argument as above proves statement (2). □

Now, we will bound the α and β. Recall that for any $j \in R$, $\mathbf{w}_j^* =: \alpha$ and for any $j \in U$, $\mathbf{w}_j^* := \beta$. The proof idea is using the gradient equal to zero and the properties of the logistic loss. Then, we can show that the value of β is compatible with the value of α.

Lemma 7. *Let α, β be values defined in the Lemma 6. Then, we have $0 < \beta < \alpha < \frac{1}{\sqrt{r}}$. Moreover, $\frac{\alpha}{\beta} < \frac{3}{4\gamma}$.*

Proof. By Lemma 6

$$\begin{aligned}\mathcal{L}^\lambda(\mathbf{w}^*) =& \mathbb{E}\left[\ell\left(\alpha\sum_{j\in R}\mathbf{z}_j + \beta\sum_{j\in U}\mathbf{z}_j\right)\right] + \frac{\lambda}{2}(r\alpha^2 + (d-r)\beta^2) &(15)\\ =&\mathcal{L}^\lambda(\alpha,\beta). &(16)\end{aligned}$$

By Lemma 3, we have for any $j \in [d]$

$$\frac{\partial\mathcal{L}^\lambda(\mathbf{w}^*)}{\partial\mathbf{w}_j^*} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{id}}}\left[\ell'(yf_{\mathbf{w}}^*(\mathbf{x}))\mathbf{z}_j\right] + \lambda\mathbf{w}_j^* = 0. \quad (17)$$

We first prove $\beta < \alpha$. For any $j \in R, j' \in U$, we have

$$\begin{aligned}\lambda\alpha =& \lambda\mathbf{w}_j^* &(18)\\ =& -\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{id}}}\left[\ell'(yf_{\mathbf{w}}^*(\mathbf{x}))\mathbf{z}_j\right] &(19)\\ >& -\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{id}}}\left[\ell'(yf_{\mathbf{w}}^*(\mathbf{x}))\mathbf{z}_{j'}(\mathbf{x},y)\right] &(20)\\ =& \lambda\mathbf{w}_{j'}^* = \lambda\beta. &(21)\end{aligned}$$

Then, we prove $\beta \geq 0$ by contradiction. Suppose $\beta < 0$,

$$\mathcal{L}^{\lambda}(\alpha, \beta) - \mathcal{L}^{\lambda}(\alpha, -\beta) \tag{22}$$

$$= \mathbb{E}\left[\ell\left(\alpha \sum_{j \in R} \mathbf{z}_j + \beta \sum_{j \in U} \mathbf{z}_j\right)\right] - \mathbb{E}\left[\ell\left(\alpha \sum_{j \in R} \mathbf{z}_j - \beta \sum_{j \in U} \mathbf{z}_j\right)\right]. \tag{23}$$

Note that for any $j, j' \in U, j \neq j'$, the norm of $\mathbf{z}_j$ is independent with its sign and $\mathbf{z}_j, \mathbf{z}_{j'}(\mathbf{x}, y)$ are independent. From $\gamma > 0$, we can get $\mathbb{P}[\mathbf{z}_j > 0] > \frac{1}{2}$. Thus, by ℓ strictly decreasing we have

$$\mathbb{P}\left[\ell\left(\alpha \sum_{j \in R} \mathbf{z}_j + \beta \sum_{j \in U} \mathbf{z}_j\right) \geq z\right] > \mathbb{P}\left[\ell\left(\alpha \sum_{j \in R} \mathbf{z}_j - \beta \sum_{j \in U} \mathbf{z}_j\right) \geq z\right], \tag{24}$$

where β case is strictly stochastically dominate $-\beta$ case. Thus, $\mathcal{L}^{\lambda}(\alpha, \beta) - \mathcal{L}^{\lambda}(\alpha, -\beta) > 0$. This is contradicted by β being the optimal value. Thus, we have $\beta \geq 0$.

Now, we prove $\alpha < \frac{1}{\sqrt{r}}$, for any $k \in R$,

$$\lambda\alpha = -\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_{\text{id}}}\left[\ell'\left(\alpha \sum_{j \in R} \mathbf{z}_j + \beta \sum_{j \in U} \mathbf{z}_j\right) \mathbf{z}_k\right] \tag{25}$$

$$\leq -\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_{\text{id}}}\left[\ell'\left(\alpha \sum_{j \in R, j \neq k} \mathbf{z}_j + \beta \sum_{j \in U} \mathbf{z}_j\right) \mathbf{z}_k\right] \tag{26}$$

$$= -\mathbb{E}\left[\ell'\left(\alpha \sum_{j \in R, j \neq k} \mathbf{z}_j + \beta \sum_{j \in U} \mathbf{z}_j\right)\right] \mathbb{E}[\mathbf{z}_k] \tag{27}$$

$$= -\frac{1}{2}\mathbb{E}\left[\ell'\left(\alpha \sum_{j \in R, j \neq k} \mathbf{z}_j + \beta \sum_{j \in U} \mathbf{z}_j\right) \middle| \sum_{j \in U} \mathbf{z}_j > 0\right] \mathbb{P}\left[\sum_{j \in U} \mathbf{z}_j > 0\right] \tag{28}$$

$$-\frac{1}{2}\mathbb{E}\left[\ell'\left(\alpha \sum_{j \in R, j \neq k} \mathbf{z}_j + \beta \sum_{j \in U} \mathbf{z}_j\right) \middle| \sum_{j \in U} \mathbf{z}_j \leq 0\right] \mathbb{P}\left[\sum_{j \in U} \mathbf{z}_j \leq 0\right] \tag{29}$$

$$\leq -\frac{1}{2}\mathbb{E}\left[\ell'\left(\alpha \sum_{j \in R, j \neq k} \mathbf{z}_j\right)\right] + \frac{1}{2}\exp\left(-\frac{(d-r)\gamma^2}{2}\right), \tag{30}$$

where the last inequality is from $\beta \geq 0$ and $\ell'(z) \in (-1, 0)$. Using Lemma 5 one more time, we have

$$-\mathbb{E}\left[\ell'\left(\alpha \sum_{j \in R, j \neq k} \mathbf{z}_j\right)\right] \tag{31}$$

$$= -\mathbb{E}\left[\ell'\left(\alpha \sum_{j \in R, j \neq k} \mathbf{z}_j \middle| \sum_{j \in R, j \neq k} \mathbf{z}_j > \frac{r-1}{4}\right)\right] \mathbb{P}\left[\sum_{j \in R, j \neq k} \mathbf{z}_j > \frac{r-1}{4}\right] \tag{32}$$

$$-\mathbb{E}\left[\ell'\left(\alpha \sum_{j \in R, j \neq k} \mathbf{z}_j \middle| \sum_{j \in R, j \neq k} \mathbf{z}_j \leq \frac{r-1}{4}\right)\right] \mathbb{P}\left[\sum_{j \in R, j \neq k} \mathbf{z}_j \leq \frac{r-1}{4}\right] \tag{33}$$

$$\leq -\ell'\left(\frac{\alpha(r-1)}{4}\right) + \frac{1}{2}\exp\left(-\frac{r-1}{8}\right) \tag{34}$$

$$= \frac{1}{1 + \exp\left(\frac{\alpha(r-1)}{4}\right)} + \frac{1}{2}\exp\left(-\frac{r-1}{8}\right). \tag{35}$$

Thus, we have

$$\lambda\alpha \leq \frac{1}{2\left(1+\exp\left(\frac{\alpha(r-1)}{4}\right)\right)} + \frac{1}{4}\exp\left(-\frac{r-1}{8}\right) + \frac{1}{2}\exp\left(-\frac{(d-r)\gamma^2}{2}\right). \tag{36}$$

Suppose $\alpha \geq \frac{1}{\sqrt{r}}$, we have contradiction,

$$\text{RHS} < O\left(\exp\left(-\frac{\sqrt{r}}{5}\right)\right) < \text{LHS}. \tag{37}$$

Thus, we get $\alpha < \frac{1}{\sqrt{r}}$.

Now, we prove $\frac{\alpha}{\beta} \leq \frac{3}{4\gamma}$, for any $k \in R, l \in U$, denote $Z = \alpha \sum_{j\in R, j\neq k} \mathbf{z}_j + \beta \sum_{j \in U, j \neq l} \mathbf{z}_j$, by Lemma 2, we have

$$\begin{aligned}
\frac{\alpha}{\beta} &= \frac{-\mathbb{E}\left[\ell'\left(\alpha\sum_{j\in R}\mathbf{z}_j + \beta\sum_{j\in U}\mathbf{z}_j\right)\mathbf{z}_k\right]}{-\mathbb{E}\left[\ell'\left(\alpha\sum_{j\in R}\mathbf{z}_j + \beta\sum_{j\in U}\mathbf{z}_j\right)\mathbf{z}_l\right]} && (38)\\
&\leq \frac{-\mathbb{E}\left[\ell'\left(Z\right)\mathbf{z}_k\right]}{-\mathbb{E}\left[\ell'\left(Z+2\alpha\right)\mathbf{z}_l | \mathbf{z}_l \geq 0\right]\mathbb{P}[\mathbf{z}_l \geq 0] - \mathbb{E}\left[\ell'\left(Z\right)\mathbf{z}_l | \mathbf{z}_l < 0\right]\mathbb{P}[\mathbf{z}_l < 0]} && (39)\\
&= \frac{-\mathbb{E}\left[\ell'\left(Z\right)\right]}{-\mathbb{E}\left[\ell'\left(Z+2\alpha\right)\right]\left(\frac{1}{2}+\gamma\right) + \mathbb{E}\left[\ell'\left(Z\right)\right]\left(\frac{1}{2}-\gamma\right)} && (40)\\
&\leq \frac{-\mathbb{E}\left[\ell'\left(Z\right)\right]}{-\exp(-2\alpha)\mathbb{E}\left[\ell'\left(Z\right)\right]\left(\frac{1}{2}+\gamma\right) + \mathbb{E}\left[\ell'\left(Z\right)\right]\left(\frac{1}{2}-\gamma\right)} && (41)\\
&= \frac{1}{\exp(-2\alpha)\left(\frac{1}{2}+\gamma\right) - \left(\frac{1}{2}-\gamma\right)} && (42)\\
&\leq \frac{1}{\exp\left(\frac{-2}{\sqrt{r}}\right)\left(\frac{1}{2}+\gamma\right) - \left(\frac{1}{2}-\gamma\right)} && (43)\\
&\leq \frac{1}{2\gamma - \left(1 - \exp\left(\frac{-2}{\sqrt{r}}\right)\right)} && (44)\\
&\leq \frac{1}{2\gamma - \frac{2}{\sqrt{r}}} && (45)\\
&< \frac{3}{4\gamma}, && (46)
\end{aligned}$$

where the second inequality follows Lemma 2 and the second last inequality follows $1 + z \leq \exp(z)$ for $z \in \mathbb{R}$ and $\gamma > \frac{3}{\sqrt{r}}$. □

A.3 Optimal Solution of ERM-rank on ID Task

We show that the optimal solution of the ERM-rank objective will only encode the features which have a strong correlation with labels (intrinsic features).

Lemma 8. *Consider ID setting with ERM-rank objective function. Denote R_{rank} is any subset of R with size $|R_{\text{rank}}| = B_{\text{rank}}$, we have an optimal $\mathbf{w}^*$ for the ERM-rank objective function following the condition (1) for any $j \in R_{\text{rank}}$, $\mathbf{w}_j^* > 0$ and (2) for any $j \notin R_{\text{rank}}$, $\mathbf{w}_j^* = 0$.*

Proof. For any $j \in U$, if $\mathbf{w}_j^* = \theta \neq 0$, there exists $k \in R$ s.t. $\mathbf{w}_k^* = 0$ by objective function condition. When we reassign $\mathbf{w}_j^* = 0, \mathbf{w}_k^* = |\theta|$, the objective function becomes smaller. This is a contradiction. Thus, we finish the proof. □

A.4 OOD Gap Between Two Objective Function

Based on the property of two optimal solutions, we will show the performance gap between these two optimal solutions on the OOD task. The idea is that the spurious features may change their correlation to the labels in different tasks.

Proposition 9 (Restatement of Proposition 1)**.** *Assume* $1 \leq B_{\text{rank}} \leq r, \lambda > \Omega\left(\frac{\sqrt{r}}{\exp\left(\frac{\sqrt{r}}{5}\right)}\right), d > \frac{r}{\gamma^2} + r, r > C$*, where* C *is some constant* < 20*. The optimal solution for the ERM-rank objective function on the ID tasks has 100% OOD test accuracy, while the optimal solution for the ERM-*ℓ_2 *objective function on the ID tasks has OOD test accuracy at most* $\exp\left(-\frac{r}{10}\right) \times 100\%$ *(much worse than random guessing).*

Proof. We denote $\mathbf{w}^*_{rank}$ as the optimal solution for the ERM-rank objective function. By Lemma 8, the test accuracy for the ERM-rank objective function is

$$\begin{aligned}
\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ood}}}[y f_{\mathbf{w}^*_{rank}}(\mathbf{x}) \geq 0] =& \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ood}}}\left[\sum_{j\in R}\mathbf{w}^*_{rank,j}\mathbf{z}_j + \sum_{j\in U}\mathbf{z}_j\mathbf{w}^*_{rank,j} \geq 0\right] && (47)\\
=& \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ood}}}\left[\sum_{j\in R}\mathbf{w}^*_{rank,j}\mathbf{z}_j \geq 0\right] && (48)\\
=& 1. && (49)
\end{aligned}$$

We denote $\mathbf{w}^*_{\ell_2}$ as the optimal solution for the ERM-rank objective function. We have α, β defined in Lemma 7. By Lemma 7, the test accuracy for the ERM-ℓ_2 objective function is

$$\begin{aligned}
\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ood}}}[y f_{\mathbf{w}^*_{\ell_2}}(\mathbf{x}) \geq 0] =& \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ood}}}\left[\alpha\sum_{j\in R}\mathbf{z}_j + \beta\sum_{j\in U}\mathbf{z}_j \geq 0\right] && (50)\\
\leq& \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ood}}}\left[\frac{3}{4\gamma}\sum_{j\in R}\mathbf{z}_j + \sum_{j\in U}\mathbf{z}_j \geq 0\right] && (51)\\
=& \mathbb{P}\left[\frac{3}{4\gamma}\sum_{j\in R}\left(\mathbf{z}_j - \frac{1}{2}\right) + \sum_{j\in U}(\mathbf{z}_j + \gamma) \geq -\frac{3r}{8\gamma} + (d-r)\gamma\right] && (52)
\end{aligned}$$

By Hoeffding's inequality and $d > \frac{r}{\gamma^2} + r > 5r$, we have

$$\begin{aligned}
&\mathbb{P}\left[\frac{3}{4\gamma}\sum_{j\in R}\left(\mathbf{z}_j - \frac{1}{2}\right) + \sum_{j\in U}(\mathbf{z}_j + \gamma) \geq -\frac{3r}{8\gamma} + (d-r)\gamma\right] && (53)\\
\leq& \exp\left(-\frac{2\left(-\frac{3r}{8\gamma} + (d-r)\gamma\right)^2}{4d}\right) && (54)\\
=& \exp\left(-\frac{\frac{9r^2}{32\gamma^2} + 2(d-r)^2\gamma^2 - \frac{3r}{2}(d-r)}{4d}\right) && (55)\\
\leq& \exp\left(-\frac{2(d-r)^2\gamma^2 - \frac{3r}{2}(d-r)}{5(d-r)}\right) && (56)\\
=& \exp\left(-\frac{4(d-r)\gamma^2 - 3r}{10}\right) && (57)\\
\leq& \exp\left(-\frac{r}{10}\right). && (58)
\end{aligned}$$

□