Group Robustness via Adaptive Class-Specific Scaling

Seonguk Seo [1] **Bohyung Han** [1 2]

Abstract

Existing group robustness approaches have apparently improved robust accuracy, but in fact these performance gains mainly come from trade-offs at the expense of average accuracy. To address the limitation, we first propose a simple class-specific scaling strategy to control the trade-off between robust and average accuracies flexibly and efficiently, which is directly applicable to existing debiasing algorithms without additional training. We also develop an instance-wise adaptive scaling technique for overcoming the trade-off and improving the performance even further in terms of both accuracies. Our approach reveals that a naïve ERM baseline matches or even outperforms the recent debiasing methods by only adopting the class-specific scaling. Then, we employ this technique to evaluate the performance of existing algorithms in a comprehensive manner by introducing a novel unified metric that summarizes the trade-off between the two accuracies as a scalar value. By considering the inherent trade-off and providing a performance landscape, our approach delivers meaningful insights in existing robust methods beyond comparing only the robust accuracy. We verify the effectiveness on the datasets in computer vision and natural language processing domains.

1. Introduction

Machine learning models have achieved remarkable performance in various tasks via empirical risk minimization (ERM). However, they often suffer from spurious correlation and dataset bias, failing to learn proper knowledge about minority groups despite their high overall accuracies. Since it is well-known that spurious correlation leads to poor generalization performance in minority groups, group distributionally robust optimization (Sagawa et al., 2020) has been widely used to mitigate algorithmic bias. Numerous approaches (Huang et al., 2016; Sagawa et al., 2020; Seo et al., 2022; Nam et al., 2020; Liu et al., 2021) have presented high robust accuracies such as worst-group or unbiased accuracies in a variety of tasks and datasets, but, although they clearly sacrifice the average accuracy, comprehensive evaluation jointly with average accuracy has not been explored actively. Refer to Figure 1 about the existing trade-offs of algorithms.

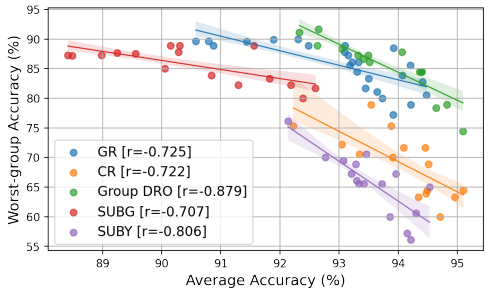

Figure 1: The scatter plots that illustrate trade-off between robust and average accuracies of existing algorithms on the CelebA dataset using ResNet-18. We visualize the results from multiple runs of each algorithm and present the relationship between the two accuracies. The lines denote the linear regression results of individual algorithms and r in the legend indicates the Pearson coefficient correlation.

This paper addresses the limitations of the current research trends and starts with introducing a simple post-processing technique, *robust scaling*, which efficiently performs class-specific scaling on prediction scores and conveniently controls the trade-off between robust and average accuracies at test time. It allows us to identify any desired performance points, *e.g.*, ones in terms of average accuracy, unbiased accuracy, worst-group accuracy, or balanced accuracy, on the accuracy trade-off curve obtained from a single model with marginal computational overhead. The proposed robust-scaling method can be easily plugged into various existing debiasing algorithms to improve the desired target objectives within the trade-off. One interesting observation is that, by adopting the proposed robust scaling, even the ERM baseline accomplishes competitive performance compared to the recent group distributionally robust optimization ap-

[1]ECE, Seoul National University [2]IPAI, Seoul National University. Correspondence to: Bohyung Han <bhhan@snu.ac.kr>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA.

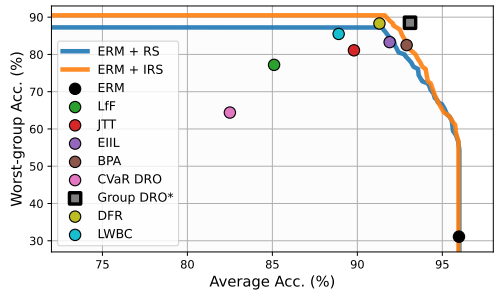

Figure 2: Comparison between the baseline ERM and existing debiasing approaches on CelebA dataset using ResNet-50. Existing works have achieved meaningful improvements in robust accuracy over ERM, but our robust scaling strategies (RS, IRS) enable ERM to catch up with or even outperform them without any training.

proaches (Liu et al., 2021; Nam et al., 2020; Sagawa et al., 2020; Kim et al., 2022; Seo et al., 2022; Creager et al., 2021; Levy et al., 2020; Kirichenko et al., 2022; Zhang et al., 2022) without extra training, as illustrated in Figure 2. We will present the results from other debiasing algorithms in the experiment section.

By taking advantage of the robust scaling technique, we develop a novel comprehensive evaluation metric that consolidates the trade-off of the algorithms for group robustness, leading to a unique perspective of group distributionally robust optimization. To this end, we first argue that comparing the robust accuracy without considering the average accuracy is incomplete and a unified evaluation of debiasing algorithms is required. For a comprehensive performance evaluation, we introduce a convenient measurement referred to as *robust coverage*, which considers the trade-off between average and robust accuracies from the Pareto optimal perspective and summarizes the performance of each algorithm with a scalar value. Furthermore, we propose a more advanced robust scaling algorithm applicable to each example adaptively based on its cluster membership at test time to maximize performance. Our instance-wise adaptive scaling strategy is effective to overcome the trade-off itself and achieve further performance gains for both accuracies.

Contribution We present a simple but effective approach for group robustness by analyzing trade-off between robust and average accuracies. Our framework captures the full landscape of robust-average accuracy trade-offs, facilitates understanding the behavior of existing debiasing techniques, and provides a way for optimizing the arbitrary objectives along the trade-off without extra training. Our main contributions are summarized as follows.

- We propose a training-free class-specific scaling strategy to capture and control the trade-off between robust and average accuracy with marginal computational cost. This approach allows us to optimize a debiasing algorithm for arbitrary objectives within the trade-off on top of any existing models.
- We introduce a novel comprehensive performance evaluation metric via the robust scaling that summarizes the trade-off as a scalar value from the Pareto optimal perspective.
- We develop an instance-wise robust scaling algorithm by extending the original class-specific scaling with joint consideration of feature clusters. This technique is effective to overcome the trade-off and improve both robust and average accuracy further.

2. Framework

2.1. Problem Setup

Consider a triplet (x, y, a) with an input feature $x \in \mathcal{X}$, a target label $y \in \mathcal{Y}$, and an attribute $a \in \mathcal{A}$. We construct a group based on a pair of a target label and an attribute, $g := (y, a) \in \mathcal{Y} \times \mathcal{A} =: \mathcal{G}$. We are given n training examples without attribute annotations, *e.g.*, $\{(x_1, y_1), ..., (x_n, y_n)\}$, while m validation examples have group annotations for model selection, *e.g.*, $\{(x_1, y_1, a_1), ..., (x_m, y_m, a_m)\}$.

Our goal is to learn a model $f_\theta(\cdot) : \mathcal{X} \to \mathcal{Y}$ that is robust to group distribution shifts. To measure the group robustness, we typically refer to the robust accuracy such as unbiased accuracy (UA) and worst-group accuracy (WA). The definitions of UA and WA require the group-wise accuracy (GA), which is formally given by

$$\text{GA}_{(r)} := \frac{\sum_i \mathbb{1}(f_\theta(\mathbf{x}_i) = y_i, g_i = r)}{\sum_i \mathbb{1}(g_i = r)}, \tag{1}$$

where $\mathbb{1}(\cdot)$ denotes an indicator function and $\text{GA}_{(r)}$ is the accuracy of the r^{th} group samples. Then, the robust accuracies are defined by

$$\text{UA} := \frac{1}{|\mathcal{G}|} \sum_{r \in \mathcal{G}} \text{GA}_{(r)} \quad \text{and} \quad \text{WA} := \min_{r \in \mathcal{G}} \text{GA}_{(r)}. \tag{2}$$

2.2. Class-Specific Robust Scaling

As shown in Figure 1, there exists a clear trade-off between robust and average accuracies for each algorithm. To understand its exact behavior, we design a simple non-uniform class-specific scaling of the scores corresponding to individual classes. This strategy may change the final decision of the classifier; if we upweight the prediction scores of minority classes, the samples will have more chances to be classified into minority classes, thus worst-group accuracy increases at the expense of average accuracy, resulting in a desirable trade-off for group robustness. Formally, the

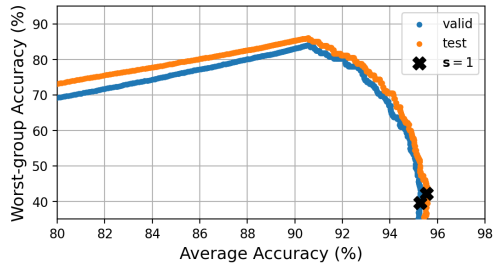

Figure 3: The relation between the robust and average accuracies obtained by varying the class-specific scaling factor $\mathbf{s}$ with ERM model on CelebA. The black marker denotes the original point, where the uniform scaling is applied.

prediction with the class-specific scaling is given by

$$\arg\max_{c} \ (\mathbf{s} \odot \hat{\mathbf{y}})_c, \tag{3}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^C$ is a prediction score vector over C classes, $\mathbf{s} \in \mathbb{R}^C$ is a C-dimensional scaling coefficient vector, and $\odot$ denotes the element-wise product operator.

Based on the ERM model, we obtain a set of the average and robust accuracy pairs using a wide range of the class-specific scaling factors and illustrate their relations in Figure 3. The black markers indicate the point with a uniform scaling, *i.e.*, $\mathbf{s} = (1, \dots, 1) \in \mathbb{R}^C$. The graphs show that a simple class-specific scaling effectively captures the landscape of the trade-off of the two accuracies. This validates that we can identify the desired Pareto optimal points between robust and average accuracies in the test set by following a simple strategy: 1) finding the optimal class-specific scaling factors that maximize the target objective in the validation set, and 2) apply the scaling factors to the test set. We refer to this scaling strategy for robust prediction as *robust scaling*.

To identify the optimal scaling factor $\mathbf{s}$, we adopt a greedy search[1] and the entire process takes less than a second. Note that the robust scaling is a post-processing method, so it can be easily applied to any kinds of existing robust optimization methods without extra training. Moreover, our method can find any desired performance points on the trade-off envelop using a single model. For example, there may be scenarios in which multiple objectives are required to solve a problem, but our robust scaling is flexible enough to handle the situation as we only need to apply a robust scaling optimized for each target metric to a single model. Meanwhile, other robust optimization methods have limited flexibility and require training of separate models for each target objective.

[1]We search for the scaling factor of each class in a greedy manner. Specifically, we first find the best scaling factor for a class and then determine the optimal factors of the remaining classes sequentially conditioned on the previously estimated ones.

2.3. Instance-wise Robust Scaling

The optimal scaling factor can be applied adaptively to each test example and the instance-specific scaling has the potential to overcome the trade-off and improve accuracy even further. Previous approaches (Seo et al., 2022; Sohoni et al., 2020) have shown the capability to identify hidden spurious attributes via clustering on the feature space for debiased representation learning. Likewise, we take advantage of feature clustering for adaptive robust scaling; we obtain the optimal class-specific scaling factors based on the cluster membership for each sample. The overall algorithm of instance-wise robust scaling (IRS) is described as follows.

1. Perform clustering with validation data on the feature space and store the cluster centroids.
2. Find the optimal scaling factor for each cluster.
3. Apply the estimated scaling factor to the test example based on its cluster membership.

In step 1, we use a simple K-means clustering algorithm, where the number of clusters K is set to the value that gives the highest robust coverage in the validation set. Empirically, numbers larger than 10, *i.e.*, $K > 10$, yield stable and superior results, compared to the original class-specific scaling.

2.4. Robust Coverage

Although the robust scaling identifies the desired performance point on the trade-off curve, from the perspective of performance evaluation, it still reflects only a single point on the trade-off curve while ignoring all other possible Pareto optima. For a more comprehensive evaluation of an algorithm, we propose a convenient evaluation metric that yields a scalar summary of the robust-average accuracy trade-off. Formally, we define the *robust coverage* as

$$\text{Coverage} := \int_{c=0}^{1} \max_{\mathbf{s}} \left\{ \text{WA}^{\mathbf{s}} | \text{AA}^{\mathbf{s}} \geq c \right\} dc \tag{4}$$

where $\text{WA}^{\mathbf{s}}$ and $\text{AA}^{\mathbf{s}}$ denote the worst-group and average accuracies given by the scaled prediction using (3) with the scaling factor $\mathbf{s}$. The robust coverage measures the area under the Pareto frontier of the robust-average accuracy trade-off curve, where the maximum operation in (4) finds the Pareto optimum for each threshold. We can use either WA or UA as the target objective of robust coverage in (3).

3. Experiments

3.1. Results

CelebA Table 1 presents the experimental results of our robust scaling methods (RS, IRS) on top of the existing

Table 1: Experimental results of the robust scaling (RS) and instance-wise robust scaling (IRS) on the CelebA dataset using ResNet-18 with the average of three runs (standard deviations in parenthesis), where RS and IRS are applied to maximize each target metric independently. RS can maximize all target metrics consistently and IRS further boosts the performance.

Method	Group Supervision	Robust Coverage		Accuracy (%)					
		Worst-group	Unbiased	Worst-group	(Gain)	Unbiased	(Gain)	Average	(Gain)
ERM		-	-	34.5 (6.1)	-	77.7 (1.8)	-	95.5 (0.4)	-
ERM + RS		83.0 (0.8)	88.1 (0.6)	82.8 (3.3)	+47.7 (7.8)	91.2 (0.5)	+13.3 (2.0)	**95.8 (0.2)**	+0.4 (0.2)
ERM + IRS		**83.4 (0.1)**	**88.4 (0.4)**	**87.2 (2.0)**	+52.7 (3.3)	**91.7 (0.2)**	+13.8 (1.6)	**95.8 (0.1)**	+0.4 (0.3)
JTT (Liu et al., 2021)		-	-	75.1 (3.6)	-	85.9 (1.4)	-	89.8 (0.8)	-
JTT + RS		77.3 (0.7)	81.9 (0.7)	82.9 (2.3)	+7.8 (3.0)	87.6 (0.5)	+1.7 (0.4)	90.3 (1.3)	+0.6 (0.1)
JTT + IRS		**78.9 (2.1)**	**82.1 (1.5)**	**84.9 (4.5)**	+9.8 (3.7)	**88.5 (0.8)**	+2.5 (0.8)	**91.0 (1.8)**	+1.2 (0.5)
GR		-	-	88.6 (1.9)	-	92.0 (0.4)	-	92.9 (0.8)	-
GR + RS	✓	86.9 (0.4)	88.4 (0.2)	90.0 (1.6)	+1.4 (1.1)	92.4 (0.5)	+0.5 (0.4)	93.8 (0.4)	+0.8 (0.5)
GR + IRS		**87.0 (0.2)**	**88.6 (0.2)**	**90.0 (2.3)**	+1.4 (1.8)	**92.6 (0.6)**	+0.6 (0.4)	**94.2 (0.3)**	+1.3 (1.0)
SUBG (Idrissi et al., 2022)		-	- -	87.8 (1.2)	-	90.4 (1.2)	-	91.9 (0.3)	-
SUBG + RS	✓	83.6 (1.6)	87.5 (0.7)	88.3 (0.7)	+0.5 (0.4)	90.9 (0.5)	+0.5 (0.5)	93.9 (0.2)	+1.9 (0.6)
SUBG + IRS		**84.5 (0.8)**	**87.9 (0.1)**	**88.7 (0.6)**	+0.8 (0.7)	**91.0 (0.3)**	+0.6 (0.9)	**94.0 (0.2)**	+2.1 (1.0)
Group DRO (Sagawa et al., 2020)		-	-	88.4 (2.3)	-	92.0 (0.4)	-	93.2 (0.8)	-
Group DRO + RS	✓	87.3 (0.2)	88.3 (0.2)	89.7 (1.2)	+1.4 (1.0)	92.3 (0.1)	+0.4 (0.2)	93.9 (0.3)	+0.7 (0.5)
Group DRO + IRS		**87.5 (0.4)**	**88.4 (0.2)**	**90.0 (2.3)**	+2.6 (1.8)	**92.6 (0.6)**	+0.6 (0.4)	**94.7 (0.3)**	+1.5 (1.1)

Table 2: Experimental results of RS and IRS on the Waterbirds dataset using ResNet-50.

Method	Group Supervision	Robust Coverage		Accuracy (%)					
		Worst-group	Unbiased	Worst-group	(Gain)	Unbiased	(Gain)	Average	(Gain)
ERM		-	-	76.3 (0.8)	-	89.4 (0.6)	-	97.2 (0.2)	-
ERM + RS		76.1 (1.4)	82.6 (1.3)	81.6 (1.9)	+5.3 (1.3)	89.8 (0.5)	+0.4 (0.4)	97.5 (0.1)	+0.4 (0.2)
ERM + IRS		**83.4 (1.1)**	**86.9 (0.4)**	**89.3 (0.5)**	+13.0 (0.9)	**92.7 (0.4)**	+3.3 (0.7)	**97.5 (0.3)**	+0.3 (0.4)
GR		-	-	86.1 (1.3)	-	89.3 (0.9)	-	95.1 (1.3)	-
GR + RS	✓	83.7 (0.3)	86.8 (0.7)	**89.3 (1.3)**	+3.2 (2.0)	92.0 (0.7)	+2.7 (1.3)	95.4 (1.3)	+0.4 (0.2)
GR + IRS		**84.8 (1.7)**	**87.4 (0.4)**	89.1 (0.8)	+3.0 (1.6)	**92.2 (1.0)**	+2.9 (1.6)	**95.6 (0.8)**	+0.6 (0.3)
SUBG		-	-	86.5 (0.9)	-	88.2 (1.2)	-	87.3 (1.1)	-
SUBG + RS	✓	80.6 (2.0)	82.3 (2.0)	87.1 (0.7)	+0.6 (0.5)	**88.5 (1.2)**	+0.3 (0.3)	91.3 (0.4)	+4.0 (0.9)
SUBG + IRS		**82.2 (0.8)**	**84.1 (0.8)**	**87.3 (1.3)**	+0.8 (0.6)	88.2 (1.2)	+0.0 (0.2)	**93.5 (0.4)**	+6.2 (1.5)
Group DRO		-	-	88.0 (1.0)	-	92.5 (0.9)	-	95.8 (1.8)	-
Group DRO + RS	✓	83.4 (1.1)	87.4 (1.4)	89.1 (1.7)	+1.1 (0.8)	92.7 (0.8)	+0.2 (0.1)	96.4 (1.5)	+0.5 (0.5)
Group DRO + IRS		**86.3 (2.3)**	**90.1 (2.6)**	**90.8 (1.3)**	+2.8 (1.5)	**93.9 (0.2)**	+1.4 (0.9)	**97.1 (0.4)**	+1.2 (0.8)

approaches (JTT, Group DRO, GR, SUBG) on the CelebA dataset. In this table, RS and IRS are applied to maximize each target metric (e.g. worst-group acc) independently. We ran the experiments three times and reported the average with the standard deviation of the results from each algorithm. Group supervision indicates that the method requires training examples with group supervision. As shown in the tables, no matter what the backbone method is, our robust scaling strategy consistently improves the performance for all target metrics. Based on the robust coverage and robust accuracy after scaling, JTT is not superior to ERM on the CelebA dataset, though their initial robust accuracies without scaling are much higher than ERM. On the other hand, the methods that leverage group supervision (Group DRO, GR) achieve better robust coverage results than the others, which verifies that group supervision helps to improve the trade-off itself. For the group-supervised methods, our scaling technique provides relatively small performance gains in robust accuracy, as the gaps between robust and average accuracies are already small and the original results are already close to the optimal for robust accuracy.

Waterbirds Table 2 demonstrates that our frameworks achieve outstanding performance with all baselines on the Waterbirds dataset consistently. Among the compared algorithms, GR and SUBG are reweighting and subsampling baselines based on group frequency, respectively. Although the two baseline approaches achieve competitive robust accuracy, the average accuracy of SUBG is far below than GR. This is mainly because SUBG drops a large portion of training samples (95.3%) to make all groups have the same size, resulting in significant performance degradation in average accuracy. Although subsampling generally helps to achieve high initial robust accuracy, it degrades the overall trade-off as well as the average accuracy and consequently hinders the benefits of robust scaling. Note that GR outperforms SUBG in terms of all worst-group, unbiased, and average accuracies after adopting RS or IRS. This consistently demonstrates the effectiveness of our framework and supports our main claim; considering only the robust accuracy is incomplete and comprehensive evaluation is needed to understand the exact behavior.

References

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *ICML*, 2021.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Learning deep representation for imbalanced classification. In *CVPR*, 2016.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *CLeaR*, 2022.

Kim, N., Hwang, S., Ahn, S., Park, J., and Kwak, S. Learning debiased classifier with biased committee. *arXiv preprint arXiv:2206.10843*, 2022.

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

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. In *NeurIPS*, 2020.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *ICML*, 2021.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: Training debiased classifier from biased classifier. In *NeurIPS*, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Seo, S., Lee, J.-Y., and Han, B. Unsupervised learning of debiased representations with pseudo-attributes. In *CVPR*, 2022.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *NeurIPS*, 2020.

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