Robustness to Spurious Correlations via Distributionally Robust Neural Network

Rong Jiang*, Chenfeng Li*, Charles Mayville*

May 2023

1 Introduction

In classical machine leraning, models are typically trained to minimize the average loss of a training dataset, and then evaluated on a test dataset that is often assumed to be from the same distribution as the training one. In practice, such methodology is problematic due to two reasons: 1) the training distribution and the test distribution can be different; 2) methods that optimize over average performance may suffer low performance on rare and atypical instances. For example, it is observed that models can experience significant downgrading performance when distribution shift between training and test data occurs [2]. Besides, models that achieve high average performance can still fail on atypical examples or rely on spurious correlations [4]. These issues raise safety and fairness concerns for building trustworthy machine learning systems in areas such as medical diagnosis and autonomous vehicles.

To tackle these challenges, [1] develops a distributionally robust stochastic optimization (DRO) framework, which seeks to optimize the worst case performance over a family of potential test distributions. While the DRO framework considered in [1] addresses the issues of distribution shift and worst-case performance, it is restricted to convex predictive models with limited capacity, and therefore cannot satisfy the needs of modern machine learing where deep neural networks play a central role.

DRO in the context of overparameterized neural networks is studied in [5] They consider group DRO, where they train models to minimize the worst-case loss over groups in the training data, so as to prevent the model from learning spurious correlations and suffering high loss on certain groups of data. Under the overparameterized regime, they find that models trained through both group DRO and empirical risk minimization (ERM) can achieve nearly zero training loss and high average test accuracy, but the worst-group test accuracies are low. They suggest this behavior is because the average generalization gap is small while the worst group one is large. However, by applying regularization to group

*Alphabetical order.

DRO models, they show the former can outperform regularized and unregularized ERM models in worst-group test accuracies while retaining high average test accuracies. Their findings indicate while overparameterization might allow for good average performance, regularization is still needed for good worst-case performance in neural networks.

In this report, we critically evaluate the method in [5] on a newly generated dataset that their work does not cover. We consider the colored MNIST dataset where we assign a binary label to the image according to the digit and color the image based on its label. We find that regularized DRO models can outperform regularized and unregularized ERM models in worst-group test accuracies while maitaining relatively high average test accuracies, which corroborates the results in [5]. One limitation of [5] is that they assume the test groups are all seen during training. This assumption might not be realistic in practice so we add unseen groups to our test data to further examine the performance of group DRO. We observe that group DRO with regularization has worst-group and average test accuracies very close to the ones in test data with all-seen groups. This shows the ability of regularized group DRO to prevent models from learning spurious correlations.

The report is organized in the following way: section 2 formally introduces the settings of ERM and group DRO. Section 3 contains the experiment results of our colored MNIST data. Section 4 conludes and discusses future directions.

2 Setup

Let $x \in \mathcal{X}$ be the input feature and $y \in \mathcal{Y}$ be the label. Denote our hypothesis family as Θ, and $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_+$ is the loss function. We assume our training data is sampled from some distribution P. In standard machine learning, we use empirical risk minimization (ERM) to find

$$\hat{\theta}_{\mathrm{ERM}} = \arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim \hat{P}}[l(\theta; (x,y))], \tag{1}$$

where $\hat{P}$ is the empirical distribution of P.

Rather than minimizing the average loss, DRO minimizes the worst-case expected loss over a set of distributions $\mathcal{Q}$:

$$\min_{\theta \in \Theta} \{R(\theta) = \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim Q}[l(\theta; (x,y))]\}. \tag{2}$$

Ideally, we want $\mathcal{Q}$ to contain the test distribution so that our model can perform well on it. One way to choose $\mathcal{Q}$ is to consider a divergence ball around the training distribution [1].

In this report, we consider the group DRO formulation as in [5]. Prior knowledge of spurious correlations is used to define groups over the training data and $\mathcal{Q}$ is then defined using these groups. Concretely, let $\mathcal{Q} = \{\sum_{g=1}^{m} q_g P_g : q \in \Delta_m\}$, where Δ_m is the $(m-1)$-dimensional probability simplex and P_g is the distribution of group g. In other words, $\mathcal{Q}$ is a set of mixture distributions and each component of the mixture distribution represents a group. Since a linear program attains its optimum at the vertex, the worst-case risk in (2) is equivalent to the worst group risk

$$R(\theta) = \sup_{g \in [m]} \mathbb{E}_{(x,y)\sim P_g}[l(\theta; (x, y))]. \tag{3}$$

Thus, we can learn a group DRO model

$$\hat{\theta}_{\text{DRO}} = \arg\min_{\theta \in \Theta} \{\hat{R}(\theta) = \sup_{g \in [m]} \mathbb{E}_{(x,y)\sim \hat{P}_g}[l(\theta; (x, y))]\} \tag{4}$$

where $\hat{P}_g$ is the empirical distribution of P_g. Here, we assume the training data consists of (x, y, g) triplets but g is not assumed to be observed at test time. Group DRO prevents the model from learning spurious correlations by optimizing the performance of the worst group where the misleading heuristics might not hold. However, (4) does not imply good test worst-group performance due to the worst-group generalization gap $\delta = R(\theta) - \hat{R}(\theta)$. We will examine the effect of regularization on δ in the following section.

3 Experiment

3.1 Data

We start by describing how we construct our dataset from the MNIST database. For each image, we label it as one of $\mathcal{Y} = \{0, 1\}$ based on the digit ($y = 0$ for digits 0-4 and $y = 1$ for digits 5-9) and color it as one of $\mathcal{A} = \{\text{red}, \text{green}\}$, with 0 more frequently appearing on red images and 1 more frequently appearing on green images. We thereby form $m = |\mathcal{A}| \times |\mathcal{Y}| = 4$ groups, one for each $(a, y) \in \mathcal{A} \times \mathcal{Y}$. It is expected that models which learn the correlation between label and color in the training data will perform badly on groups where the relation does not hold and consequently on the worst-group loss $R(\theta)$.

Specifically, we color a image as red if its label is 0 and green if its label is 1. We then flip the color of each image with a probability of 0.2. We uniformly split the data into training and test sets so that the portion of each group within training data and test data are roughly the same, with a training size of 20000 and test size of 5000. Besides, we incoporate corruption into the data by flipping the label of each image in both training and test data with a probability of 0.25. So the optimal accuracy should be $1 - 0.25 = 0.75$ and an accuracy higher than that in the training process would indicate that the model relies on the spurious correlation or memorizes the data.

3.2 ERM v.s. DRO without regularization

We start by comparing the performance of ERM models and DRO models without regularization added.

ERM. The ERM models achieve 80.0% average training accuracy and 80.3% average test accuracy [1]. However, the worst-group training accuracy is only 0.2% and the worst-group test accuracy is 0.3%[2]. Clearly, ERM models merely learn the spurious correlation between label and color and therefore fail on the groups where this relation do not hold.

DRO. The DRO models obtain 91.7% average training accuracy and 72.8% average test accuracy. The worst-group training accuracy is 89.8% and the worst-group test accuracy is 44.4% (Table 1, Figure 1). We find that though DRO without regularization can outperform ERM in terms of worst-group accuracies, the worst-group generalization gap between training and test data is still large.

	Average Accuracy		Worst-Group Accuracy	
	ERM	DRO	ERM	DRO
Train	80.0%	91.7%	0.2%	89.8%
Test	80.3%	72.8%	0.3%	44.4%

Table 1: Comparison of ERM and DRO models without regularization.

	Average Accuracy		Worst-Group Accuracy	
	ERM	DRO	ERM	DRO
Train	80.2%	73.9%	0.0%	70.7%
Test	80.3%	71.3%	0.0%	67.5%

Table 2: Comparison of ERM and DRO models with regularization.

3.3 ERM v.s. DRO with regularization

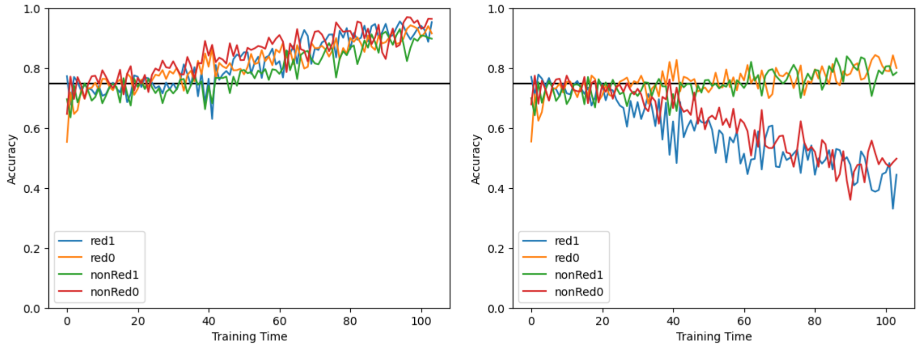
Next, we regularize both models by using l_2 penalty and compare their performance.

ERM. The ERM models get 80.2% average training accuracy and 80.3% average test accuracy. Still, the worst-group training accuracy is only 0.0% and the worst-group test accuracy is 0.0%. We find that adding regularization does not improve the performance of ERM.

DRO. Under regularization, DRO models reach 73.9% average training accuracy and 71.3% average test accuracy. Note that we flip 25% of the labels

[1] The test accuracy is slightly higher here because the portion of worst groups in test data is slightly less than that of the training data.

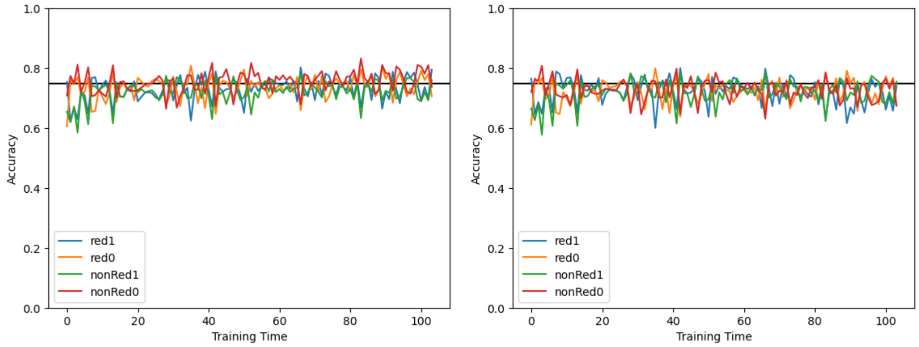
[2] Unlike [5] where they obtain nearly perfect worst-group training accuracy using ResNet50 as the underlying learner which is able to memorize the data, in our case ERM performs badly even in the worst group of the training data because we use a moderately-sized convolutional neural network as our underlying learner due to computational constraint.



(a) DRO training accuracy

(b) DRO test accuracy

Figure 1: Unregularized DRO accuracies by group (the black line at 0.75 represents the theoretical optimal accuracy)



(a) DRO training accuracy

(b) DRO test accuracy

Figure 2: Regularized DRO accuracies by group

in both training and test data so the optimal accuracy should be 75% theoretically, and an accuracy higher than it suggests the model is memorizing the data. The worst-group training accuracy is 70.7% and the worst-group test accuracy is 67.5% (Table 2, Figure 2). We can see that regularization significantly reduces the worst-group generalization gap while retaining the average test accuracy.

3.4 Unseen groups

A limitation of [5] is that their test groups are all seen during training. To further evaluate DRO, we remove this condition and test it on unseen data. We do so by creating a test dataset with the originally green color replaced by random RGB values (Figure 3). Using regularized DRO models trained on data containing only red and green groups, we obtain an average test accuracy of 69.8% and a worst-group test accuracy of 67.3% under the new test set with unseen groups. This result is similar to the result of regularized DRO on test data with all-seen groups.

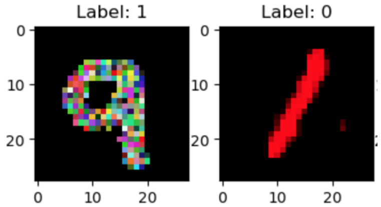


Figure 3: Unseen group where the green color is replaced by random RGB values

4 Discussion

In this report, we examine the performance of group DRO on our colored MNIST dataset. We find that group DRO with regularization can improve the worst-group performance while keeping the average performance, which is consistent with the results in [5]. When evaluated on test data containing groups unseen during training, group DRO with regularization retains its worst-group and average test performance as before. This suggests group DRO with regularization has the potential to learn the underlying true relation instead of relying on spurious correlations. Nevertheless, more complicated real world data is still required to fully test the effectiveness of it. Besides, instead of focusing on the worst-group performance, methods which reweight different groups based on their training errors are also interesting solutions to the group distributional robustness problem [3]. These directions are beyond the scope of this report and we leave them as future work.

References

[1] John Duchi and Hongseok Namkoong. *Learning Models with Uniform Performance via Distributionally Robust Optimization.* 2020. arXiv: 1810.08750 [stat.ML].

[2] David J. Hand. “Classifier Technology and the Illusion of Progress”. In: *Statistical Science* 21.1 (2006), pp. 1–14. DOI: 10.1214/088342306000000060. URL: https://doi.org/10.1214/088342306000000060.

[3] Yachuan Liu et al. *Ranking Reweighting Improves Group Distributional Robustness.* 2023. arXiv: 2305.05759 [cs.LG].

[4] Nicolai Meinshausen and Peter Bühlmann. “Maximin effects in inhomogeneous large-scale data”. In: *The Annals of Statistics* 43.4 (Aug. 2015). DOI: 10.1214/15-aos1325. URL: https://doi.org/10.1214%2F15-aos1325.

[5] Shiori Sagawa* et al. “Distributionally Robust Neural Networks”. In: *International Conference on Learning Representations.* 2020. URL: `https://openreview.net/forum?id=ryxGuJrFvS`.