Towards Group Robustness in the Presence of Partial Group Labels

Anonymous Authors[1]

Abstract

Learning invariant representations is a fundamental requirement for training machine learning models that are influenced by spurious correlations. These spurious correlations, present in the training datasets, wrongly direct the neural network predictions resulting in reduced performance on certain groups, especially the minority groups. Robust training against such correlations requires the knowledge of group membership on every training sample. This need is impractical in situations where the data labeling efforts, for minority/rare groups, are significantly laborious or where the individuals comprising the dataset choose to conceal sensitive information pertaining to the groups. On the other hand, the presence of data collection efforts often results in datasets that contain partially labeled group information. Recent works, addressing the problem, have tackled fully unsupervised scenarios where no labels for groups are available. We aim to fill a missing gap in the literature that addresses a more realistic setting by leveraging partially available group information during training. First, we construct a constraint set and derive a high probability bound for the group assignment to belong to the set. Second, we propose an algorithm that optimizes for a worst-off group assignment from the constraint set. Through experiments on image and tabular datasets, we show improvements in the minority group's performance while preserving overall accuracy across groups.

1. Introduction

Neural networks being overly biased to certain groups of the data is an increasing concern within the machine learning community (Agarwal et al., 2018). A primary cause for bias against specific groups is the presence of extraneous attributes in the datasets that wrongly direct the model responses (Xie et al., 2017). An inevitable consequence of such correlations to extraneous attributes is disparities in performance across different groups. Specifically, if certain groups form a minority, a model can simply *cheat* by having a high overall aggregate accuracy but poor minority group accuracy (Oakden-Rayner et al., 2020).

Existing works (Arjovsky et al., 2019; Sagawa et al., 2019) operate in the regime where the number of groups are known apriori. Further, most works assume a complete knowledge of the group membership of individual samples. While these methods have been proven effective, it is not realistic to assume complete access to the group labels. For example, in a lung cancer detection problem, the label class could contain many unrecognized subgroups such as solid/subsolid tumors and central/peripheral neoplasms (Oakden-Rayner et al., 2020). These unrecognized subgroups are difficult to label and form a minority in the dataset, thereby resulting in an unbalanced performance across the different subgroups (Sohoni et al., 2020). In this work, we consider a setting where a significant portion of the training data is devoid of group labels. We choose to fill a missing gap in the literature where several works bifurcate into methods that either are fully supervised or fully unsupervised.

We answer the question using a framework of distributionally robust optimization (DRO) (Shapiro et al., 2021; Namkoong & Duchi, 2016). DRO optimizes for the worst-case training loss over predefined set of groups and is closely connected to the Rawlsian criterion (Rawls, 2001). The application of DRO to the partial group label setting poses some challenges: (1) the lack of group label makes it infeasible to compute the worst-off group loss; (2) optimizing only for the high-loss samples, by considering them as a worst-off group, discards considerable portion of the training data thereby impacting the overall accuracy; and (3) inferring missing group labels with pseudo-label methods is a cause for ethical concerns.

The third challenge above suggests a straightforward way of handling partial labels wherein we directly estimate the group label for unlabelled samples. However, this approach could be harmful in the context of fairness problems because the estimated labels are susceptible to misuse by a wrongdoer. For example, when the groups are indicative

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

of sensitive information such as age or gender, an incorrect estimation would wrongly designate the demographics of an individual. Moreover, when it's desirable to conceal such sensitive information, a direct estimation of groups would be violation of privacy. Thus we **cautiously avoid** building or utilizing pseudo-label based methods in this paper.

In light of above challenges, we make the following **contributions**. We propose a method that defines a constraint set of group assignments and optimize over worst-off assignments within the set. We show that the constraint set encompasses the ground-truth group labels of the unlabeled data with high probability. Since worst-off assignments do not directly relate to ground-truth, our approach is safe and does not violate privacy. As we shall see, our method assigns high loss samples to groups with low marginal probability and does not discard any samples. We show experiments on three imaging datasets and one tabular dataset.

2. Methodology

We revisit the GroupDRO in Section 2.1 and detail our method, Worst-off DRO, in Section 2.2 and 2.3. In Section 2.4, we describe a practical algorithm for optimization.

2.1. Preliminary: Group DRO

Let $x \in \mathcal{X} \subset \mathbb{R}^d$ be data descriptors, $y \in \mathcal{Y} \subset \{0, 1\}$ be target labels, and $g \in \mathcal{G} \subset \{1, ..., M\}$ be group labels.We assume training a neural network parameterized by the weights w that corresponds to a per-sample loss $l(x, y; w)$. Given data triplets $\{(x_i, y_i, g_i)\}_{i=1}^N$, we seek to optimize w for the Rawlsian criterion (Rawls, 2001; Hashimoto et al., 2018), which minimizes the loss of the worst-off group, as follows:

$$\min_w \max_{g \in \mathcal{G}} \mathbb{E}\big[l(x, y; w) | g\big]. \tag{1}$$

Sagawa et al. (2019) proposed a practical algorithm to solve (1), called *Group DRO*. This method optimizes a weighted expected loss across all groups. These weights over the groups, denoted by q, are drawn from a simplex Δ^M. The objective function $\mathcal{L}_{\text{GDRO}}$ is as described below,

$$\min_w \max_{q \in \Delta^M} \sum_{j=1}^{M} \overbrace{q_j}^{\text{Group Weights}} \underbrace{\Big[\frac{\sum_{i=1}^N \overbrace{\mathbb{1}\{g_i = j\}}^{\text{Indicator function}} \; l(x_i, y_i, w)}{\sum_{i=1}^N \mathbb{1}\{g_i = j\}}\Big]}_{\text{Per-group average loss}} \tag{2}$$

2.2. Our Approach: Worst-off DRO

In this work, we are interested in training a robust neural network when group labels are only partially available. That is, our training dataset consists of fully-labeled $\{(x_i, y_i, g_i^\star)\}_{i=1}^K$ and task-labeled samples $\{(x_i, y_i, -)\}_{i=K+1}^N$, where $-$ indicates the missing groups.

As noted in (2), the Group DRO requires group labels of entire dataset. When some of them are missing, we propose to optimize for the following objective for $\mathcal{L}_{\text{WDRO}}(\mathcal{C})$:

$$\min_w \max_{q \in \Delta^M} \max_{\{\hat{g}\} \in \mathcal{C}} \sum_{j=1}^{M} q_j \Big[\frac{\sum_{i=1}^N \mathbb{1}\{\hat{g}_i = j\} l(x_i, y_i, w)}{\sum_{i=1}^N \mathbb{1}\{\hat{g}_i = j\}}\Big] \tag{3}$$

where $\mathcal{C}$ is a set of group assignments $\{\hat{g}_i\}_{i=1}^N$ satisfying $\hat{g}_i = g_i^\star, \forall i \leq K$. We call the objective in (3) **Worst-off DRO** method; as it optimizes over the worst-off assignments in a certain constraint set $\mathcal{C}$ (more details on $\mathcal{C}$ soon).

Note that the Worst-off DRO objective forms an upper bound to the Group DRO objective evaluated at the ground-truth group labels if $\{g_i^\star\}_{i=1}^N \in \mathcal{C}$. Under identical parameters w and q, this is rather a straightforward consequence from the fact that the ground-truths $\{g_i^\star\}_{i=1}^N$ falls within the constraint set $\mathcal{C}$. However, the following lemma generalizes the upper bound relationship between Worst-off DRO and Group DRO objectives for all w and q parameters.

Lemma 2.1. *Denote $\mathcal{L}_{\text{GDRO}}$ at a specific w and q as $\mathcal{L}_{\text{GDRO}(w,q)}$. Similarly $\mathcal{L}_{\text{WDRO}}(\mathcal{C})$ at a fixed w and q is denoted by $\mathcal{L}_{\text{WDRO}(w,q)}(\mathcal{C})$. When the ground-truth group assignment $\{g_i^\star\}_{i=1}^N \in \mathcal{C}$, we have*

$$\min_w \max_{q \in \Delta^M} \mathcal{L}_{\text{GDRO}(w,q)} \leq \min_w \max_{q \in \Delta^M} \mathcal{L}_{\text{WDRO}(w,q)}(\mathcal{C}) \tag{4}$$

The proof is in Appendix A.3. For safety-critical applications, such as learning a fair classifier, it is important that the optimal objective (i.e., Group DRO with a ground-truth group assignment) is bounded by the objective used in optimization as in Lemma 2.1. This is simply because optimizing the proposed objective **guarantees** that the corresponding lower bound Group DRO is also optimized.

2.3. Reducing Constraint Set with Marginal Distribution Constraint

Clearly the constraint set plays an important role in connecting Worst-off DRO to Group DRO. We see that Worst-off DRO can be made closer to Group DRO by further restricting $\mathcal{C}$ as long as $\mathcal{C}$ contains the ground-truths. To achieve this goal, we utilize marginal distribution constraints. These constraints may be given as a side information or could be estimated from the small set of partial group labels. Let $\mathcal{C}_{\mathbf{p},\epsilon}$ be subset of $\mathcal{C}$ whose elements $\{g_i\}_{i=1}^N$ satisfy,

$$g_i = g_i^\star, \forall i \leq K, \tag{5}$$

$$|\frac{1}{N}\sum_{i=1}^N \mathbb{1}\{g_i = j\} - \mathbf{p}_j| \leq \epsilon, \forall j \leq M, \tag{6}$$

where (5) implies that the true group labels are assigned whenever available, and (6) implies that the data marginal distribution should be close to the marginal distribution $\mathbf{p}$. Then, for any marginal distribution $\mathbf{p}$ and $\epsilon > 0$, it is easy to show $\mathcal{L}_{\text{WDRO}}(\mathcal{C}_{\mathbf{p},\epsilon}) \leq \mathcal{L}_{\text{WDRO}}(\mathcal{C})$ as $\mathcal{C}_{\mathbf{p},\epsilon} \subset \mathcal{C}$. Moreover, we will see in Lemma 2.2 that, with high probability, the

Algorithm 1 Worst-off DRO Algorithm

Input: Fully-labelled dataset $\{(x_i, y_i, g_i^\star)\}_{i=1}^K$, task-labeled dataset $\{(x_i, y_i, -)\}_{i=K+1}^N$
Initialization: learning rates η_w and η_q, Marginal distribution $\bar{\mathbf{p}}$, Permissible variance ϵ
Parameters: Group Weights q_j, Worst-off DRO group assignments $\hat{g}$, Neural network parameter w
for $t = 0, 1, 2, ..., T$ **do**
  $\{\hat{g}^t\} \leftarrow \max_{\{\hat{g}\}\in\mathcal{C}_{\bar{\mathbf{p}},\epsilon}} \sum_{j=1}^M q_j^t \frac{\sum_i \hat{g}_{ij} l(x_i,y_i;w^t)}{\sum_i \hat{g}_{ij}}$ where, $\mathcal{C}_{\bar{\mathbf{p}},\epsilon}$ as defined in (10).
  Gradient descent on w: $w^{t+1} \leftarrow w^t - \eta_w \nabla_w \sum_{j=1}^M q_j^t \frac{\sum_i \hat{g}_{ij}^t l(x_i,y_i;w)}{\sum_i \hat{g}_{ij}^t}$
  Exponential ascent on q: $q^{t+1} \leftarrow q^t \exp\left(\eta_q \nabla_q \sum_{j=1}^M q_j \frac{\sum_i \hat{g}_{ij}^t l(x_i,y_i;w^{t+1})}{\sum_i \hat{g}_{ij}^t}\right)$
end for
Output: Trained neural network parameters w^{T+1}

constraint set $\mathcal{C}_{\mathbf{p}^\star,\epsilon}$ with the true marginal distribution $\mathbf{p}^\star$ contains the true group assignment $\{g^\star\}$.

Lemma 2.2. *The constraint set $\mathcal{C}_{\mathbf{p}^\star,\epsilon}$ contains the true group labels $\{g_i^\star\}_{i=1}^N$ with high probability:*

$$P(\{g_i^\star\}_{i=1}^N \in \mathcal{C}_{\mathbf{p}^\star,\epsilon}) \geq 1 - 2Me^{-2N\epsilon^2} \tag{7}$$

The proof is in Appendix A.3. As in (7), the probability of the constraint set containing the true group labels gets closer to 1 by allowing a larger variance (ϵ) from the true marginals. For fixed $\epsilon > 0$, the probability gets closer to 1 as we increase the number of *unlabeled* data (N). Finally, this implies that $\mathcal{L}_{\text{WDRO}}(\mathcal{C}_{\mathbf{p}^\star,\epsilon})$ is an upper bound to $\mathcal{L}_{\text{GDRO}}$:

$$\boxed{\mathcal{L}_{\text{GDRO}} \underset{\text{w.h.p}}{\leq} \mathcal{L}_{\text{WDRO}}(\mathcal{C}_{\mathbf{p}^\star,\epsilon}) \leq \mathcal{L}_{\text{WDRO}}(\mathcal{C})}$$

In practice, however, the true marginal distribution $\mathbf{p}^\star$ may not be available. For our setting where group labels are partially available, with an assumption that group labels are missing completely at random (MCAR) (Rubin, 1976), the true marginal distribution could be estimated from the labelled subset. This again allows us to formulate a constraint set that contains the ground-truth group assignment with high probability. Let $\bar{\mathbf{p}}$ be the estimate of the marginal distribution from $\{(x_i, y_i, g_i^\star)\}_{i=1}^K$.

Lemma 2.3. *The constraint set $\mathcal{C}_{\bar{\mathbf{p}},\delta+\epsilon}$ contains the true group labels $\{g_i^\star\}_{i=1}^N$ with high probability:*

$$P(\{g_i^\star\}_{i=1}^N \in \mathcal{C}_{\bar{\mathbf{p}},\delta+\epsilon}) \geq 1 - 2Me^{-2N\epsilon^2} - 2Me^{-2K\delta^2} \tag{8}$$

Proof in Appendix A.3. Here, δ accounts for estimation error from the true marginals $\mathbf{p}^\star$.

2.4. A Practical Optimization Algorithm

We are interested in solving the optimization problem $\mathcal{L}_{\text{WDRO}}(\mathcal{C}_{\mathbf{p},\epsilon})$. Unfortunately, the inner maximization problem with respect to the group assignments $\{\hat{g}\}$ in (3) is challenging as variables are discrete and the objective cannot be decomposed due to the marginal distribution constraint. Thus, we propose to use a soft group assignments. Specifically, for each sample, we retain a soft group assignment $\hat{g}_i \in \Delta^M$, and optimize the Worst-off DRO objective over the constraint set $\mathcal{C}_{\bar{p},\epsilon}$ as defined below:

$$\min_w \max_{q\in\Delta^M} \max_{\{\hat{g}\}\in\mathcal{C}_{\bar{\mathbf{p}},\epsilon}} \sum_{j=1}^M q_j \left[\frac{\sum_{i=1}^N \hat{g}_{ij} l(x_i, y_i, w)}{\sum_{i=1}^N \hat{g}_{ij}}\right] \tag{9}$$

$$\mathcal{C}_{\bar{\mathbf{p}},\epsilon} = \left\{ \{\hat{g}_i\}_{i=1}^N \middle| \begin{array}{l} \hat{g}_i \in \Delta^M, \forall i \leq N, \\ \hat{g}_{i(g_i^\star)} = 1, \forall i \leq K, \\ |\frac{1}{N}\sum_{i=1}^N \hat{g}_{ij} - \bar{\mathbf{p}}_j| \leq \epsilon, \forall j \leq M \end{array} \right\} \tag{10}$$

The first condition ensures that the assignments form a probability simplex, second one assures consistency with labeled data, and the third one validates the data marginal distribution follows the provided distribution. The third constraint also provides for a mitigation strategy when $\bar{\mathbf{p}}$ is misspecified (likely when data is not MCAR). We alternate optimization over w, q and $\{\hat{g}\}$ as shown in Algorithm 1.

Next, we will see how the worst-off assignments computed by the algorithm look to be. For simplicity, consider the case $\epsilon = 0$ and $K = 0$, (i.e., no group-labelled samples). Denoting $\frac{q_j}{\sum_i^N \hat{g}_{ij}} = \frac{Nq_j}{\bar{\mathbf{p}}_j}$ in (9) as θ_j and $l(x_i, y_i, w)$ as l_i, we can re-write the maximization over $\{\hat{g}_{ij}\}$ as,

$$\max_{\{\hat{g}_{ij}\}\in\mathcal{C}_{\bar{\mathbf{p}},\epsilon=0}} \sum_{i=1,j=1}^{N,M} \hat{g}_{ij} \times \theta_j \times l_i \tag{11}$$

The constraints ensure $\sum_{i=1}^N \hat{g}_{ij} = N\bar{\mathbf{p}}_j$ and $\sum_{i=j}^M \hat{g}_{ij} = 1, \hat{g}_{ij} \geq 0$ for all $i \leq N$ and $j \leq M$ respectively. The linear program (11) sets the highest mass on $\hat{g}_{ij}$ for i and j that maximize θ_j and sample loss l_i. A large θ_j represents groups with a high group weight q_j and low marginal probability $\bar{\mathbf{p}}_j$, characteristic of a worst-off group. In summary, *high loss samples are assigned to groups with high group weights and low marginal probabilities*.

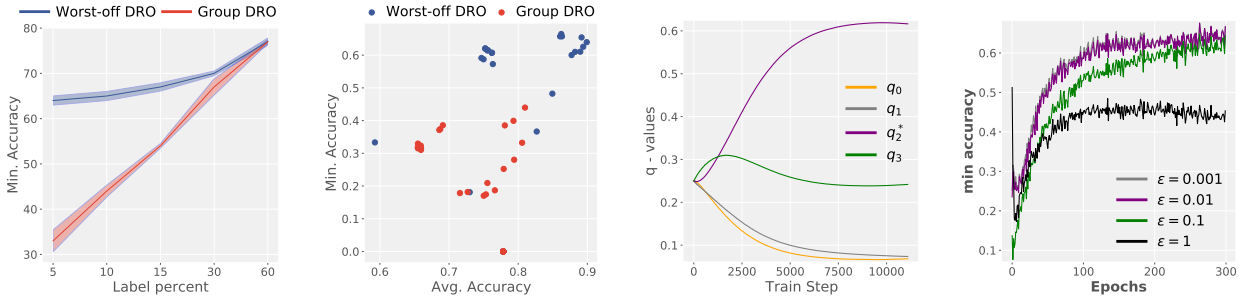
3. Experiments

We test the efficacy of our method on image and tabular datasets[1]. We use Waterbirds (Sagawa et al., 2019), Group CMNIST (Arjovsky et al., 2019), Group Adult (Dua et al.,

[1]CMNIST and Adult datasets differ from their previous instantiations in (Creager et al., 2021). These datasets are used to assess the group-robustness criterion, hence same pre-defined groups are used in both training and testing. More details in Appendix A.11

Table 1: **Quantitative Results.** Here, the labelled samples are about 10% of total samples. For baselines, we consider an ERM, Unsup DRO (Hashimoto et al., 2018), Group DRO (Partial) for partly labelled Group DRO (Sagawa et al., 2019) method, Group DRO (Oracle) for the fully supervised model. Our method Worst-off DRO improves the minority group's accuracy (**min**) while maintaining a similar overall accuracy (**avg**) relative to baselines. The accuracies are computed on the test set and are an average over three random runs.

	Waterbirds		CMNIST		Adult		CelebA	
	min	**avg**	**min**	**avg**	**min**	**avg**	**min**	**avg**
Group DRO (Oracle)	82.9±0.5	92.0±0.0	49.7±0.6	75.3±0.3	82.0±0.9	87.5±0.9	79.6±1.3	94.3±0.3
ERM	59.7±1.0	87.4±0.1	12.7±1.5	79.3±0.5	67.9±1.1	92.0±0.3	44.8±3.0	94.9±0.1
Unsup DRO	64.8±0.9	88.2±0.1	9.6±1.4	79.5±0.6	67.6±1.7	92.1±0.2	38.9±1.8	95.5±0.0
Group DRO (Partial)	44.0±1.3	81.0±1.3	35.5±0.8	75.6±0.2	67.0±0.3	90.4±0.3	39.6±4.5	95.0±0.1
Worst-off DRO	**65.4±1.0**	89.2±0.2	**39.4±1.1**	77.0±0.4	**71.3±0.2**	90.8±0.1	**48.7±2.1**	95.0±0.1



(a) Increasing labels (b) Model selection (c) Group Weights (d) Constraint set size

Figure 1: Ablation experiments on Waterbirds dataset: **(a)** Group DRO (Partial) and Worst-off DRO algorithms improve the minority group accuracies as the number of group labels are increased. Also, the Worst-off DRO method has relatively higher accuracy values than Group DRO. **(b)** Evaluations for different hyper-parameter choices are plotted for Worst-off DRO and Group DRO (Partial) methods. Worst-off DRO models are concentrated in the top-right corner of the plots. This is desirable indicating high accuracies across the two metrics. **(c)** The progression of $q-$values (see Algorithm 1) is plotted for each group. The $q-$values for the minority groups increases gradually while those of the majority groups reduce. A high $q-$value indicates that the corresponding group receives a higher weight relative to other groups (minority group indicated by $*$). **(d)** The marginal constraint is gradually relaxed by increasing the ϵ parameter. Models with $\epsilon \leq 0.01$ have similar performance. Accuracies start to drop when increasing ϵ beyond 0.01.

2017), and CelebA (Liu et al., 2015). We compare our method against ERM, Unsup DRO (Hashimoto et al., 2018) and Group DRO. We consider a variant of Group DRO that only uses samples with group labels at train time. We call the method *Group DRO (Partial)*, to contrast with fully-labelled counterpart, *Group DRO (Oracle)*. Our quantitative results in Table 1 show that the minority group accuracy is higher relative to the baselines for Worst-off DRO method[2]. Moreover, the average accuracy across all the groups is within a 2% window for all the methods.

Ablation Experiments. In Figure 1a, we show that the minority group accuracy increases with more labelled samples. More labels provides two benefits. Firstly, the standard deviation of errors in estimating the marginal probabilities reduces (Wasserman, 2004) ($\approx \sqrt{\#\text{ samples}}$ rate). Secondly, labelled groups reinforce an accurate evaluation of the Rawlsian objective in (2). In Figure 1b, we plot the minority/average group accuracies for different hyper-parameter choices. Due to distributional differences amongst the groups, a mild tradeoff exists between the two accuracy measures. Evidently, the top-right regions in the plot are desirable, at which the Worst-off DRO models are positioned. Next, in Figure 1c, we plot group weights ($q-$values) in the training phase. The plots show a high $q-$value on the minority groups indicating that the loss values on them are upweighted relative to other groups. Lastly, Figure 1d describes experiments where the constraint set size is gradually increased by varying the ϵ parameter of $\mathcal{C}_{\bar{\mathbf{p}},\epsilon}$. Increasing ϵ parameter accommodates the case where $\bar{\mathbf{p}}$ is misspecified.

[2]Minor differences, in Group DRO accuracies, due to using random sampling (unlike weighted sampling in (Sagawa et al., 2019)). Weighted sampling is noisy when group labels are missing.

4. Conclusion

We present Worst-off DRO, an invariant learning method for partially labelled datasets. Worst-off DRO extends Group DRO by optimizing the loss against the worst-off group assignments in a constraint set. By reducing the constraint set with the marginal distribution, we reduce the optimization parameter space while maintaining the objective to be an upper bound to that of the Group DRO with true group assignments. By harnessing both labeled and unlabeled data in terms of group, we demonstrate in experiments that the Worst-off DRO outperforms both ERM, UnsupDRO, which do not make use of available group labels, as well as the Group DRO (Partial), which does not use unlabeled data.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *ICML*, 2018.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

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

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *JMLR*, 17 (83):1–5, 2016.

Dua, D., Graff, C., et al. Uci machine learning repository. 2017.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *ICML*, 2018.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *NeurIPS*, 2020.

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

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *arXiv preprint arXiv:2010.05893*, 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.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *ICCV*, 2015.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching. In *International Conference on Machine Learning*, pp. 7313–7324. PMLR, 2021.

Mohan, K. and Pearl, J. Graphical models for recovering probabilistic and causal queries from missing data. *Advances in Neural Information Processing Systems*, 27: 1520–1528, 2014.

Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., and Ver Steeg, G. Invariant representations without adversarial training. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/415185ea244ea2b2bedeb0449b926802-Paper.pdf.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, 2016.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proc ACM conference on health, inference, and learning*, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 2019.

Rahimian, H. and Mehrotra, S. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Rawls, J. *Justice as fairness: A restatement*. Harvard University Press, 2001.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *ICLR*, 2019.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *NeurIPS*, 2020.

Wasserman, L. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-ucsd birds 200. technical report cns-tr-2010-001. *California Institute of Technology*, 2010.

Xie, Q., Dai, Z., Du, Y., Hovy, E., and Neubig, G. Controllable invariance through adversarial feature learning. *arXiv preprint arXiv:1705.11122*, 2017.

Zhao, Y. and Udell, M. Matrix completion with quantified uncertainty through low rank gaussian copula. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20977–20988. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f076073b2082f8741a9cd07b789c77a0-Paper.pdf.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *PAMI*, 40(6):1452–1464, 2017.

A. Appendix

A.1. Ethics Statement

Machine Learning (ML) models that perform poorly on a minority group or environment have raised a lot of concerns within the AI community and broader society in recent years. To democratize ML in real world, learning ML models that perform robustly across groups or environments has become an important venue of research. The proposed Worst-off DRO is a versatile method that could be employed to train an invariant classifier across groups even when the group information is available only for the portion of the data. This is a rather practical scenario as the group information could be missing for various reasons during the data collection. We further emphasize the importance of theoretical result showing the objective of Worst-off DRO being an upper bound to that of Group DRO with complete group information for safety-critical ML applications.

A.2. Reproducibility

We write our experimental code from scratch using PyTorch library (Paszke et al., 2019). Due to its similarity, our implementation may closely follow that of Group DRO (Sagawa et al., 2019).[3] One of the key differentiation of Worst-off DRO is the inner maximization solver for the worst-off group assignments $\{\hat{g}\}$, which we elaborate the exact code using CVXPY solver (Diamond & Boyd, 2016) in Algorithm 2 of Appendix. Additional implementation details, including the neural network architectures, as well as value for hyperparameters including the learning rate, weight decay, batch size, number of training epochs, and algorithm-specific parameters are summarized in Table 2 and Section A.8 and A.11.

A.3. Proof of Lemmas

Lemma A.1. *Denote $\mathcal{L}_{\text{GDRO}}$ at a given w and q parameters as $\mathcal{L}_{\text{GDRO}(w,q)}$. Similarly $\mathcal{L}_{\text{WDRO}}(\mathcal{C})$ at a fixed w and q as $\mathcal{L}_{\text{WDRO}(w,q)}(\mathcal{C})$. When the ground-truth group assignment $\{g_i^\star\}_{i=1}^N \in \mathcal{C}$, we have*

$$\min_w \max_{q\in\Delta^M} \mathcal{L}_{\text{GDRO}(w,q)} \leq \min_w \max_{q\in\Delta^M} \mathcal{L}_{\text{WDRO}(w,q)}(\mathcal{C}) \tag{12}$$

Proof. Under the case $\{g_i^\star\}_{i=1}^N \in \mathcal{C}$, due to the max over $\mathcal{C}$, we have

$$\mathcal{L}_{\text{GDRO}(w,q)} \leq \mathcal{L}_{\text{WDRO}(w,q)}(\mathcal{C}) \quad \forall w, q \tag{13}$$

Define $q^*_{\text{WDRO}} = \arg\max_{q\in\Delta^M} \mathcal{L}_{\text{WDRO}(w,q)}(\mathcal{C})$ and $q^*_{\text{GDRO}} = \arg\max_{q\in\Delta^M} \mathcal{L}_{\text{GDRO}(w,q)}$.
From the above definitions, we have,

$$\mathcal{L}_{\text{WDRO}(w,q)}(\mathcal{C}) \leq \mathcal{L}_{\text{WDRO}(w,q^*_{\text{WDRO}})}(\mathcal{C}) \tag{14}$$

Moreover,

$$\mathcal{L}_{\text{GDRO}(w,q^*_{\text{GDRO}})} \leq \mathcal{L}_{\text{WDRO}(w,q^*_{\text{GDRO}})}(\mathcal{C}) \quad \text{from (13)} \tag{15}$$

$$\mathcal{L}_{\text{WDRO}(w,q^*_{\text{GDRO}})}(\mathcal{C}) \leq \mathcal{L}_{\text{WDRO}(w,q^*_{\text{WDRO}})}(\mathcal{C}) \quad \text{from (14)} \tag{16}$$

$$\implies \mathcal{L}_{\text{GDRO}(w,q^*_{\text{GDRO}})} \leq \mathcal{L}_{\text{WDRO}(w,q^*_{\text{WDRO}})}(\mathcal{C}) \tag{17}$$

Minimizing (17) over w, we obtain,

$$\min_w \max_{q\in\Delta^M} \mathcal{L}_{\text{GDRO}(w,q)} \leq \min_w \max_{q\in\Delta^M} \mathcal{L}_{\text{WDRO}(w,q)}(\mathcal{C})$$

□

Lemma A.2. *The constraint set $\mathcal{C}_{\mathbf{p}^\star,\epsilon}$ contains the true group labels $\{g_i^\star\}_{i=1}^N$ with high probability:*

$$P(\{g_i^\star\}_{i=1}^N \in \mathcal{C}_{\mathbf{p}^\star,\epsilon}) \geq 1 - 2Me^{-2N\epsilon^2}$$

[3] https://github.com/kohpangwei/group_DRO

Proof. The probability of the true group assignment $\{g_i^\star\}_{i=1}^N$ in the constraint set $\mathcal{C}_{\mathbf{p}^\star,\epsilon}$ is written as follows:

$$P(\{g_i^\star\}_{i=1}^N \in \mathcal{C}_{\mathbf{p}^\star,\epsilon}) = P\Big(\big|p_j^\star - \frac{1}{N}\sum\nolimits_{i=1}^N \mathbb{1}\{g_i^\star = j\}\big| \leq \epsilon \quad \forall j\Big) \geq 1 - 2Me^{-2N\epsilon^2} \tag{18}$$

where (18) holds true from the Hoeffding's inequality. □

Lemma A.3. *The constraint set $\mathcal{C}_{\bar{\mathbf{p}},\delta+\epsilon}$ contains the true group labels $\{g_i^\star\}_{i=1}^N$ with high probability:*

$$P(\{g_i^\star\}_{i=1}^N \in \mathcal{C}_{\bar{\mathbf{p}},\delta+\epsilon}) \geq 1 - 2e^{-2N\epsilon^2} - 2e^{-2K\delta^2}$$

Proof. Using Hoeffding's inequality, we can show that the estimation error of the marginal distribution is bounded by δ with high probability as follows:

$$P(|p_j^\star - \bar{p}_j| \leq \delta) = P\Big(\big|p_j^\star - \frac{1}{K}\sum\nolimits_{i=1}^K \mathbb{1}\{g_i^\star = j\}\big| \leq \delta\Big) \geq 1 - 2e^{-2K\delta^2} \tag{19}$$

Furthermore, we show using Hoeffding's inequality that

$$P\Big(\big|p_j^\star - \frac{1}{N}\sum\nolimits_{i=1}^N \mathbb{1}\{g_i = j\}\big| \leq \epsilon\Big) \geq 1 - 2e^{-2N\epsilon^2} \tag{20}$$

$$\begin{aligned}
&P\Big(\big|\bar{p}_j - \frac{1}{N}\sum\nolimits_{i=1}^N \mathbb{1}\{g_i = j\}\big| \leq \delta + \epsilon\Big) && (21)\\
&\qquad \geq P\Big(\Big\{\big|p_j^\star - \bar{p}_j\big| \leq \delta\Big\} \cap \Big\{\big|p_j^\star - \frac{1}{N}\sum\nolimits_{i=1}^N \mathbb{1}\{g_i = j\}\big| \leq \epsilon\Big\}\Big) && (22)\\
&\qquad \geq P\Big(\big|p_j^\star - \bar{p}_j\big| \leq \delta\Big) + P\Big(\big|p_j^\star - \frac{1}{N}\sum\nolimits_{i=1}^N \mathbb{1}\{g_i = j\}\big| \leq \epsilon\Big) - 1 && (23)\\
&\qquad \geq 1 - 2e^{-2K\delta^2} - 2e^{-2N\epsilon^2} && (24)
\end{aligned}$$

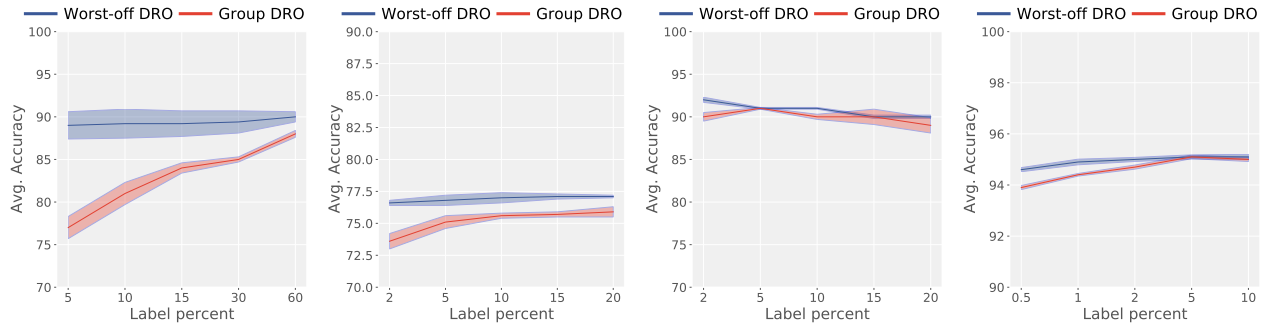
where (22) is due to that the intersection of events and is a subset of an event in (21). Then, (23) is derived using union bound. Now, the probability of the true group assignment $\{g_i^\star\}_{i=1}^N$ in the constraint set $\mathcal{C}_{\bar{\mathbf{p}},\delta+\epsilon}$ is written as follows:

$$\begin{aligned}
P(\{g_i^\star\}_{i=1}^N \in \mathcal{C}_{\bar{\mathbf{p}},\delta+\epsilon}) &= P\Big(\big|\bar{p}_j - \frac{1}{N}\sum\nolimits_{i=1}^N \mathbb{1}\{g_i = j\}\big| \leq \delta + \epsilon \quad \forall j\Big) && (25)\\
&= 1 - P\Big(\bigcup_{j=1}^M \big|\bar{p}_j - \frac{1}{N}\sum\nolimits_{i=1}^N \mathbb{1}\{g_i = j\}\big| > \delta + \epsilon\Big) && (26)\\
&\text{from union bound, we get,} && (27)\\
&\geq 1 - \sum_{j=1}^M P\Big(\big|\bar{p}_j - \frac{1}{N}\sum\nolimits_{i=1}^N \mathbb{1}\{g_i = j\}\big| > \delta + \epsilon\Big) && (28)\\
&\text{from (23), we have} && (29)\\
&\geq 1 - 2Me^{-2K\delta^2} - 2Me^{-2N\epsilon^2} && (30)
\end{aligned}$$

□

A.4. Notes on the Optimization Procedure

When using CVXPY to solve for the Worst-off DRO assignments, we simplify the problem by replacing the data marginal distribution $\sum_{i=1}^N \hat{g}_{ij}$ in the denominator of (10) to $\bar{\mathbf{p}}_j$, thus providing us with a convex optimization problem. The code for the solver is available in Algorithm 2.



(a) Waterbirds (b) CMNIST (c) Adult (d) CelebA

Figure 2: **Increasing the labelled samples - Average Group Accuracy.** We plot the average group accuracies as a function of labelled samples. These accuracies remain fairly similar as the count of labelled samples grows.

Table 2: **Grid search for Table 1.** The range of values for each hyper-parameter is listed. A grid search over these hyper-parameters is conducted to identify the best performing model. Models outside these range values were observed to be either unstable or not converging. Model selection is done based on NVP (novel validation procedure) where first the models, with higher overall accuracies, are selected. From the top five such performing models, the one with the highest minority group accuracy is picked.

Waterbirds		CMNIST	
Learning Rate	$0.0001, 0.00001, 0.000001$	Learning Rate	$0.001, 0.0001, 0.00001$
Weight Decay	$1.5, 1.0, 0.1$	Weight Decay	$0.01, 0.001, 0.0001$
η_{UDRO}	$0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3$	η_{UDRO}	$0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3$
η_{GDRO}	$0.1, 0.01, 0.001$	η_{GDRO}	$0.01, 0.001, 0.0001$
η_{WDRO}	$0.1, 0.01, 0.001$	η_{WDRO}	$0.01, 0.001, 0.0001$

Adult		CelebA	
Learning Rate	$0.001, 0.0001, 0.00001$	Learning Rate	$0.0001, 0.00001, 0.000001$
Weight Decay	$0.01, 0.001, 0.0001$	Weight Decay	$1.0, 0.1, 0.01$
η_{UDRO}	$0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3$	η_{UDRO}	$0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3$
η_{GDRO}	$0.01, 0.001, 0.0001$	η_{GDRO}	$0.1, 0.01, 0.001$
η_{WDRO}	$0.01, 0.001, 0.0001$	η_{WDRO}	$0.1, 0.01, 0.001$

A.5. An example of worst-off assignments

Using three samples, we provide an example of the worst-off assignments made by our algorithm,

Example A.4. Consider three samples with loss values $l_1 > l_2 > l_3$ and two predefined groups. Assume the marginal probabilities $\bar{\mathbf{p}}_1 = 0.6$ and $\bar{\mathbf{p}}_2 = 0.4$. Without loss in generality, assume $\frac{q_1}{\bar{\mathbf{p}}_1} > \frac{q_2}{\bar{\mathbf{p}}_2}$. With constraint $\mathcal{C}_{\bar{\mathbf{p}},\epsilon=0}$ and solving for Worst-off DRO objective results in the following group assignments, $\{\hat{g}^t\} = \begin{pmatrix} 1 & 0 \\ 0.8 & 0.2 \\ 0 & 1 \end{pmatrix}$. Here, the ith row indicates the assignment given to sample l_i.

The group assignments can be derived by identifying a $\{\hat{g}\}$ that satisfies the constraints $\sum_{i=1}^{N} \hat{g}_{i1} \leq N\bar{\mathbf{p}}_1$ and $\sum_{i=1}^{N} \hat{g}_{i2} \leq N\bar{\mathbf{p}}_2$, where $N = 3$, $\bar{\mathbf{p}}_1 = 0.6$ and $\bar{\mathbf{p}}_1 = 0.4$, and correspondingly maximizes Worst-off DRO objective. The above example informs us that group assignments depend on the magnitude of loss values in addition to the group weights and marginal probabilities. As indicated in the paper, we find that *high loss samples are assigned to groups with high group weights and low marginal probabilities*, characteristic of a worst-off group.

Marginal constraints form a key ingredient of our algorithm as per the above example. Without the marginal constraints, the group assignments $\{\hat{g}^t\} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$. That is, the assignments would have been made independent of the loss values and sparsely restricted to the group with large $\frac{q_j}{\bar{\mathbf{p}}_j}$ value.

Table 3: **Hyperparamter choices for Table 1.** We list the hyper-parameters selected using the NVP procedure (see Section 3) after performing grid-search. Learning rate and weight decay are an important set of parameters that influences the minority group performance. Each baseline has it's algorithm-specific hyper-parameter such as step-size of the simplex weights in Group DRO (η_{GDRO}), the loss threshold in Unsup DRO (η_{UDRO}) and the step size for the group weights in Worst-off DRO (η_{WDRO}). The symbol "-" for batchsize in CMNIST experiments indicate the use of full-batch data for training.

Dataset	Method	Architecture	Learning Rate	Weight Decay	Batch Size	# Epochs	Other params
Waterbirds	ERM	ResNet50	0.0001	0.1	128	300	-
Waterbirds	Unsup DRO	ResNet50	0.0001	0.1	128	300	η=0.3
Waterbirds	Group DRO-(Oracle)	ResNet50	0.00001	1.0	128	300	η=0.001
Waterbirds	Group DRO-(Partial)	ResNet50	0.00001	0.1	128	300	η=0.001
Waterbirds	Worst-off DRO	ResNet50	0.00001	1.0	128	300	η=0.001
CMNIST	ERM	MLP(390,390)	0.001	0.01	-	500	-
CMNIST	Unsup DRO	MLP(390,390)	0.00001	0.001	-	500	η=0.4
CMNIST	Group DRO-(Oracle)	MLP(390,390)	0.0001	0.001	-	500	η=0.001
CMNIST	Group DRO-(Partial)	MLP(390,390)	0.001	0.01	-	500	η=0.001
CMNIST	Worst-off DRO	MLP(390,390)	0.0001	0.01	-	500	η=0.0001
Adult	ERM	MLP(64,32)	0.0001	0.001	128	200	-
Adult	Unsup DRO	MLP(64,32)	0.0001	0.001	128	200	η=0.3
Adult	Group DRO-(Oracle)	MLP(64,32)	0.0001	0.001	128	200	η=0.0001
Adult	Group DRO-(Partial)	MLP(64,32)	0.0001	0.01	128	200	η=0.001
Adult	Worst-off DRO	MLP(64,32)	0.00001	0.001	128	200	η=0.0001
CelebA	ERM	ResNet50	0.0001	0.01	128	50	-
CelebA	Unsup DRO	ResNet50	0.0001	0.01	128	50	η=0.6
CelebA	Group DRO-(Oracle)	ResNet50	0.00001	0.1	128	50	η=0.1
CelebA	Group DRO-(Partial)	ResNet50	0.00001	0.01	128	50	η=0.1
CelebA	Worst-off DRO	ResNet50)	0.00001	0.1	128	50	η=0.001

A.6. Discussion on Unsupervised DRO methods

In this section, we contrast Worst-off DRO method against Unsup DRO (Hashimoto et al., 2018) and CVaR DRO (Levy et al., 2020). CVaR DRO (Levy et al., 2020) is a coherent risk measure (Rahimian & Mehrotra, 2019) that optimizes over a certain fixed-sized sub-populations within the training dataset. In essence, CVaR DRO is alike Unsup DRO where the size of the sub-population is controlled by a threshold on the loss value. In both CVaR DRO and Unsup DRO, the size of the selected sub-population needs to be close to the size of the smallest group as identified in Section 3.2.2 of (Liu et al., 2021). Such a requirement demands wider hyper-parameter search space for α/η parameters that control the size of the sub-populations. Our experiments justify this need, Table 2 of Appendix A.8 shows that the search space of Unsup DRO is twice relative to Worst-off DRO in order to attain comparable average group accuracies. Clearly, a wider search space contributes to a harder model selection procedure. Moreover, scenarios where extensive search is not possible (eg, small validation set/dataset regimes) could result in incorrect/unstable model selection. From the perspective of the methodology, CVaR DRO / Unsup DRO train only on the highest loss samples while discarding the remaining samples. In contrast, Worst-off DRO does not discard any sample rather downweights/upweights as per the worst-off group assignment. This property aids in maintaining a high overall accuracy besides reaching good minority group accuracy.

A.7. Discussion on MAR case

The δ gap in Lemma 2.3 captures the error in misspecification of $\bar{\mathbf{p}}$ in relation to $\mathbf{p}^\star$. When $\bar{\mathbf{p}}$ is misspecified due to the data being Missing at Random (MAR) rather than MCAR (Missing Completely at Random), a solution could be to estimate the propensity of missingness from other features; then use inverse propensity weighting to get a consistent estimate of the fraction of samples in each group as discussed in (Zhao & Udell, 2020). Alternatively, if provided with the knowledge of the data-generation process, the core effort in extending our method simply involves using off-the-shelf estimators to characterize the probability distributions (see (Mohan & Pearl, 2014) for example.

Algorithm 2 Group Assignment Solver using CVXPY library

```
import cvxpy as cp
import numpy as np

class Solver(object):
  def __init__(self, n_controls, bsize, marginals, epsilon, labeled=None):
    """Group assignment solver.

    Arguments:
      n_controls: An integer for the number of groups.
      bsize: An integer for the batch size.
      marginals: A 2D array for the marginal distribution.
      epsilon: A float for the variance.
      labeled: A tuple for labeled data indices and their value.
    """
    self.X = cp.Variable((bsize, n_controls))
    self.l = cp.Parameter((bsize, 1))
    self.p = cp.Parameter((n_controls, 1), value=marginals)
    self.q = cp.Parameter(n_controls)
    if labeled is not None:
      labeled_idx, labeled_value = labeled
    counts = cp.sum(self.X, axis=0, keepdims=True)

    obj = ((self.l.T @ self.X) / self.p.T) @ self.q
    constraints = [self.X >= 0,
                   cp.sum(self.X, axis=1, keepdims=True) == np.ones((bsize, 1)),
                   cp.abs(cp.sum(self.X, axis=0, keepdims=True) / bsize - self.p.T) <= epsilon]
    if labeled is not None:
      constraints += [self.X[labeled_idx] == labeled_value]

    self.prob = cp.Problem(cp.Maximize(obj), constraints)

  def cvxsolve(self, losses, weights):
    """Solver.

    Arguments:
       losses: A 2D array for loss values.
       weights: A 1D array for group weights q.

    Returns:
      A 2D array for soft group assignments.
    """
    self.l.value = losses
    self.q.value = weights
    self.prob.solve()
    return self.X.value
```

A.8. Hyper-parameter Tuning

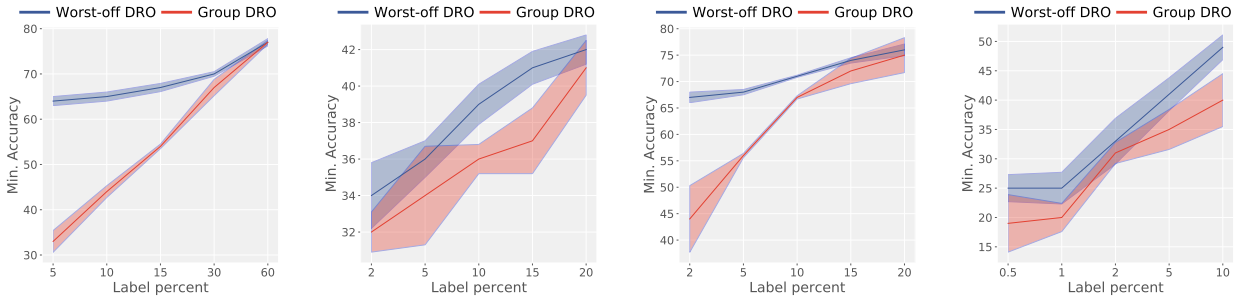
Hyper-parameters were selected for each algorithm by performing an NVP procedure (see Section 3). The best performing model was identified on the validation set associated with each dataset. All the measures were computed and averaged over three random runs. A list of all the hyper-parameters that were tuned for are available in Table 2. The final hyper-parameters selected for each method can be viewed from Table 3.

A.9. Additional ablation experiments

In this section, we extend the ablation experiments presented in the main paper. Specifically, results over all four datasets is provided in Figures 3,4,5. As we see in the plots, all the datasets exhibit similar behaviour across the ablation experiments.

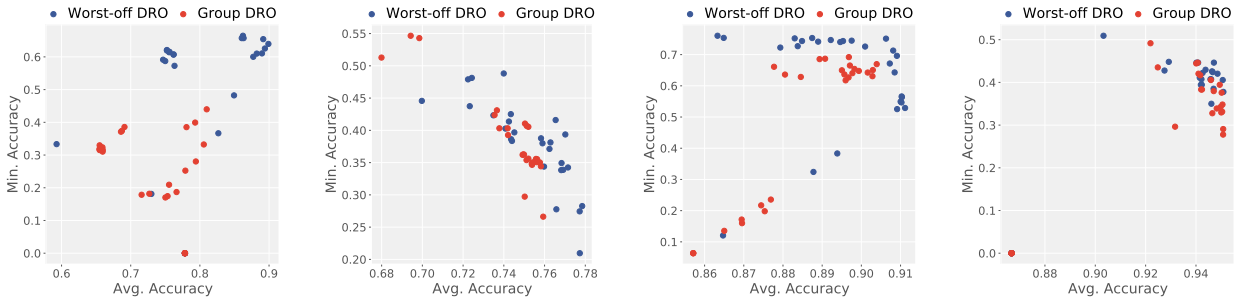
A.10. Ablation study on increasing the constraint set size.

We conduct experiments on Worst-off DRO method for different values of the ϵ parameter in the set $\{0, 0.001, 0.01, 0.1, 1\}$. The test set accuracies on the minority group and average group are reported in Figure 6. Increasing the ϵ value also increases the constraint set size because the marginal constraint is gradually relaxed. Figure 6 shows that the both minority group accuracy and average group accuracy values reduce with increase in ϵ value beyond 0.1 threshold. The accuracy values for $\epsilon \leq 0.01$ are comparable. A similar trend holds on other datasets as well.



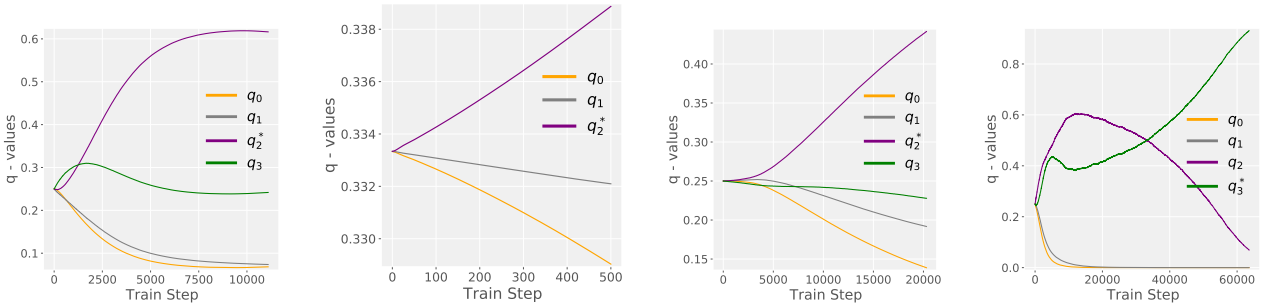
(a) Waterbirds (b) CMNIST (c) Adult (d) CelebA

Figure 3: **Increasing the labelled samples.** Minority group accuracies are plotted at different counts of the labelled samples. Both, Group DRO (Partial) and Worst-off DRO algorithms improve the minority group accuracies with more training labels. Also, the Worst-off DRO method has higher accuracy values than Group DRO method. The average group accuracies are shown in the Appendix Figure 2.



(a) Waterbirds (b) CMNIST (c) Adult (d) CelebA

Figure 4: **Minority Group vs. Average Accuracy.** Evaluations for different hyper-parameter choices are plotted for Worst-off DRO and Group DRO (Partial) methods. Models from Worst-off DRO training are concentrated in the top-right corner of the plots. This is desirable indicating a high accuracies across the two metrics. For model selection from among the possible choices, we adopt the NVP procedure (see Section 3).



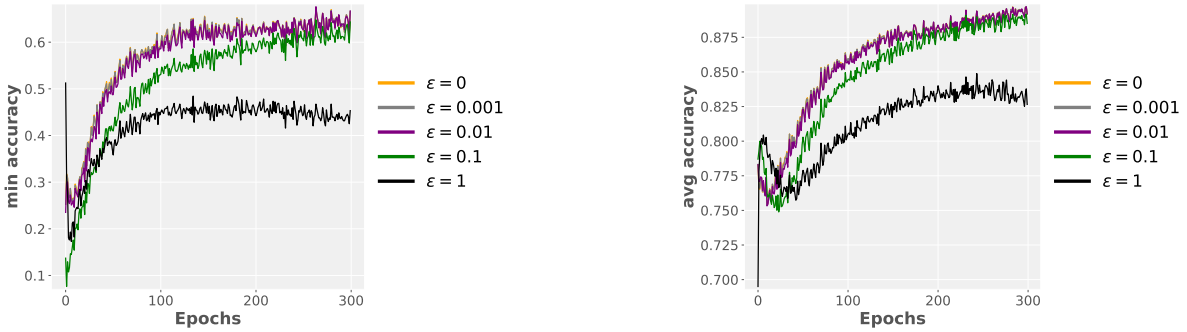
(a) Waterbirds (b) CMNIST (c) Adult (d) CelebA

Figure 5: **Progression of group weights.** The evolution of $q-$values (see Algorithm 1) is plotted for each group. The $q-$values for the minority groups increases gradually while those of the majority groups reduce. A high $q-$value indicates that the corresponding group receives a higher weight relative to other groups. In the plots, minority group is indicate by a $*$ on q.

A.11. More details on the datasets

A.11.1. Waterbirds

The dateset, used in (Sagawa et al., 2019), comprises of 4795 images of birds from the CUB dataset (Welinder et al., 2010) and the backgrounds taken from the Places dataset (Zhou et al., 2017). Each image in the dataset has a background of land or water. The target labels are either "landbirds" or "waterbirds". In this dataset the groups "landbirds" on water and "waterbirds" on land form a minority. A ResNet50 model, pre-trained with ImageNet weights, has been used for training in experiments on this dataset. No data augmentation has been applied for any of the Algorithms.



(a) Waterbirds Minority Group Accuracy

(b) Waterbirds Average Group Accuracy

Figure 6: **Varying the ϵ parameter in the constraints.** The marginal constraint is gradually relaxed by increasing the ϵ parameter. The accuracies in the plots are computed on the test sets. Performance of the models with $\epsilon \leq 0.01$ are similar, however, the accuracies drop when increasing ϵ beyond 0.01 threshold.

A.11.2. Group CMNIST

CMNIST, derived from an MNIST (LeCun et al., 1998), is a digit recognition dataset where each image is colored either red or green. Digits $< 5/ \geq 5$ are considered as label $0/1$. We consider three groups in our experiments. In the first two groups, label 0 images are predominantly colored red and vice versa. In the third group, which forms a minority, we switch coloring such that the label 1 images are predominantly colored red. Specifically, for the first two groups, the color id is sampled by flipping the target label with probabilities 0.2 and 0.1 respectively, while the third group with probability 0.9. Both training and testing sets contain three groups. The overall setup for generating a given group is similar to (Arjovsky et al., 2019). However, unlike (Arjovsky et al., 2019) the training and testing phases share the set of three pre-defined groups so as to evaluate for group-robustness criterion.

A.11.3. Group Adult

We use a semi-synthetic version of the Adult dataset (Dua et al., 2017) for this experiment. Similar to (Lahoti et al., 2020), we consider race and sex as the four demographic groups. The target label is income $> 50K\$$. Similar to the CMNIST dataset, each group has a different correlation strength to the target label. For the purposes of the experiment, we exaggerate these spurious correlations caused by group membership close to (Creager et al., 2021). However, distinct from (Creager et al., 2021), the training and testing phases share same set of groups. For samples with group label as Afican-American, we undersample examples with probability $P(y = 1 \mid \text{group}) = 0.06$ whereas for the non African-American group labels, we oversample examples with probability $P(y = 1 \mid \text{group}) = 0.94$.

A.11.4. CelebA

CelebA (Liu et al., 2015) is a dataset containing about 200k celebrity faces curated from the internet. Similar to (Sagawa et al., 2019), we aim to predict the target attribute *Blond Hair* that is spuriously correlated to the Gender attribute. The minority group in this dataset are the images with attributes (blond, male). For this dataset, the official train-val-test splits as recommended by (Liu et al., 2015) has been used. Similar to the Waterbirds experiments, a pre-trained ImageNet-based ResNet50 model has been used for the implementations.

A.12. Related work

Group Robust Optimization. Methods in the literature handling robustness to extraneous attributes can be broadly categorized into two classes. The first class, domain generalization methods (Arjovsky et al., 2019; Mahajan et al., 2021; Moyer et al., 2018), aim at learning representations invariant to a predefined set of extraneous attributes or groups. The goal is to be able to generalize to unseen domains or environments in the testing phase. On the other hand, the second class of methods, called as the group robust methods (Oakden-Rayner et al., 2020; Sagawa et al., 2019; Liu et al., 2021), seek to improve the worst-off or the minority group performance within the set of pre-defined groups. Here the training and testing phases share the same set of groups. Our approach falls into the second class of methods.

Robust Optimization with Demographics. When group information is known at train time, Group DRO (Hu et al., 2018;

Sagawa et al., 2019) or Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) could be employed to improve the performance over multiple groups. Specifically, Group DRO proceeds by minimizing the loss of the group with the largest loss, while IRM enforces a shared predictor across multiple environments to be optimal.

Robust Optimization without Demographics. Several works have focused on developing methods that remove the dependence on the group labels. (Hashimoto et al., 2018) proposed to minimize the loss of the samples with losses larger than a certain threshold. (Lahoti et al., 2020) developed an adversarial re-weighting scheme that assigns large weight to high loss samples. Recently, (Liu et al., 2021) proposed a simple yet effective two-stage approach called Just-Train-Twice (JTT) that trains a model by upweighting samples with high losses from the initial ERM model. Lastly, EIIL (Creager et al., 2021) also proposed a two-stage method where in the first stage the group or environment labels are inferred and in the second stage a Group DRO or an IRM optimization is employed on inferred labels.

Drawbacks of two-stage methods. While two-stage methods, like JTT (Liu et al., 2021) and EIIL (Creager et al., 2021), have demonstrated significant improvements in minority group accurcies, they bear few drawbacks in relation to single-stage methods. Firstly, two-stage methods introduce additional set of hyper-parameters. For example, it's crucial for JTT to tune for the number of epochs in its first stage. Similarly, several hyper-parameters are introduced in EIIL method especially in optimizing the EI objective and in identifying a pre-trained reference model. Secondly, in a two-stage model, a failed first stage leads to an unsuccessful second stage. This is because the errors from the first stage are propagated to the later stages. For example, a first stage model could fail in JTT due to model overfitting, and likewise an inaccurate group inference in the EIIL method may block second-stage invariant learning besides raising ethical issues on pseudo-label misuse. In summary, efforts to reduce a two-stage model to a single-stage method are beneficial and, as we shall see shortly, our proposal fits in class of single stage methods.