The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Distributionally Robust Optimization with Probabilistic Group

Soumya Suvra Ghosal, Yixuan Li

Department of Computer Sciences, University of Wisconsin – Madison
{sghosal, sharonli}@cs.wisc.edu

Abstract

Modern machine learning models may be susceptible to learning spurious correlations that hold on average but not for the atypical group of samples. To address the problem, previous approaches minimize the empirical worst-group risk. Despite the promise, they often assume that each sample belongs to *one and only one group*, which does not allow expressing the uncertainty in group labeling. In this paper, we propose a novel framework **PG-DRO**, which explores the idea of probabilistic group membership for distributionally robust optimization. Key to our framework, we consider soft group membership instead of hard group annotations. Our framework accommodates samples with group membership ambiguity, offering stronger flexibility and generality than the prior art. We comprehensively evaluate PG-DRO on both image classification and natural language processing benchmarks, establishing superior performance.

1 Introduction

A major challenge in training robust models is the presence of *spurious correlations*—misleading heuristics imbibed within the training dataset that are correlated with most examples but do not hold in general. For example, consider the Waterbirds dataset (Sagawa et al. 2020a), which involves classifying bird images as waterbird or landbird. Here, the target label (bird type) is spuriously correlated with the background, *e.g.*, an image of WATERBIRD has a higher probability to appear on WATER background. Machine learning models, when trained on such biased datasets using empirical risk minimization (ERM), can achieve high average test accuracy but fail significantly on rare and untypical test examples lacking those heuristics (such as WATERBIRD on LAND) (Sagawa et al. 2020a; Geirhos et al. 2019; Sohoni et al. 2020). Such disparities in model prediction can lead to serious ramifications in applications where fairness or safety are important, such as facial recognition (Buolamwini and Gebru 2018) and medical imaging (Oakden-Rayner et al. 2020). This calls for the need of ensuring group robustness, *i.e.*, high accuracy on the under-represented groups.

Over the past few years, a line of algorithms (Sagawa et al. 2020a; Zhang et al. 2020; Mohri, Sivek, and Suresh 2019; Goel et al. 2021) have been proposed to improve group robustness. The core idea behind one of the most common algorithms, G-DRO (Sagawa et al. 2020a), involves minimizing the loss for the worst-performing group during training. Despite the promise, existing approaches suffer from a fundamental limitation — they assume each sample belongs to *one and only one group*, which does not allow expressing the uncertainty in group labeling. For example, in Figure 1, we show samples from the Waterbirds dataset (Sagawa et al. 2020a), where the background consists of features of *both* land and water, displaying inevitable ambiguity. In such cases of group ambiguity, using hard group labels can result in the loss of information in robust optimization, and disproportionally penalize the model. Taking the LANDBIRD in Figure 1 (right) as an example, a hard assignment of the water attribute would incur an undesirably high sample-wise loss for being associated with the land attribute. To date, few efforts have been made to resolve this.

Motivated by this, we propose a novel framework **PG-DRO**, which performs a distributionally robust optimization using *probabilistic groups*. Our framework emphasizes the uncertain nature of group membership, while minimizing the worst-group risk. Our key idea is to introduce the "soft" group membership, which relaxes the hard group membership used in previous works. The probabilistic group membership can allow input to be associated with multiple groups, instead of always selecting one group during robust optimization. We formalize the idea as a new robust optimization objective PG-DRO, which scales the loss for each sample based on the probability of a sample belonging to each group. Our formulation thus accommodates samples with group membership ambiguity, offering stronger flexibility and generality than G-DRO (Sagawa et al. 2020a).

As an integral part of our framework, PG-DRO tackles the challenge of estimating the group probability distribution. We aim to provide a solution with (almost) full autonomy without expensive manual group annotation. Our proposed training algorithm PG-DRO can be flexibly used in conjunction with a variety of pseudo labeling approaches generating group probabilities. Our method henceforth alleviates the heavy data requirement in G-DRO, where each sample needs to be annotated with a group label. In particular, we showcase multiple instantiations through pseudo labeling approach using a small amount of group-labeled data (Section 5) or even without

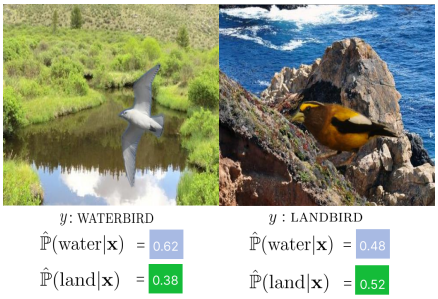

Figure 1: Visual illustration group membership ambiguity. Images are from Waterbirds (Sagawa et al. 2020a) dataset.

any group-annotated data (Appendix A). In all cases, PG-DRO achieves performance comparable to or outperforms G-DRO (Sagawa et al. 2020a), while significantly reducing group annotation.

Extensive experiments support the superiority of using group probabilities to minimize the worst-group loss. We comprehensively evaluate PG-DRO on both image classification (Section 4.1) and natural language processing (Section 4.2) benchmarks, where PG-DRO establishes *state-of-the-art* performance. On CelebA (Liu et al. 2015), using only 5% of validation data (988 samples), PG-DRO achieves worst-group accuracy of 89.4% and outperforms G-DRO (88.7%) requiring group-annotation on entire train and validation set (182637 samples). Lastly, we provide a better understanding of the benefits of our method (Section 6), and validate that PG-DRO effectively learns a more robust decision boundary than G-DRO. Training code is available at https://github.com/deeplearning-wisc/PG-DRO.

Our **key contributions** are summarized as follows:

1. We propose PG-DRO, a novel distributionally robust optimization framework for enhancing group robustness. During training, PG-DRO leverages probabilistic group membership, which allows expressing the ambiguity and uncertainty in group labeling. To the best of our knowledge, we are the first to explore the idea of assigning group probabilities as opposed to hard group membership.
2. We perform extensive experimentation on a set of vision and language processing datasets to understand the efficacy of PG-DRO. Specifically, we compare PG-DRO against the five best-performing learning algorithms to date. For both computer vision and NLP tasks, PG-DRO consistently outperforms competing methods and establishes superior performance.
3. We provide extensive ablations to understand the impact of each component in our proposed framework: pseudo group labeling and robust optimization. We observe that irrespective of the pseudo group labeling approach used, PG-DRO consistently outperforms G-DRO in all cases.

2 Preliminaries

In this paper, we consider the problem of learning a classifier when the training data has correlations between true labels and spurious attributes. More generally, spurious attributes refer to statistically informative features that work for the majority of training examples but do not necessarily capture cues related to the labels (Sagawa et al. 2020a; Geirhos et al. 2019; Goel et al. 2021; Tu et al. 2020). Recall the setup in WATERBIRD vs LANDBIRD classification problem, where the majority of the training samples has target label (WATERBIRD or LANDBIRD) spuriously correlated with the background features (WATER or LAND background). Sagawa *et al.* (Sagawa et al. 2020a) showed that deep neural networks can rely on these spurious features to achieve high accuracy on average, but fail significantly for groups where such correlations do not hold.

Problem Setup. Formally, we consider a training set $\mathcal{D}_{\text{train}}$ consisting of N training samples: $\{\mathbf{x}_i, y_i\}_{i=1}^N$. The samples are drawn *i.i.d.* from a probability distribution: $\mathcal{P}_{X,Y}$. Here, $X \in \mathcal{X}$ is a random variable defined in the input space, and $Y \in \mathcal{Y} = \{1, \dots, K\}$ represents its label. We further assume that the data is sampled from a set of E environments $\mathcal{E} = \{e_1, e_2, \cdots, e_E\}$. The training data has spurious correlations, if the input $\mathbf{x}_i$ is generated by a combination of invariant features $\mathbf{z}_i^{\text{inv}} \in \mathbb{R}^{d_{\text{inv}}}$, which provides essential cues for accurate classification, and environmental features $\mathbf{z}_i^e \in \mathbb{R}^{d_e}$ dependent on environment e:

$$\mathbf{x}_i = \rho(\mathbf{z}_i^{\text{inv}}, \mathbf{z}_i^e).$$

Here ρ represents a function transformation from the feature space $[\mathbf{z}_i^{\text{inv}}, \mathbf{z}_i^e]^\top$ to the input space $\mathcal{X}$. Under the data model, we form groups $g = (y, e) \in \mathcal{Y} \times \mathcal{E} =: \mathcal{G}$ that are jointly determined by the label y and environment e.

Based on the data model, each sample $(\mathbf{x}, y, g)$ can be denoted as a tuple consisting of input $\mathbf{x} \in \mathcal{X}$, label $y \in \mathcal{Y}$, and group label $g \in \mathcal{G}$. The standard aim is to train a parameterized model $f_\theta : \mathcal{X} \to \mathcal{Y}$ that minimizes the expected loss $\mathbb{E}_{(\mathbf{x},y,g)\sim\mathcal{P}}[l(f_\theta(\mathbf{x}), y)]$ under the training distribution $\mathcal{P}$, for some loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.

G-DRO. Sagawa et al. (2020a) proposed group distributionally robust optimization (G-DRO), which minimizes the maximum of the expected loss among the groups:

$$\mathcal{R}_{\text{G-DRO}}(\theta) = \max_{g\in\mathcal{G}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_g}[l(f_\theta(\mathbf{x}), y)],$$

where $\mathcal{P}_g$ indexed by $g \in \mathcal{G} = \{1, 2, \cdots, |\mathcal{G}|\}$ denotes group-conditioned data distribution. Given a dataset consisting of N training points $\{\mathbf{x}_i, y_i, g_i\}_{i=1}^N$, G-DRO minimizes the empirical worst-group risk:

$$\hat{f}_{\text{G-DRO}} = \arg\min_{\theta\in\Theta}\{\max_{g\in\mathcal{G}} \frac{1}{n_g}\sum_{i=1}^N \mathbb{I}\{g_i = g\} l(f_\theta(\mathbf{x}_i), y_i)\},$$

where n_g represents the number of samples in each group $g \in \mathcal{G}$. In particular, G-DRO assumes that a given sample $\mathbf{x}$ can belong to *only one group* g, which does not allow expressing the uncertainty in group labeling. To see this, in Figure 1, we show examples from Waterbirds dataset (Sagawa et al. 2020a), where the background is associated with both LAND and WATER. Using hard group labels in such cases, can encode imprecise information about environmental attributes.

3 Proposed Method

We propose a novel distributional robust optimization framework with probabilistic group (dubbed **PG-DRO**). Our framework emphasizes the probabilistic nature of group membership, while minimizing the worst-group risk. We proceed to describe the method in detail.

3.1 Robust Optimization Objective

Probabilistic Group. We first introduce the notion of *probabilistic group membership*, which relaxes the hard group annotation used in G-DRO. Our key idea is to consider the "soft" probabilities for an input $\mathbf{x}$ to be in environments $\mathcal{E} = \{e_1, e_2, \cdots, e_E\}$. Formally, we denote $\hat{\mathbb{P}}(e_i|\mathbf{x})$ as the *estimated* probability of an input $\mathbf{x}$ associated with environment e_i, where:

$$\sum_{i=1}^{E} \hat{\mathbb{P}}(e_i|\mathbf{x}) = 1. \tag{1}$$

Here we temporarily assume we have some estimation of the probability and we will describe the means of estimation in the following Section 3.2.

Given $\hat{\mathbb{P}}(e_i|\mathbf{x})$, the probabilistic group label is defined as a vector $\hat{\mathbb{Q}}(\mathbf{x}) \in \mathbb{R}^{|\mathcal{G}|}$ such that:

$$\hat{\mathbb{Q}}(\mathbf{x})_g = \begin{cases} \hat{\mathbb{P}}(e_i|\mathbf{x}) & \text{if } g = (y, e_i), \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $g = \{1, 2, \cdots, |\mathcal{G}|\}$ represents group index, and $|\mathcal{G}| = |\mathcal{Y}| \times |\mathcal{E}|$ indicates total number of groups. Note that our definition generalizes the hard group labeling used in G-DRO, where $\hat{\mathbb{Q}}(\mathbf{x})_g = 1$ for the assigned group and 0 elsewhere.

Probabilistic Group-DRO. With the definition above, we are ready to introduce our new learning objective called Probabilistic Group-DRO (PG-DRO). Specifically, given N training examples comprising triplets $\{\mathbf{x}_i, y_i, \hat{\mathbb{Q}}(\mathbf{x}_i)\}_{i=1}^{N}$, we define the empirical worst-group risk as :

$$\hat{\mathcal{R}}_{\text{PG-DRO}}(\theta) = \max_{g \in \mathcal{G}} \left\{ \frac{1}{\widetilde{n}_g} \sum_{i=1}^{N} \hat{\mathbb{Q}}(\mathbf{x}_i)_g \cdot l(f_\theta(\mathbf{x}_i), y_i) + \frac{C}{\sqrt{\widetilde{n}_g}} \right\}, \tag{3}$$

where $\hat{\mathbb{Q}}(\mathbf{x}_i)_g$ represents the probability of an input $\mathbf{x}_i$ belonging to group g, and C is a hyper-parameter modulating model capacity. In effect, the group probability is used as the co-efficient to scale the sample-wise loss. For any group $g \in \mathcal{G}$, we define $\widetilde{n}_g = \sum_{i=1}^{N} \hat{\mathbb{Q}}(\mathbf{x}_i)_g$ as the sum of membership of all samples to group g. The scaling term $1/\sqrt{\widetilde{n}_g}$ constraints the model from overfitting the larger groups and focus on smaller groups. The adjustment parameter $C/\sqrt{\widetilde{n}_g}$ helps reduce the generalization gap for each group (Sagawa et al. 2020a). Finally we obtain the optimal model by minimizing the risk in Equation 3:

$$\hat{f}_{\text{PG-DRO}} = \arg\min_{\theta \in \Theta} \hat{\mathcal{R}}_{\text{PG-DRO}}(\theta).$$

Note, when $\hat{\mathbb{Q}}(\mathbf{x})$ tends to one-hot encoding for all samples in the training set, indicating that a given input $\mathbf{x}$ is absolutely certain to belong to a particular group, $\hat{\mathcal{R}}_{\text{PG-DRO}}(\theta)$ reduces to $\hat{\mathcal{R}}_{\text{G-DRO}}(\theta)$. PG-DRO henceforth provides a more general and flexible formulation than G-DRO.

3.2 Pseudo Group Labeling

In PG-DRO, towards estimating the probabilistic group labeling $\hat{\mathbb{Q}}(\mathbf{x})$, our aim is to propose a solution with (almost) full autonomy without expensive group annotation manually. Our method alleviates the heavy requirement in G-DRO, where each sample needs to be annotated with group information.

Specifically, given an input $\mathbf{x}$ sampled from a set of environments $\mathcal{E} = \{e_1, e_2, ..., e_E\}$, we would like to first estimate each $\hat{\mathbb{P}}(e_i|\mathbf{x})$. For this, we propose training a parameterized classifier $f_\phi : \mathcal{X} \to \mathcal{E}$, that can predict the spurious environment attribute e. To train the model, we use a small set of group-labeled samples, $\mathcal{D}_{\text{L}} = \{(\mathbf{x}_1, y_1, g_1), \cdots, (\mathbf{x}_m, y_m, g_m)\}$, such that $m \ll N$. The overall objective function consists of a supervised loss for samples in $\mathcal{D}_{\text{L}}$:

$$\hat{f}_{\text{env}} = \arg\min_{\phi \in \Phi} \{\mathbb{E}_{(\mathbf{x},y,g) \sim \mathcal{D}_L} [\ell_{\text{CE}}(f_\phi(\mathbf{x}), e)]\}, \tag{4}$$

where $g = (y, e)$ and ℓ_{CE} represents standard cross-entropy loss. We use weighted sampling based on the frequency of the samples in $\mathcal{D}_{\text{L}}$. Finally, for every sample $\mathbf{x} \in \mathcal{D}_{\text{train}}$, we use the trained model, $\hat{f}_{\text{env}}$, to predict the probability estimate $\hat{\mathbb{P}}(e|\mathbf{x})$ for each environment $e \in \{e_1, e_2, ..., e_E\}$. We obtain group probabilities, $\hat{\mathbb{Q}}(\mathbf{x})$, as defined in Equation 2 and use it for training PG-DRO.

Remark. Note that our proposed training objective PG-DRO can be flexibly used in conjunction with other pseudo labeling approaches generating group probabilities. Beyond pseudo labeling with supervised loss above, we also showcase the strong feasibility of alternative group labeling approaches. We provide further details in Section 5 and Appendix A.

4 Experiments

In this section, we comprehensively evaluate PG-DRO on both computer vision tasks (Section 4.1) and natural language processing (Section 4.2) containing spurious correlations.

4.1 Evaluation on Image Classification Benchmarks

Datasets. In this study, we consider two common image classification benchmarks: Waterbirds (Sagawa et al. 2020a) and CelebA (Liu et al. 2015).

(1) CelebA (Liu et al. 2015): Training samples in CelebA have spurious associations between target label and demographic information such as gender. We use the label space $\mathcal{Y} = \{\text{BLOND HAIR}, \text{DARK HAIR}\}$ and gender as the spurious feature, $\mathcal{E} = \{\text{MALE}, \text{FEMALE}\}$. The training data consists of 162770 images with 1387 in the smallest group, *i.e.*, MALE with BLOND HAIR.

(2) Waterbirds (Sagawa et al. 2020a): Introduced in (Sagawa et al. 2020a), this dataset contains spurious correlation between the background features and target label $y \in \{\text{WATERBIRD}, \text{LANDBIRD}\}$. The dataset is constructed by

Method	Dataset with group label	Waterbirds		CelebA	
		Avg. Acc	Worst Group Acc	Avg. Acc	Worst Group Acc
ERM (Vapnik 1991)	None	97.3	63.2	95.6	47.2
CVaR DRO (Levy et al. 2020)	val. set	96.0	75.9	82.5	64.4
LfF (Nam et al. 2020)	val. set	91.2	78.0	85.1	77.2
EIIL (Creager, Jacobsen, and Zemel 2021)	val. set	96.9	78.7	91.9	83.3
JTT (Liu et al. 2021)	val. set	93.3	86.7	88.0	81.1
SSA (Nam et al. 2022)	val. set	92.2	89.0	92.8	89.8
PG-DRO (Ours)	val. set	$92.5_{\pm 0.5}$	$\mathbf{91.0}_{\pm 0.6}$	$92.2_{\pm 0.4}$	$\mathbf{90.0}_{\pm 0.9}$
G-DRO (Sagawa et al. 2020a)	train & val. set	$92.4_{\pm 0.2}$	$90.7_{\pm 0.9}$	$92.8_{\pm 0.3}$	$88.7_{\pm 1.5}$

Table 1: Comparison of average and worst-group test accuracies for different methods when evaluated on image classification datasets: Waterbirds (Sagawa et al. 2020a) & CelebA (Liu et al. 2015). We obtain the results of CVaR DRO, LfF, EIIL, JTT and SSA from (Nam et al. 2022). Results (mean and std) of our method are estimated over 3 random runs. Best-performing results (in terms of worst-group accuracy) are marked in bold.

Method	Waterbirds			CelebA		
	100%	10% (Val. set)	5%	100%	10% (Val. set)	5%
JTT	86.7	86.9	76.0	81.1	81.1	82.2
SSA	89.0	88.9	87.1	89.8	90.0	86.7
PG-DRO	**91.0**	**90.3**	**89.2**	**90.0**	**90.6**	**89.4**

Table 2: Worst-group accuracy on Waterbirds and CelebA under varying fractions of the group-annotated validation set. Our proposed framework, PG-DRO, outperforms state-of-art methods even under reduced group annotation. Results of JTT (Liu et al. 2021) and SSA (Nam et al. 2022) are from (Nam et al. 2022).

selecting bird photographs from the Caltech-UCSD Birds-200-2011 (CUB) (Wah et al. 2011) dataset and then superimposing on $\mathcal{E} = \{\text{WATER}, \text{LAND}\}$ background selected from the Places dataset (Zhou et al. 2017). The dataset consists of $n = 4795$ training examples, with the smallest group size 56 (*i.e.*, WATERBIRD on LAND background).

Training details. For experimentation on image classification datasets, following prior works (Sagawa et al. 2020a) we use ResNet-50 (He et al. 2016) initialized from ImageNet pre-trained model. Models are selected by maximizing the worst-group accuracy on the validation set. *We provide detailed description regarding hyper-parameter in Appendix B and C.*

Metrics. For all methods, we report two standard metrics: (1) average test accuracy and (2) worst-group test accuracy. In particular, the worst-group test accuracy indicates the model's generalization performance for groups where the correlation between the label y and environment e does not hold. High worst-group test accuracy indicates a model's less reliance on the spurious attribute.

PG-DRO outperforms strong baselines. In Table 1, we show that our approach PG-DRO significantly outperforms all the rivals on vision datasets, measured by the worst-group accuracy. For comparison, we include the best-performed learning algorithms to date: CVaR DRO (Levy et al. 2020), LfF (Nam et al. 2020), EIIL (Creager, Jacobsen, and Zemel 2021), JTT (Liu et al. 2021), and SSA (Nam et al. 2022)—all of which are developed without assuming the availability of group-labeled training data. Similar to ours, these methods utilize a validation set that contains the group labeling information. The comparisons are thus fair given the same amount of information. In addition, we also compare with G-DRO (Sagawa et al. 2020a), which requires the entire training dataset to be labeled with group attribute and hence is significantly more expensive from an annotation perspective.

We highlight two salient observations: **(1)** On CelebA, PG-DRO outperforms G-DRO by 1.3% in terms of worst-group accuracy. The result signifies the advantage of using probabilistic group labeling compared to hard group labeling. Moreover, our method achieves overall better performance than G-DRO while using *significantly less group annotation* (validation set only). **(2)** Among methods that do not use group-labeled training samples, PG-DRO outperforms the current SOTA methods, JTT (Liu et al. 2021) and SSA (Nam et al. 2022), by 4.3% and 2% on Waterbirds dataset respectively. As expected, while ERM consistently achieves the best average accuracy among all methods, its worst-group accuracy suffers the most.

PG-DRO remains competitive under reduced group annotation. In Table 2, we show that the benefits of using group probabilities persist even when the model is trained with reduced group annotation. Specifically, we perform controlled experiments and compare PG-DRO with two latest SOTA methods, SSA (Nam et al. 2022) and JTT (Liu et al. 2021), in settings where the group labels are available on 100%, 10% and 5% of the validation set. We observe that even under reduced group annotation, PG-DRO consistently outperforms both SSA and JTT. In particular, under the challenging setting with only 5% of the validation set, PG-DRO outperforms both SSA and JTT by 2.1% and 13.2% on Waterbirds (Sagawa et al. 2020a) dataset, in terms of worst-group accuracy.

Further, we also include additional qualitative evidence via GradCAM (Selvaraju et al. 2017) visualizations in *Appendix*

E. We observe that for the Waterbirds dataset, PG-DRO consistently focuses on semantic regions representing essential cues for accurately identifying the foreground object such as claw, wing, beak, and fur. In contrast, baseline methods tend to output higher salience for spurious background attribute pixels.

4.2 Evaluation on Natural Language Processing Tasks

Datasets. To validate the effectiveness of PG-DRO, we perform experiments on two natural language processing datasets: MultiNLI and CivilComments-WILDS containing spurious correlations.

(1) MultiNLI (Williams, Nangia, and Bowman 2018): The Multi-Genre Natural Language Inference (MultiNLI) dataset is a crowdsourced collection of sentence pairs with the premise and hypothesis. The label indicates whether the hypothesis is entailed by, contradicts, or is neutral to the premise. Hence, the label space is defined as $\mathcal{Y} = \{\text{ENTAILED}, \text{NEUTRAL}, \text{CONTRADICTORY}\}$. Previous study (Gururangan et al. 2018) has shown the presence of spurious associations between the target label CONTRADICTORY and set of negation words, $\mathcal{E}$=$\{$NOBODY, NO, NEVER, NOTHING$\}$.

(2) CivilComments-WILDS (Borkan et al. 2019; Koh et al. 2021): For this dataset, we use a similar setup as defined in (Nam et al. 2022). Each instance in this dataset corresponds to an online comment generated by users which is labeled as either toxic or not toxic, $\mathcal{Y} = \{\text{TOXIC}, \text{NON-TOXIC}\}$. The spurious attribute is set as $\mathcal{E} = \{\text{IDENTITY}, \text{NO IDENTITY}\}$, where IDENTITY indicates comment associated with any of the demographic identities {male, female, White, Black, LGBTQ, Muslim, Christian, other religion}.

Training details. For experimentation on language processing tasks, we use a pre-trained BERT model (Devlin et al. 2019). Specifically, we use the Hugging Face PyTorch-Transformers (Wolf et al. 2020) implementation of the BERT `bert-base-uncased` model. We use the default tokenizer and model settings. The evaluation metrics are the same as Section 4.1. *Refer Appendix B and C for detailed description regarding hyper-parameters.*

PG-DRO achieves superior performance. In Table 3, we provide a comprehensive comparison against an array of well-known learning algorithms designed to tackle spurious correlations in training data, including CVaR DRO (Levy et al. 2020), LfF (Nam et al. 2020), EIIL (Creager, Jacobsen, and Zemel 2021), JTT (Liu et al. 2021), and SSA (Nam et al. 2022). For fair evaluation, all competing methods utilize group annotations on the validation set. We also compare against standard baselines such as ERM which does not require group labeling, and G-DRO (Sagawa et al. 2020a) which assumes group labeling information for both train and validation set.

Similar to our observations on vision datasets, PG-DRO upholds the trend of outperforming competing baselines achieving superior performance on both natural language processing tasks. In particular, PG-DRO improves the current best method SSA (Nam et al. 2022) by 3.7% and 2.3% on the CivilComments-WILDS and MultiNLI dataset respectively.

5 Further Ablations

Given the strong performance of PG-DRO on both computer vision (Section 4.1) and NLP (Section 4.2) tasks, we take a step further to understand the advantages of using probabilistic group membership over hard group annotations. Specifically, in this section, we design controlled experiments to ablate the role of each component in our proposed framework: pseudo group labeling and robust optimization. In particular, for comprehensive evaluation and comparison, we train the spurious attribute classifier ($\hat{f}_{\text{env}}$) using: (1) supervised approach described in Section 3.2, (2) a new zero-shot approach (see details in Appendix A), and (3) semi-supervised learning (Lee et al. 2013; Nam et al. 2022). The semi-supervised objective additionally requires using group unlabelled training samples, in addition to a small number of labeled ones.

PG-DRO outperforms G-DRO under different pseudo group labeling methods. In Table 4, we provide a comprehensive comparison between PG-DRO and G-DRO for different pseudo group labeling objectives. We observe that for both Waterbirds and CelebA datasets, PG-DRO consistently outperforms G-DRO under different approaches for pseudo group labeling. These results further highlight that the advantages of using probabilistic group membership can be leveraged irrespective of the pseudo group labeling approach used. Moreover, PG-DRO continues its dominance over G-DRO under the most challenging setting, when group annotations are available for only 5% of the validation set. These experiments further underwrite the importance of using probabilistic group membership.

6 Understanding Benefits of PG-DRO

The results in Section 4 and Section 5 validate the improved performance of our proposed framework PG-DRO. In this section, we seek to provide a better understanding for the improved performance.

6.1 Setup

Data distribution. We construct a synthetic dataset to better understand the advantages of using group-probabilities over hard group annotations. Compared to the complex real datasets studied in Section 4, this synthetic data helps us directly visualize and understand each component of PG-DRO, and its influence on the decision boundary. Specifically, the training dataset ($\mathcal{D}_{\text{train}}$) consists of n training samples: $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$, where the label $y = \{-1, 1\}$ is spuriously correlated with the environment attribute $e = \{-1, 1\}$. Replicating the real datasets studied in Section 4, we divide the training data into four groups accordingly: two majority groups with $y = e$, each of size $n_{\text{maj}}/2$, and two minority groups with $y = -e$, each of size $n_{\text{min}}/2$. Further, we set $n = n_{\text{maj}} + n_{\text{min}}$ as the total number of training points, and $p = n_{\text{maj}}/n$ as the fraction of majority examples. A higher value of p indicates a stronger correlation between target label y and environment attribute e.

Method	Dataset with group label	MultiNLI		CivilComments-WILDS	
		Avg. Acc	Worst Group Acc	Avg. Acc	Worst Group Acc
ERM (Vapnik 1991)	None	82.6	66.4	92.6	58.4
CVaR DRO (Levy et al. 2020)	val. set	82.0	68.0	92.5	60.5
LfF (Nam et al. 2020)	val. set	80.8	70.2	92.5	58.8
EIIL (Creager, Jacobsen, and Zemel 2021)	val. set	79.4	70.9	90.5	67.0
JTT (Liu et al. 2021)	val. set	78.6	72.6	91.1	69.3
SSA (Nam et al. 2022)	val. set	79.9	76.6	88.2	69.9
PG-DRO (Ours)	val. set	$81.0_{\pm 0.4}$	$\mathbf{78.9}_{\pm 0.9}$	$89.2_{\pm 1.4}$	$\mathbf{73.6}_{\pm 1.8}$
G-DRO (Sagawa et al. 2020a)	train & val. set	$81.4_{\pm 0.1}$	$77.5_{\pm 1.2}$	$87.7_{\pm 0.6}$	$69.1_{\pm 0.8}$

Table 3: Comparison of average and worst-group test accuracies for different methods when evaluated on language processing datasets: MultiNLI (Williams, Nangia, and Bowman 2018) & CivilComments-WILDS (Borkan et al. 2019; Koh et al. 2021). We obtain the results of CVaR DRO, LfF, EIIL, JTT and SSA from (Nam et al. 2022). Results (mean and std) of our method are estimated over 3 random runs. Best performing results (in terms of worst-group accuracy) are marked in bold.

Pseudo Group Labeling Method	Robust Optimization Method	Dataset with group label	Waterbirds		CelebA	
			Avg. Acc	Worst Group Acc	Avg. Acc	Worst Group
Supervised	G-DRO	val. set	92.4	88.0	91.6	88.6
	PG-DRO (Ours)	val. set	92.5	**91.0**	92.2	**90.0**
Supervised	G-DRO	5% of val. set	92.5	87.5	92.5	87.5
	PG-DRO (Ours)	5% of val. set	91.7	**89.2**	92.0	**89.4**
Semi-Supervised	G-DRO	val. set	92.2	89.0	92.8	**89.8**
	PG-DRO (Ours)	val. set	92.4	**91.0**	92.5	**89.8**
Semi-Supervised	G-DRO	5% of val. set	92.6	87.1	92.8	86.7
	PG-DRO (Ours)	5% of val. set	91.9	**89.2**	91.7	**87.8**
Zero-shot	G-DRO	None	91.2	87.6	93.2	87.8
	PG-DRO (Ours)	None	91.5	**89.4**	92.7	**88.8**

Table 4: Comparison of average and worst-group test accuracies for different combinations of training objectives used during pseudo group labeling and robust optimization. Best-performing results (in terms of worst-group accuracy) are marked in bold.

An input sample $\mathbf{x} = [z_{\text{inv}}, z_{\text{e}}] \in \mathbb{R}^2$ comprises of invariant features $z_{\text{inv}} \in \mathbb{R}$ generated from target label y, and environmental/spurious features $z_{\text{e}} \in \mathbb{R}$ generated from environment e. In particular, we generate:

$$z_{\text{inv}} \sim \mathcal{N}(y, \sigma^2_{\text{inv}})$$
$$z_{\text{e}} \sim \mathcal{N}(e, \sigma^2_{\text{e}}),$$

The randomness in the invariant and environmental features are controlled through their respective variances. For all experiments in Section 6.2, we fix the total number of training points $n = 4000$ and majority fraction $p = 0.95$. Further, σ^2_{inv} and σ^2_{e} are set as 0.5 and 0.05 respectively to encourage the model to use the environmental features over the invariant features.

Model. Both for the purpose of pseudo group labeling and robust optimization, we use a simple four layered neural network with ReLU activation. We provide further details of model training in Appendix.

6.2 Insights

PG-DRO leads to a more robust decision boundary than G-DRO. In this study, we train a group predictor on a small group-labeled dataset ($\mathcal{D}_{\text{L}}$) consisting of 100 samples.

Next, the trained group classifier is used to generate both hard group labels and group probabilities for every sample $\mathbf{x} \in \mathcal{D}_{\text{train}}$ for further training with G-DRO and PG-DRO respectively. In case of G-DRO, $\hat{\mathbb{Q}}(\mathbf{x})$ has all its mass in the predicted group. In Figure 2, we visualize the decision boundary plot of models trained using G-DRO vs. PG-DRO. The vertical axis in the plots indicate the environmental/spurious feature (z_{e}) that highly correlates with y, and the horizontal axis is the invariant feature (z_{inv}). We observe that the PG-DRO trained model (Figure 2 (Right)) learns a more robust decision boundary (more dependent on invariant features) as compared to the model trained using G-DRO (Figure 2 (Left)). Moreover, the improved worst-group test accuracy of using PG-DRO further validates the advantages of using group probabilities.

PG-DRO better captures uncertainty in spurious attribute predictor. In Figure 3, we contrast the spurious attribute's prediction under soft (left) vs. hard (right) encoding. We can observe that the classifier $\hat{f}_{\text{env}}$ can be uncertain on samples belonging to the minority groups, where spurious correlations do not hold. Hard labeling, in this scenario, does not consider the ambiguity in the predictions of the group

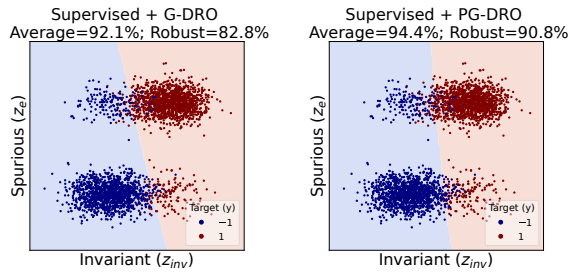

Figure 2: Consider data points $\mathbf{x}$ in $\mathbb{R}^2$ with two classes y. The vertical axis of $\mathbf{x}$ is the environment/spurious feature that highly correlates with y, and the horizontal axis is the invariant feature. The data consists of four groups, where the *top-left and lower-right are two minority groups*. The robust accuracy is test worst-group accuracy. Model trained using PG-DRO (Right) is more robust as compared to the G-DRO (Left). Color of points encodes ground truth label y. Color of background shade indicates model's prediction $\hat{y}$.

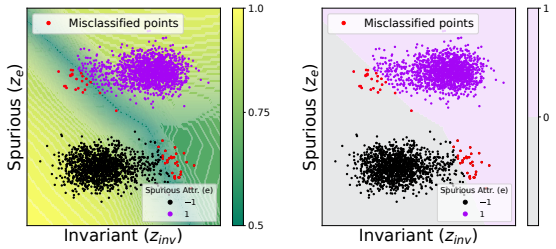

Figure 3: Consider data points $\mathbf{x} \in \mathcal{D}_{\text{train}}$. Color of points encodes the true spurious attribute e. (Left) Confidence plot of the trained spurious attribute classifier. We visualize the maximum confidence ($\max_i \hat{\mathbb{P}}(e_i|x)$) of the spurious predictor. (Right) Decision boundary based on hard labeling. Color of background shade indicates the model's predicted environment ($\arg\max_i \hat{\mathbb{P}}(e_i|x)$). Mis-classified points are marked in red, which are erroneously penalized with hard labeling. Using probabilistic group better captures uncertainty in spurious attribute prediction.

classifier. As a result, the model can disproportionally penalize the samples, leading to poor worst-group robustness (See Figure 2 (Left)). Using probabilistic group membership is more advantageous since it allows capturing group ambiguity and uncertainty in the spurious predictor.

7 Related Works

Distributionally robust optimization. A line of works (Ben-Tal et al. 2013; Wiesemann, Kuhn, and Sim 2014; Blanchet and Murthy 2019; Lam and Zhou 2015; Namkoong and Duchi 2017; Mohajerin Esfahani and Kuhn 2018; Bertsimas, Gupta, and Kallus 2018) proposed algorithms to minimize the worst loss within a ball centered around the empirical distribution over training data. Recent works (Sagawa et al. 2020a; Zhang et al. 2022) have raised concerns regarding these algorithms, as they tend to optimize for both the worst-group and average accuracy during training. Specifically, Sagawa *et al.* (Sagawa et al. 2020a) have shown that in the regime of over-parameterized models, DRO performs no better than empirical risk minimization.

Improving robustness with group annotations. When the group annotations are available for training data, one can leverage the information to alleviate the model's reliance on spurious correlations. Among this line of research, Sagawa et al. (2020a), Zhang et al. (2020), Mohri, Sivek, and Suresh (2019) have proposed designing objective functions to minimize the worst group training loss. In this study, we specifically constrain our attention to Group-DRO (G-DRO), an online optimization algorithm proposed by Sagawa et al. (2020a), that focuses on training updates over data points from higher-loss groups. Goel et al. (2021) proposed CAMEL, which trains a CycleGAN (Zhu et al. 2017) model to learn data transformations and then apply it to extend the minority group through data augmentation synthetically. A group of methods (Shimodaira 2000; Byrd and Lipton 2019; Sagawa et al. 2020b) aims to improve the worst group performance through re-weighting or re-sampling the given dataset to balance the distribution of each group. Although these methods show promising results and improvement in worst-group performance over standard empirical risk minimization (ERM), there is a major caveat: all these methods assume the availability of group annotations for samples in the entire dataset—which can be expensive and prohibitive in practice.

Improving robustness without group annotations. Several recent works tackle the problem of reducing the dependency on group annotations for the entire training dataset. In particular, Nam et al. (2020) proposed to train a pair of models simultaneously ("biased" and "debiased" versions), such that their relative cross-entropy losses on each training example determine their importance weights in the overall training objective. Liu et al. (2021) first train an ERM model for a fixed number of epochs to identify samples being misclassified, and then train another model by up-weighting the misclassified samples. Similarly, EIIL (Creager, Jacobsen, and Zemel 2021) and SSA (Nam et al. 2022) first train a model to infer the group labels, and then use the generated group labels for robust training with G-DRO (Sagawa et al. 2020a). Another line of work includes re-weighting training samples based on weights determined through meta-learning (Ren et al. 2018) or by learning explicitly using a small amount of group labeled samples (Shu et al. 2019). Our proposed approach is more related to methods belonging to this branch of study (which do not require group labels for the training set).

8 Conclusion

In this paper, we propose PG-DRO, a novel robust learning framework targeting the spurious correlation problem. Our framework explores the idea of probabilistic group membership for distributionally robust optimization. We broadly evaluate PG-DRO on both computer vision and NLP benchmarks, establishing superior performance among methods not using training set group annotations. With minimal group annotations, PG-DRO can favorably match and even outperform G-DRO (using group annotations for both training and validation set).

Acknowledgements

We would like to thank Yifei Ming, Ziyang (Jack) Cai and Gabriel Gozum for insightful discussions and comments. We gratefully acknowledge the support of the AFOSR Young Investigator Award under No. FA9550-23-1-0184; Philanthropic Fund from SFF; Wisconsin Alumni Research Foundation; faculty research awards from Google, Meta, and Amazon. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements either expressed or implied, of the sponsors.

References

Ben-Tal, A.; Den Hertog, D.; De Waegenaere, A.; Melenberg, B.; and Rennen, G. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 341–357.

Bertsimas, D.; Gupta, V.; and Kallus, N. 2018. Data-driven robust optimization. *Mathematical Programming*, 235–292.

Blanchet, J.; and Murthy, K. 2019. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 565–600.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 491–500.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Byrd, J.; and Lipton, Z. 2019. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, 872–881. PMLR.

Creager, E.; Jacobsen, J.-H.; and Zemel, R. 2021. Environment inference for invariant learning. In *International Conference on Machine Learning*, 2189–2200. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

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

Goel, K.; Gu, A.; Li, Y.; and Ré, C. 2021. Model Patching: Closing the Subgroup Performance Gap with Data Augmentation. In *International Conference on Learning Representations*.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 5637–5664. PMLR.

Lam, H.; and Zhou, E. 2015. Quantifying uncertainty in sample average approximation. In *2015 Winter Simulation Conference (WSC)*, 3846–3857. IEEE.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 896.

Levy, D.; Carmon, Y.; Duchi, J. C.; and Sidford, A. 2020. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*.

Liu, E. Z.; Haghgoo, B.; Chen, A. S.; Raghunathan, A.; Koh, P. W.; Sagawa, S.; Liang, P.; and Finn, C. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 6781–6792. PMLR.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3730–3738.

Mohajerin Esfahani, P.; and Kuhn, D. 2018. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 115–166.

Mohri, M.; Sivek, G.; and Suresh, A. T. 2019. Agnostic federated learning. In *International Conference on Machine Learning*, 4615–4625. PMLR.

Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*.

Nam, J.; Kim, J.; Lee, J.; and Shin, J. 2022. Spread Spurious Attribute: Improving Worst-group Accuracy with Spurious Attribute Estimation. In *International Conference on Learning Representations*.

Namkoong, H.; and Duchi, J. C. 2017. Variance-based regularization with convex objectives. *Advances in neural information processing systems*.

Oakden-Rayner, L.; Dunnmon, J.; Carneiro, G.; and Ré, C. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, 151–159.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, 4334–4343. PMLR.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020a. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*.

Sagawa, S.; Raghunathan, A.; Koh, P. W.; and Liang, P. 2020b. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, 8346–8356. PMLR.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 227–244.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*.

Sohoni, N.; Dunnmon, J.; Angus, G.; Gu, A.; and Ré, C. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*.

Tu, L.; Lalwani, G.; Gella, S.; and He, H. 2020. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. *Transactions of the Association for Computational Linguistics*, 621–633.

Vapnik, V. 1991. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wiesemann, W.; Kuhn, D.; and Sim, M. 2014. Distributionally robust convex optimization. *Operations Research*, 1358–1376.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Zhang, J.; Menon, A.; Veit, A.; Bhojanapalli, S.; Kumar, S.; and Sra, S. 2020. Coping with label shift via distributionally robust optimisation. *arXiv preprint arXiv:2010.12230*.

Zhang, M.; Sohoni, N. S.; Zhang, H. R.; Finn, C.; and Ré, C. 2022. Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations. *arXiv preprint arXiv:2203.01517*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 1452–1464.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.