Outlier-Robust Group Inference via Gradient Space Clustering

Yuchen Zeng[1]* **Kristjan Greenewald**[2] **Luann Jung**[3] **Kangwook Lee**[1]

Justin Solomon[3] **Mikhail Yurochkin**[2]

[1]UW-Madison, [2]MIT-IBM Watson AI Lab, [3]MIT

Abstract

Traditional machine learning models focus on achieving good performance on the overall training distribution, but they often underperform on minority groups. Existing methods can improve the worst-group performance, but they can have several limitations: (i) they require group annotations, which are often expensive and sometimes infeasible to obtain, and/or (ii) they are sensitive to outliers. Most related works fail to solve these two issues simultaneously as they focus on conflicting perspectives of minority groups and outliers. We address the problem of learning group annotations in the presence of outliers by clustering the data in the space of gradients of classification loss w.r.t. the model parameters. We show that data in the gradient space has a simpler structure while preserving information about minority groups and outliers, making it suitable for standard clustering methods like DBSCAN. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art both in terms of group identification and downstream worst-group performance.

1 Introduction

Empirical Risk Minimization (ERM), i.e., the minimization of average training loss over the set of model parameters, is the standard training procedure in machine learning. It yields models with strong in-distribution performance[2] but does not guarantee satisfactory performance on groups with fewer training samples (*minority groups*) that contribute relatively few data points to the training loss function (Sagawa et al., 2019; Koh et al., 2021). For instance, consider Waterbirds dataset (Sagawa et al., 2019), where the task is to predict the bird type in a given image featuring a bird on either a water or land background. Note that groups with matched bird types and backgrounds have significantly more samples than groups with mismatched bird types and backgrounds (see Table 1). Such data can lead to problematic and unfair spurious correlations between the background and bird types. Consequently, ERM learns to associate bird types with backgrounds (Creager et al., 2021; Sagawa et al., 2019), resulting in low accuracy on minority groups with mismatched backgrounds and bird types. A related phenomenon is *subpopulation shift* (Koh et al., 2021), i.e., when the test distribution differs from the train distribution in terms of group

# Samples (group-wise accuracy)	Waterbird	Landbird
Water Background	1057 (93.98%)	184 (82.19%)
Land Background	56 (54.89%)	3498 (98.50%)

Table 1: **Sample distribution and group-wise accuracy in Waterbirds dataset, with groups defined by background and bird type.** Groups in which bird types and backgrounds align, comprising most of the training samples, are categorized as *majority groups*, while those with mismatched bird types and backgrounds, containing only a small fraction of samples, are referred to as *minority groups*. The numbers in parentheses indicate group-wise ERM model accuracy.

*Work performed while doing an internship at IBM research.

[2] I.e. low loss on test data drawn from the same distribution as the training dataset.

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

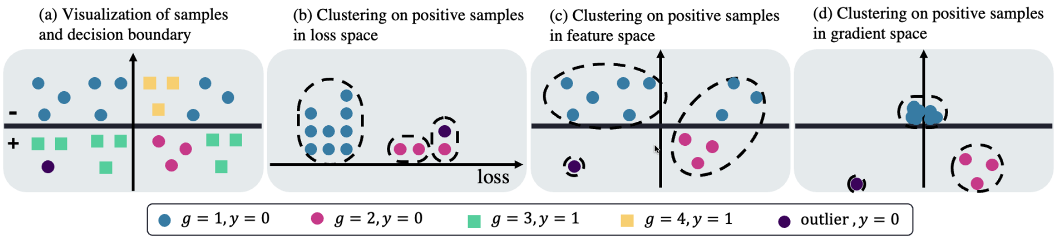


Figure 1: **An illustration of learning group annotations in the presence of outliers.** (a) A toy dataset in two dimensions. There are four groups $g = 1, 2, 3, 4$ and an outlier. $g = 1$ and $g = 3$ are the majority groups distributed as mixtures of three components each; $g = 2$ and $g = 4$ are unimodal minority groups. y-axis is the decision boundary of a logistic regression classifier. Figures (b, c, d) compare different data views for learning group annotations and detecting outliers via clustering of samples with $y = 0$. (b) loss values can confuse outliers and minority samples which both can have high loss; (c) in the original feature space it is difficult to distinguish one of the majority group modes and the minority group; (d) gradient space (bias gradient omitted for visualization) simplifies the data structure making it easier to identify the minority group and to detect outliers.

proportions. Under subpopulation shift, poor performance on the minority groups in the train data translates into poor overall test distribution performance, where these groups are more prevalent or more heavily weighted. Subpopulation shift occurs in many application domains (Tatman, 2017; Beery et al., 2018; Oakden-Rayner et al., 2020; Santurkar et al., 2020; Koh et al., 2021). This effect becomes particularly problematic when minority groups correspond to socially-protected groups, potentially leading to discrimination (Dixon et al., 2018; Garg et al., 2019; Yurochkin & Sun, 2020).

1.1 Preliminaries & Related Works

Denote $[N] = \{1, \ldots, N\}$. Consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^n$ consisting of n samples, where $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$ is the input feature and $\mathrm{y} \in \mathcal{Y} = [C]$ is the class label. The samples are categorized into G disjoint groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\} \triangleq P$, which may also take into account class label. Denote the group membership of each point in the dataset as $\{\mathrm{g}_i\}_{i=1}^n$, where $\mathrm{g}_i \in [G]$ for all $i \in [n]$. For example, in toxicity classification, a group could correspond to a toxic comment mentioning a specific identity, or, in image recognition, a group could be an animal species appearing on an atypical background (Beery et al., 2018; Sagawa et al., 2019).

The goal of learning in the presence of minority groups is to learn a model $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$ parameterized by $\boldsymbol{\theta} \in \Theta$ that performs well on all groups $\mathcal{G}_g$, where $g \in [G]$.

Group-aware setting. Many prior works study the problem of learning in the presence of minority groups assuming assume access to the group annotations. Among the state-of-the-art methods in this setting is group Distributionally Robust Optimization (gDRO) (Sagawa et al., 2019). Let $\ell : \mathcal{Y} \times \mathcal{Y}$ be a loss function. The optimization problem of gDRO is

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{g \in [G]} \frac{1}{|\mathcal{G}_g|} \sum_{(\mathbf{x}, \mathrm{y}) \in \mathcal{G}_g} \ell(\mathrm{y}, h_{\boldsymbol{\theta}}(\mathbf{x})), \tag{gDRO}$$

which aims to minimize the maximum group loss.

Group-oblivious setting. In contrast to the group-aware setting, the *group-oblivious* setting attempts to improve worst-group performance without group annotations. Methods in this group rely on various forms of DRO (Hashimoto et al., 2018; Zhai et al., 2021) or adversarial reweighing (Lahoti et al., 2020). Algorithmically, this results in up/down-weighing the contribution of the high/low-loss points. For example, Hashimoto et al. (2018) optimizes a DRO objective with respect to a chi-square divergence ball around the data distribution, which is equivalent to an ERM discounting low-loss points by a constant depending on the ball radius. Unfortunately, such methods are at risk of overfitting to *outliers*.

We define outliers as data points that significantly deviate from other observations in either the feature x (e.g. corrupted images) or label y (i.e. mislabeled data). Note that outliers differ from minority groups in that minority groups must include multiple samples that are similar to each other, while outliers are inherently isolated and unpredictable. These outliers will generally be high-loss points, indeed, existing methods for outlier-robust training propose to *ignore* the high-loss points (Shen & Sanghavi, 2019), the opposite of the approach in Hashimoto et al. (2018); Liu et al. (2021). Since outliers of this form are ubiquitous in real-world datasets, it is crucial to resolve this dilemma.

Table 2: **Summary of methods for learning in the presence of minority groups.** "-" indicates that there is no clear evidence in the prior works.

Setting		Group-aware	Group-oblivious		Group-learning			
Method	ERM	gDRO (Sagawa et al., 2019)	χ^2-DRO (Hashimoto et al., 2018)	DORO (Zhai et al., 2021)	JTT (Liu et al., 2021)	EIIL (Creager et al., 2021)	George (Sohoni et al., 2020)	GRASP (Ours)
Improves worst-group performance?	✗	✓	✓	✓	✓	✓	✓	✓
No training group annotations?	✓	✗	✓	✓	✓	✓	✓	✓
No validation group annotations?	✓	✗	✗	✗	✗	✗	✓	✓
Group inference?	✗	✗	✗	✗	✓	✓	✓	✓
Robust to outliers?	✗	-	✗	✓	✗	-	-	✓

Algorithm 1: GRASP

```
Input   : DBSCAN hyperparameters $\epsilon$ and $m$
Train the ERM classifier $\boldsymbol{\theta}_0 \leftarrow \arg\min_{\boldsymbol{\theta}\in\Theta'} \sum_{(\mathbf{x},\mathrm{y})\in\mathcal{D}} \ell(\mathrm{y}, h_{\boldsymbol{\theta}}(\boldsymbol{x}))$ ;
for $(\mathbf{x}, \mathrm{y}) \in \mathcal{D}$ do
    Compute its gradient $\boldsymbol{f} \leftarrow \frac{\partial \ell(\mathrm{y}, h_{\boldsymbol{\theta}}(\mathbf{x}))}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$;
for $y \in \mathcal{Y}$ do
    Consider all samples $\{(\mathbf{x}_i, \mathrm{y}_i)\} \subset \mathcal{D}$ with $\mathrm{y} = y$ and their corresponding gradients $\{\boldsymbol{f}_i\}$;
    Compute the distance matrix $D$, where $D_{ij}$ is the distance between $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$;
    Assign group annotations and identify outliers by performing DBSCAN clustering in
      gradient space: $\{\hat{\mathrm{g}}_i\} \leftarrow \mathrm{DBSCAN}(D, \epsilon, m)$, where $\hat{\mathrm{g}}_i = -1$ indicates outliers;
Output : Dataset with predicted group annotations $\mathcal{D}' \leftarrow \{(\mathbf{x}, \hat{\mathrm{g}}, \mathrm{y})\}_{\{\hat{\mathrm{g}}\neq -1, (\mathbf{x},\mathrm{y})\in\mathcal{D}\}}$, where the
         detected outliers are removed
```

Group-learning setting. The final category corresponds to a two-step procedure, wherein the data points are first assigned group annotations based on various criteria, followed by group-aware training typically using gDRO. In this category, Just Train Twice (JTT) (Liu et al., 2021) trains an ERM model and designates high-loss points as the minority and low-loss points as the majority group; George (Sohoni et al., 2020) seeks to cluster the data to identify groups with a combination of dimensionality reduction, overclustering, and augmenting features with loss values, and Environment Inference for Invariant Learning (EIIL) (Creager et al., 2021) finds group partition that maximizes the Invariant Risk criterion (Arjovsky et al., 2019).

2 GRASP: Gradient Space Partitioning

Our method, Gradient Space Partitioning (GraSP), belongs to the group-learning category.[3] However, GraSP differs from prior group-learning works in its ability to account for outliers in the data. In addition, most prior methods in this and the group-oblivious categories typically require validation data with *true* group annotations for model selection to achieve meaningful worst-group performance improvements over ERM, while GraSP does not need any group annotations to achieve good performance. Table 2 summarizes properties of relevant methods in each setting.

We propose to represent data using gradients of a datum's loss w.r.t. the model parameters. Such gradients tell us how a specific data point wants the parameters of the model to change to fit it better. In this gradient space, we anticipate groups (conditioned on label) to correspond to gradients forming clusters. Outliers, on the other hand, majorly correspond to isolated gradients: they are likely to want model parameters to change differently from any of the groups *and* other outliers. See Figure 1 for an illustration. The gradient space structure allows us to separate out the outliers and learn the group annotations via traditional clustering techniques such as DBSCAN (Ester et al., 1996). We use learned group annotations to train models with improved worst-group performance (measured w.r.t. the true group annotations).

Gradient Space vs Feature Space. We consider a logistic regression model to better understand how gradient space simplifies data structure and aids clustering.

Consider a binary classification problem ($\mathrm{y} \in \{0, 1\}$) and logistic regression model $\mathbb{P}(\mathrm{y} = 1|\mathbf{x}) = \sigma(\boldsymbol{w}^\top \mathbf{x} + b)$ trained on the given dataset $\mathcal{D}$, where $\sigma(\cdot)$ denotes the sigmoid function, $\boldsymbol{w}$ are the

[3] We present a taxonomy of distributionally-robust methods in Appendix Fig. 2 to clarify our problem setting.

coefficients and b is the bias. Recall that the logistic regression loss is defined as,

$$\ell(\mathrm{y}, \sigma(\boldsymbol{w}^\top \mathbf{x} + b)) = -\mathrm{y} \log(\sigma(\boldsymbol{w}^\top \mathbf{x} + b)) - (1 - \mathrm{y}) \log(1 - \sigma(\boldsymbol{w}^\top \mathbf{x} + b)),$$

and the gradient of this loss at point $(\mathbf{x}, \mathrm{y})$ w.r.t. $(\boldsymbol{w}, b)$ is

$$\boldsymbol{f} =: \nabla_{[\boldsymbol{w},b]} \ell(\mathrm{y}, \boldsymbol{w}^\top \mathbf{x} + b) = (\sigma(\boldsymbol{w}^\top \mathbf{x} + b) - \mathrm{y}) \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}. \tag{1}$$

This gradient is simply a scaling of the data vector $\mathbf{x}$ by the error $(\sigma(\boldsymbol{w}^\top \mathbf{x} + b) - \mathrm{y}) \in [-1, 1]$, padded by an additional element (the bias entry) consisting of the error alone. When $(\mathbf{x}, \mathrm{y})$ is correctly classified, the scaling is close to zero and when it is incorrectly classified, the scaling approaches 1.

We interpret this gradient (1) through the lens of Euclidean distance ($\|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2$). Recall that we apply clustering to each class independently. The scaling effect mentioned in the previous paragraph shrinks the correctly classified points towards the origin, while leaving the misclassified points almost unaffected (see Figure 1(d)). The error itself is included as an extra element (using loss as an additional feature was previously considered as a heuristic in feature clustering for learning group annotations (Sohoni et al., 2020)). Consequently, gradient clustering w.r.t. Euclidean distance should cluster the correctly classified samples into one "majority" group, and then divide the remaining points into minority groups and outliers based on the size of the error and their position in the feature space. In Appendix D.1, we additionally provide an interpretation of our gradient representation when centered cosine distance is employed.

Meanwhile, it is worth noting that the scaling term of (1), the error, is not specific to logistic regression but applies to all models due to the chain rule. Thus, the intuition behind gradient space clustering based on logistic regression model can be extended to non-linear models.

GraSP. We present the pseudocode for GRASP in Algorithm 1. We first train an ERM classifier $h_{\boldsymbol{\theta}}(\cdot)$ and collect the gradients of samples' losses w.r.t. model parameters $\boldsymbol{\theta}$. We then compute the pairwise centered cosine distances within each class $y \in \mathcal{Y}$ using gradient representations, as discussed in Sec. C.1. Lastly, to estimate the group annotations and identify outliers, we apply DBSCAN on these distance matrices for each class $y \in \mathcal{Y}$.

We discard the identified outliers and provide learned group annotations as inputs to a Group-aware method of choice. For concreteness, in this work, we solve (gDRO) with the method of Sagawa et al. (2019), a stochastic optimization algorithm equipped with convergence guarantees. Note that other choices could be appropriate. For example, methods accounting for noise in group annotations (Lamy et al., 2019; Mozannar et al., 2020; Celis et al., 2021) are also useful as they could counteract mistakes in GRASP annotations.

3 Experiments

In this section, we conduct experiments on synthetic and benchmark datasets (see Sec. E.1.1) to evaluate the performance of GRASP,[4] finding that GRASP outperforms state-of-the-art baselines in group identification quality and downstream worst-group performance while providing robustness to outliers.

3.1 Evaluation of GRASP

We assess the performance of GRASP in terms of group identification and downstream tasks of training models with comparable performance across groups, both with and without outliers. In all experiments, we consider true group annotations unknown in both train and validation data (except for "oracle" gDRO which has access to true group annotations in both train and validation data). We note that this setting is stricter

Table 3: **Group identification performance of group-learning methods measured by Adjusted Rand Index (ARI).** Higher ARI indicates higher group identification quality. GRASP significantly outperforms the group-learning baselines on all the tested datasets. Moreover, we observe that GRASP is robust to outliers.

	SYNTHETIC		WATERBIRDS		COMPAS	CIVIL-COMMENTS
OUTLIERS?	✗	✓	✗	✓		
EIIL	-.0069	-.0043	.0114	.0078	-.0025	-.0001
GEORGE	.6027	.4565	.2832	.2600	.1962	.1422
FEASP	.5133	.4946	.0418	.0418	.2956	.2093
GRASP (OURS)	**.6943**	**.6944**	**.7453**	**.7453**	**.5453**	**.2639**

[4] Our code is available at `https://github.com/yzeng58/private_demographics`.

Table 4: **Downstream worst-group accuracy and average accuracy on the test data.** The average test accuracy is a re-weighted average of the group-specific accuracies, where the weights are based on the training distribution. The results are reported on clean and contaminated versions of Synthetic and Waterbirds datasets (Sagawa et al., 2019; Wah et al., 2011), COMPAS (ProPublica, 2021) and CivilComments (Borkan et al., 2019) datasets. We observe that GRASP significantly outperforms the group-oblivious (DORO) and other group-learning approaches (EIIL, George, FeaSP) methods on Synthetic, COMPAS, and CivilComments datasets, and performs relatively well on Waterbirds datasets, while being robust to outliers.

	SYNTHETIC		WATERBIRDS		COMPAS	CIVILCOMMENTS
OUTLIERS?	✗	✓	✗	✓		
METHOD	WORST.(AVG.)	WORST.(AVG.)	WORST.(AVG.)	WORST.(AVG.)	WORST.(AVG.)	WORST.(AVG.)
ERM	.6667(.8823)	.5333(.8273)	.6075(.9673)	.5249(.9621)	.4706(.6792)	.4659(.9213)
DORO	.6667(.8823)	.6000(.8342)	.5888(.9694)	.6636(.9686)	.4706(.6801)	.4905(.9182)
EIIL	.6667(.8783)	.6000 (.8115)	.6916 (.9645)	**.7056(.9629)**	.0588 (.6046)	.6056(.9066)
GEORGE	.5333(.8732)	.6000(.8342)	**.7523**(.9612)	.5897(.9100)	.4416(.6232)	.5897(.9100)
FEASP	.6667(.8823)	.6667(.8823)	.1417(.9346)	.1417(.9346)	.4416(.6232)	.6056(.9066)
GRASP (OURS)	**.8000**(.8926)	**.8000**(.8926)	.6854(.9654)	.6798(.9004)	**.4743**(.6717)	**.6798(.9004)**
gDRO (ORACLE)	.7333(.8639)	.8000(.8755)	.8665(.9272)	.8545(.9081)	.4625(.6807)	.6941(.8767)

than the majority of prior works considering unknown group annotations (see Table 2). For example, inspecting Table 5 in Appendix B.2 of Zhai et al. (2021), we observe that DORO cannot surpass ERM without access to validation data containing true group annotations (see non-oracle model selection results).

Group annotations quality. The first experiment examines the quality of group annotations learned with GRASP. To collect the gradients of the data's losses w.r.t. the model parameters, we train a logistic regression model on the Synthetic dataset, a three-layer ReLU neural network with 50 hidden neurons per layer on the COMPAS dataset, and a BERT (Devlin et al., 2018) model on the CivilComments dataset (due to the large number of parameters in BERT, we only consider the last transformer and the subsequent prediction layer when extracting gradients). For the Waterbirds dataset, we first featurize the images using a ResNet50 pre-trained on ImageNet (Deng et al., 2009), and then train a logistic regression. We then use DBSCAN clustering with centered cosine distance. We select DBSCAN hyperparameters using standard clustering metrics that do not require knowledge of the true group annotations, see Appendix E.1.

In Table 3, we compare group identification quality of GRASP (measured with ARI) to three group learning baselines, EIIL, George, and FeaSP, across four datasets. There are two key observations supporting the claims made in this paper: (i) clustering in the gradient space (GRASP) outperforms clustering in the feature space (FeaSP and George), as well as other baselines (EIIL); (ii) GRASP is robust to outliers, i.e. it performs equally well in the presence and absence of outliers. To comment on the low ARI of EIIL, we note that the Invariant Risk criteria EIIL optimizes was designed primarily for invariant learning (i.e., learning environment labels) (Arjovsky et al., 2019; Creager et al., 2021), which may not be suitable for learning group annotations.

Worst-group performance. The standard metric when comparing methods for training ML models with comparable performance across groups (evaluated w.r.t. true group annotations) is worst-group accuracy (Sagawa et al., 2019; Koh et al., 2021). For the group-learning methods (GRASP, FeaSP, George, EIIL), we first discard identified outliers if applicable (GRASP and FeaSP), and then train gDRO with the corresponding learned group annotations. We also use the learned group annotations on the validation data to select the corresponding gDRO hyperparameters. In Appendix E.2.2 we demonstrate that GRASP worst-group performance is fairly robust to the corresponding DBSCAN hyperparameters. For ERM and DORO we used the validation set overall performance for hyperparameter selection.

For all methods, on a given dataset, we train models with the same architecture and initialization. Recall that these models can be different from the models used in estimating group annotations with any of the group-learning methods. See Appendix E.1 for details.

We summarize results in Table 4. GRASP outperforms baselines on Synthetic, COMPAS, and CivilComments datasets. For the Waterbirds dataset, GRASP also performs relatively well. Interestingly, EIIL performs best on the contaminated Waterbirds dataset, despite the poor ARI discussed earlier. It is, however, failing on the COMPAS dataset. We also notice that GRASP outperforms "oracle" gDRO on Synthetic and COMPAS datasets. This could be due to the fact that gradient space clustering helps to focus on "harder" instances, as discussed in Section C.1, while the available ("oracle") group annotations (at least on COMPAS), might be noisy.

References

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

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*, pp. 1349–1361. PMLR, 2021.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pp. 226–231, 1996.

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., and Beutel, A. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226, 2019.

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.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hodge, V. and Austin, J. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

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. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Kwon, G., Prabhushankar, M., Temel, D., and AlRegib, G. Backpropagated gradient representations for anomaly detection. In *European Conference on Computer Vision*, pp. 206–226. Springer, 2020a.

Kwon, G., Prabhushankar, M., Temel, D., and AlRegib, G. Novelty detection through model-based characterization of neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 3179–3183. IEEE, 2020b.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems*, 33:728–740, 2020.

Lamy, A., Zhong, Z., Menon, A. K., and Verma, N. Noise-tolerant fair classification. *Advances in Neural Information Processing Systems*, 32, 2019.

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 *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Mirzasoleiman, B., Cao, K., and Leskovec, J. Coresets for robust training of deep neural networks against noisy labels. *Advances in Neural Information Processing Systems*, 33:11465–11477, 2020.

Mozannar, H., Ohannessian, M., and Srebro, N. Fair learning with private demographic data. In *International Conference on Machine Learning*, pp. 7066–7075. PMLR, 2020.

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

ProPublica. Compas recidivism risk score data and analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis, 2021.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Santurkar, S., Tsipras, D., and Madry, A. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

Shen, Y. and Sanghavi, S. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pp. 5739–5748. PMLR, 2019.

Singh, K. and Upadhyaya, S. Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, 9(1):307, 2012.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Tatman, R. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pp. 53–59, 2017.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

Wang, H., Bah, M. J., and Hammad, M. Progress in outlier detection techniques: A survey. *IEEE Access*, 7:107964–108000, 2019.

Yurochkin, M. and Sun, Y. Sensei: Sensitive set invariance for enforcing individual fairness. In *International Conference on Learning Representations*, 2020.

Zhai, R., Dan, C., Kolter, Z., and Ravikumar, P. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pp. 12345–12355. PMLR, 2021.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464, 2017.

Appendix

A Background

In this section, we provide a taxonomy of the distributionally robust learning to better clarify our problem setting (see Fig. 2), the details of the Adjusted Rand Index (ARI), and describe the complete algorithm of DBSCAN.

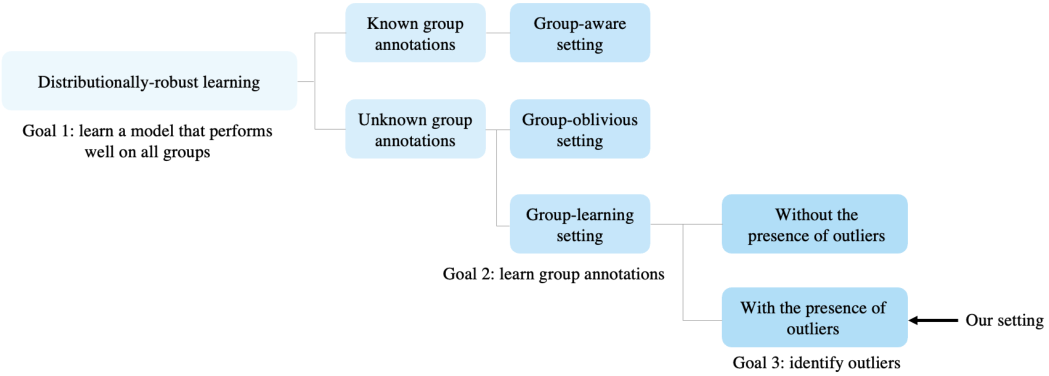

Distributionally-robust learning
Goal 1: learn a model that performs well on all groups
Known group annotations
Group-aware setting
Unknown group annotations
Group-oblivious setting
Group-learning setting
Goal 2: learn group annotations
Without the presence of outliers
With the presence of outliers
Our setting
Goal 3: identify outliers

Figure 2: A taxonomy of distributionally-robust learning illustrating our problem setting.

Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) The Adjusted Rand Index (ARI) is a measure of the degree of agreement between two data partitions and accounts for the chance grouping of elements in the data sets. In our case, consider true group partition P and estimated group partition $\hat{P}$. ARI can be computed by

$$ARI(P,\hat{P}) = \frac{\sum_{g,g'} \binom{n_{gg'}}{2} - \left[\sum_g \binom{n_g}{2} \sum'_g \binom{n_{g'}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_g \binom{n_g}{2} + \sum'_g \binom{n_{g'}}{2}\right] - \left[\sum_g \binom{n_g}{2} \sum'_g \binom{n'_g}{2}\right] / \binom{n}{2}},$$

where $n_{gg'}$ is the number of data points belonging to $\mathcal{G}_g \in P$ assigned to group $\hat{\mathcal{G}}_{g'} \in \hat{P}$, $n_g = |\mathcal{G}_g|$, $n_{g'} = |\hat{\mathcal{G}}_{g'}|$, and n is the total number of samples in the dataset.

DBSCAN (Ester et al., 1996) DBSCAN is a clustering and outlier-detecting method that does not require the number of clusters to be known. It operates on a distance matrix D. We call a sample as a "core sample" if there exist m other samples within a distance of ϵ from this sample. DBSCAN starts with a single cluster that contains an arbitrary core sample and adds core samples from the neighborhood of the cluster to the cluster until all core samples in the ϵ-neighborhood of the cluster have been visited. It then adds the remaining samples in the ϵ-neighborhood of the cluster to the cluster. Next, DBSCAN creates another cluster and expands that cluster by finding unvisited core samples. It then repeats this process of creating and expanding clusters until all core samples have been visited. Any remaining samples that are not added to a cluster are considered outliers. Note that DBSCAN clustering requires two hyperparameters (ϵ, m) and a distance matrix D as input.

B Extended Related Works

The challenge of outliers. Outliers, e.g., mislabeled samples or corrupted images, are ubiquitous in applications (Singh & Upadhyaya, 2012), and outlier detection has long been a topic of inquiry in ML (Hodge & Austin, 2004; Wang et al., 2019). For handling outliers without the presence of minority groups, recent works have proposed ignoring samples with high losses during training (Shen & Sanghavi, 2019) or using low-rank approximations of the Jacobian matrix of the model output to extract a clean subset of the data (Mirzasoleiman et al., 2020). However, outliers are much more challenging to detect when data has (unknown) minority groups, which could be hard to distinguish from outliers but require the opposite treatment: minority groups need to be upweighted while outliers must be discarded. Hashimoto et al. (2018) writes, “it is an open question whether it is possible to design algorithms which are both fair to unknown latent groups and robust [to outliers].”

We provide an illustration of a dataset with minority groups and an outlier in Figure 1(a). Figure 1(b) illustrates the problem with the methods relying on the loss values. Specifically, Liu et al. (2021) and Hashimoto et al. (2018) upweigh high-loss points, overfitting the outlier. Zhai et al. (2021) optimize Hashimoto et al. (2018)'s objective function after discarding a fraction of points with the largest loss values to account for outliers. They assume that outliers will have higher loss values than the minority group samples, which can easily be violated leading to exclusion of the minority samples, as illustrated in Figure 1(b).

Gradients as data representations. Given a model $h_{\boldsymbol{\theta}_0}(\cdot)$ and loss function $\ell(\cdot,\cdot)$, we consider an alternative representation of the data where each sample is mapped to the gradient with respect to the model parameters of the loss on this sample:

$$\boldsymbol{f}_i = \left.\frac{\partial \ell(\mathrm{y}_i, h_{\boldsymbol{\theta}}(\mathrm{x}_i))}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \quad \text{for } i = 1, \ldots, n. \tag{2}$$

We refer to (2) as *gradient representation*. For scalability and efficiency, one can consider a subset of the model parameters for large models with a high number of parameters such as ResNet-50 (He et al., 2016). Gradient representations have also found success in novelty detection (Kwon et al., 2020b), anomaly detection (Kwon et al., 2020a), out-of-distribution inputs detection (Huang et al., 2021), and robust training (Mirzasoleiman et al., 2020). In this work, we show that gradient representations are suitable for simultaneously learning group annotations *and* detecting outliers. Compared to the original feature space, gradient space simplifies the data structure, making it easier to identify minority groups. Figure 1(c) illustrates a failure of feature space clustering. Here the majority group for class $y = 0$ is a mixture of three components with one of the components being close to the minority group in the feature space. In the gradient space, for a logistic regression model, representations of misclassified points remain similar to the original features, while the representations of correctly classified points are pushed towards zero. We illustrate the benefits of clustering in the gradient space in Figure 1(d) and provide additional details in the next section.

C Extended Description of GRASP

In this section, we provide more details of Gradient Space Partitioning (GRASP), our method for identifying minority groups and outliers through clustering in the gradient space. We first demonstrate through an empirical study of synthetic and semi-synthetic datasets that gradient space is more suitable for clustering than feature space. We also present theoretical results showing that correctly partitioning minority groups alone is sufficient for achieving good worst-group performance, even if majority groups are not correctly partitioned. Lastly, the details of GRASP are provided in Sec. C.2.

C.1 Gradient Space vs Feature Space

Sufficiency of identifying minority groups. As we discussed above, representing data in gradient space helps to cluster correctly classified samples into one group and to identify the minority groups. We now theoretically prove that identifying minority groups is sufficient for achieving good worst-group performance.

Assume that input feature $\mathbf{x} \sim \mathcal{P}_{\mathbf{x}}$, group annotation $\mathrm{g} \sim \mathcal{P}_{\mathrm{g}}$, where $\mathcal{P}_{\mathrm{g}}(\mathrm{g} = g) = p_g > 0$ for all $g \in [G]$ and $\sum_{g=1}^{K} p_g = 1$. Let $\mathrm{y} \mid \mathbf{x} = \boldsymbol{x}, \mathrm{g} = g \sim \mathrm{Bern}(\eta_g(\boldsymbol{x}))$, where $g \in [K]$. Note that $\mathrm{y} \mid \mathbf{x} = \boldsymbol{x} \sim \eta(\boldsymbol{x})$, where $\eta : \mathcal{X} \to [0, 1]$ as $\eta(\boldsymbol{x}) = \sum_{g\in[K]} p_g \eta_g(\boldsymbol{x})$. Let $\alpha = \mathbb{P}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}(|\eta(\mathbf{x}) - 1/2| < \delta)$ indicate the separability of the dataset, where $\delta > 0$ is a constant. Let $\mathcal{H} = \{h : \mathcal{X} \to [0, 1]\}$ be the space of random classifiers. Given $h \in \mathcal{H}$ and data sample $\boldsymbol{x}$, we consider the following random predictions: $\hat{\mathbf{y}} \mid \mathbf{x} = \boldsymbol{x} \sim \mathrm{Bern}(h(\boldsymbol{x}))$. Given a group partition P, the risk of group $g \in [|P|] = [G]$ and gDRO classifier w.r.t. P are defined as

$$\mathcal{R}_g(h; P) \triangleq \mathbb{E}\left[h(\mathbf{x})(1 - \mathrm{y}) + (1 - h(\mathbf{x}))\mathrm{y} \mid \mathrm{g} = g\right], \tag{3}$$

$$h_{\mathrm{gDRO}} = \arg\min_{h\in\mathcal{H}} \max_{g\in[G]} \mathcal{R}_g(h; P). \tag{4}$$

We now mathematically define majority and minority groups based on the similarity between their group-specific decision boundaries η_g and overall decision boundary η.

Table 5: **Group identification quality of clustering methods in feature and gradient space measured by Adjusted Rand Index (ARI).** Higher ARI indicates higher group identification quality. Three clustering methods are considered: K-means, DBSCAN w.r.t. Euclidean distance (DBSCAN/Euclidean), and DBSCAN w.r.t. centered cosine distance (DBSCAN/Cos). We set the number of groups per class $k = 2$ for K-means. The reported numbers show that gradient space clustering noticeably outperforms feature space clustering.

	Dataset	Synthetic		Waterbirds	
	Outliers?	✗	✓	✗	✓
Feature Space	K-means	.5505	.3631	.3932	.3932
	DBSCAN/Euclidean	.5923	.6042	.0000	.0000
	DBSCAN/Cos	.5133	.4946	.0418	.0418
Gradient Space	K-means	**.8409**	.6436	.7235	.7171
	DBSCAN/Euclidean	.7724	**.7237**	.7304	.7304
	DBSCAN/Cos	.6943	.6944	**.7453**	**.7453**

Definition C.1 (Majority and minority groups)**.** Assume that group partition P can be divided into two nonempty disjoint sets: majority groups $\mathcal{A}$ and minority groups $\mathcal{B}$:

$$\mathcal{A} = \{g \in [G] : \mathbb{P}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}(|\eta_g(\mathbf{x}) - \eta(\mathbf{x})| < \delta) \geq 1 - \epsilon\}\,, \qquad \text{(Majority Groups)}$$

$$\mathcal{B} = \{g \in [G] : \mathbb{P}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}(|\eta_g(\mathbf{x}) - \eta(\mathbf{x})| < \delta) < 1 - \epsilon\}\,, \qquad \text{(Minority Groups)}$$

where $\epsilon \in [0, 1]$ captures how close the majority groups' decision boundaries are to the overall decision boundary.

A small ϵ implies better performance of the ERM classifier on the majority groups. Given Definition C.1, the following theorem states that if the dataset is well-separated (i.e., α is small) and the performance on majority groups is good (i.e., ϵ is small), applying gDRO on predicted group partition P' which only identifies minority groups correctly can guarantee good worst-group performance.

Theorem C.2 (Error Analysis)**.** *Assume a predicted group partition P' that identifies the minority groups correctly, i.e., there exists an index set $\mathcal{B}' \subset [G']$ such that $\{\mathcal{G}'_{g'}\}_{g'\in\mathcal{B}'} = \{\mathcal{G}_g\}_{g\in\mathcal{B}}$. Denote the gDRO classifier w.r.t. predicted group partition P' as h'_{gDRO}, which is defined by (4) in terms of P'. Then the worst-case risk of f'_{gDRO} satisfies the inequality*

$$\max_{g\in[G]} \mathcal{R}_g(h'_{\text{gDRO}}; P) - \min_{h\in\mathcal{H}} \max_{g\in[G]} \mathcal{R}_g(h; P) \leq |\mathcal{A}|\epsilon(1-\alpha) + \alpha.$$

Remark C.3. When the performance on majority groups is not good (indicated by a large ϵ), identifying minority groups only might not be sufficient. However, as discussed in Sec. C.2, our GraSP can identify groups that suffer from high loss. Therefore, in this case, our GraSP has the potential to correctly identify majority groups and still achieve good worst-group performance.

Quantitative comparison. We now empirically compare the group identification quality of clustering in feature space and gradient space on two datasets consisting of four groups each. We consider both clean and contaminated versions. The first dataset is Synthetic based on the Figure 1 illustration. The second dataset is known as Waterbirds (Sagawa et al., 2019). It is a semi-synthetic dataset of images of two types of birds placed on two types of backgrounds. We embed the images with a pre-trained ResNet50 (He et al., 2016) model. To obtain gradient space representations, we trained logistic regression models. See Section 3 for additional details.

We consider three popular clustering methods: K-means, DBSCAN with Euclidean distance, and DBSCAN with centered cosine distance. Group annotations quality is evaluated using the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985), a measure of clustering quality. Higher ARI indicates higher group annotations quality, and ARI $= 1$ implies the predicted group partition is identical to the true group partition. The definition of ARI is provided in Appendix A. Table 5 shows that clustering in gradient space outperforms clustering in feature space, providing empirical evidence that gradient space facilitates learning of group annotations. Visualizations of the feature and gradient spaces of the datasets can be found in Appendix E.2.1.

C.2 GraSP for Group Inference and Outlier Identification

We now describe how to use GraSP to train a distributionally and outlier robust model.

Clustering method and distance measure. Results in Table 5 indicate that both K-means and DBSCAN perform well in the gradient space. DBSCAN is a density-based clustering algorithm, where clusters are defined as areas of higher density, while the rest of the data is considered outliers. In this work, we choose to use DBSCAN for its ability to identify outliers, which is an important aspect of the problem we consider. As an additional benefit, unlike K-means, it does not require knowledge of the number of groups. See Appendix A for a detailed description of DBSCAN.

In terms of distance measure, we recommend centered cosine distance due to its better performance on the Waterbirds data, which closer resembles real data. We note that the distance and clustering method choices could be reconsidered depending on the application. For example, for Gaussian-like data without outliers, K-means performed better in Table 5.

GRASP. We present the pseudocode for GRASP in Algorithm 1. We first train an ERM classifier $h_{\boldsymbol{\theta}}(\cdot)$ and collect the gradients of samples' losses w.r.t. model parameters $\boldsymbol{\theta}$. We then compute the pairwise centered cosine distances within each class $y \in \mathcal{Y}$ using gradient representations, as discussed in Sec. C.1. Lastly, to estimate the group annotations and identify outliers, we apply DBSCAN on these distance matrices for each class $y \in \mathcal{Y}$.

Training models with improved worst-group performance in the presence of outliers using GRASP. We discard the identified outliers and provide learned group annotations as inputs to a Group-aware method of choice. For concreteness, in this work, we solve (gDRO) with the method of Sagawa et al. (2019), a stochastic optimization algorithm equipped with convergence guarantees. Note that other choices could be appropriate. For example, methods accounting for noise in group annotations (Lamy et al., 2019; Mozannar et al., 2020; Celis et al., 2021) are also useful as they could counteract mistakes in GRASP annotations.

Remark. We note that the model h_θ and parameter space Θ used for computing gradient representations $\boldsymbol{f}$ and learning group annotations with GRASP can be different from the classifier and parameter space used for the final model training. For example, one can train a logistic regression model (using features from a pre-trained model when appropriate) and collect the corresponding gradients for GRASP, and then train a deep neural network of choice with the estimated group annotations.

D Proofs

D.1 Mathematical interpretation of gradient 1 with centered cosine distance

In Section C.1, we provided a mathematical interpretation of Gradient 1 using Euclidean distance ($|\boldsymbol{f}_i - \boldsymbol{f}_j|_2$). Here, we interpret Gradient 1 using centered cosine distance ($1 - \frac{\langle \boldsymbol{f}_i - \mu_f, \boldsymbol{f}_j - \mu_f \rangle}{|\boldsymbol{f}_i - \mu_f|_2 \cdot |\boldsymbol{f}_j - \mu_f|_2}$)[5]. We compare the (class conditioned; class dependency omitted for simplicity) centering terms in the gradient space (μ_f) and feature space ($\mu_{\boldsymbol{x}}$):

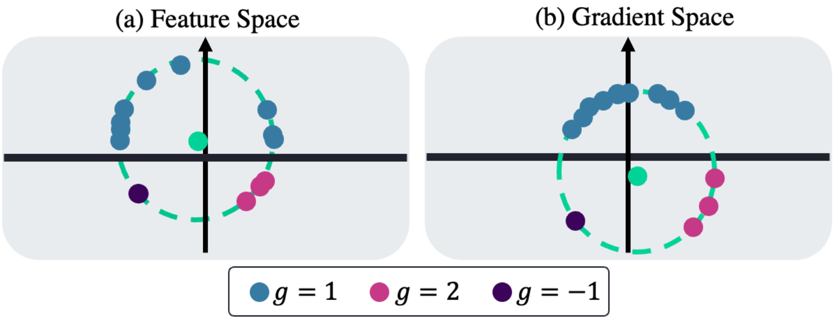
$$\mu_f = \frac{1}{n} \sum_{i:y_i=c} (\sigma(\boldsymbol{w}^\top \mathbf{x}_i + b) - \mathrm{y}_i) \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}, \quad \mu_x = \frac{1}{n} \sum_{i:y_i=c} \mathbf{x}_i.$$

Due to the underrepresentation of the minority group in the data, the feature space center is heavily biased towards the majority group which could hinder the clustering as illustrated in Figure 3(a). On the other hand, the expression of μ_f above implies that gradient space center upweighs high-loss points which are more representative of the minority groups, resulting in a center in-between minority and majority groups. Thus, centering in the gradient space facilitates learning group annotations via clustering with the cosine distance as illustrated in Fig. 3(b).

D.2 Solution of gDRO classifier

Before providing the proof of Theorem C.2 that examines the performance of the gDRO classifier w.r.t. the predicted group partition P', we will first discuss how to solve the gDRO classifier.

[5] Here, μ_f refers to the class-conditional empirical mean of $\boldsymbol{f}$.



r

Figure 3: **Normalized representations of the data from Figure 1(a) in (a) feature space and (b) gradient space.** The green points are the means (before normalization) of the corresponding representations. Gradient space makes it easier to identify groups and detect outliers via clustering with centered cosine distances.

Continuing from Sec. C.1, we define the overall risk and rewrite the risk of group $g \in [|P|] = [G]$ defined in (3) as

$$\begin{aligned}\mathcal{R}(h) &= \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}\left[\mathbb{E}_{\mathrm{y}}\left[h(\mathbf{x})(1-\mathrm{y}) + (1-h(\mathbf{x}))\mathrm{y} \mid \mathbf{x}\right]\right] \\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}[h(\mathbf{x})(1-\eta(\mathbf{x})) + (1-h(\mathbf{x}))\eta(\mathbf{x})], \quad \text{and} \\ \mathcal{R}_g(h;P) &= \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}\left[\mathbb{E}_{\mathrm{y}}\left[h(\mathbf{x})(1-\mathrm{y}) + (1-h(\mathbf{x}))\mathrm{y} \mid \mathrm{g} = g\right]\right] \\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}\left[h(\mathbf{x})(1-\eta_g(\mathbf{x})) + (1-h(\mathbf{x}))\eta_g(\mathbf{x})\right].\end{aligned}$$

The ERM classifier and gDRO classifier w.r.t. group partition P are defined as below:

$$\begin{aligned}h_{\mathrm{ERM}} &= \arg\min_{h\in\mathcal{H}} \mathcal{R}(h), \\ h_{\mathrm{gDRO}} &= \arg\min_{h\in\mathcal{H}} \max_{g\in[G]} \mathcal{R}_g(h;P).\end{aligned}$$

Given the formulation above, the following Lemma gives the closed form of ERM classifier.

Lemma D.1 (ERM classifier)**.** *The Empirical Risk Minimizer (ERM) classifier is h_{ERM}, where*

$$h_{\mathrm{ERM}}(\boldsymbol{x}) = \begin{cases} 1, & \eta(\boldsymbol{x}) \geq 1/2 \\ 0, & \textit{otherwise} \end{cases} \quad \textit{for all } \boldsymbol{x} \in \mathcal{X}.$$

Proof of Lemma D.1. By the definition of ERM classifier,

$$\begin{aligned}\arg\min_{h\in\mathcal{H}} \mathcal{R}(h) &= \arg\min_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}[h(\mathbf{x})(1-\eta(\mathbf{x})) + (1-h(\mathbf{x}))\eta(\mathbf{x})] \\ &= \arg\min_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}\left(1-2\eta(\mathbf{x})\right)h(\mathbf{x}). \qquad (5)\end{aligned}$$

Since the range of h is $[0,1]$, (5) is minimized when $h(\boldsymbol{x}) = \mathbb{I}\left\{1-2\eta(\boldsymbol{x}) < 1/2\right\}$, which implies the desired results. □

We then obtain the closed-form formula of gDRO classifier following similar steps.

Lemma D.2 (gDRO classifier)**.** *Consider a group partition as $P = \{\mathcal{G}_g\}_{g\in[G]}$. The gDRO classifier is*

$$h_{\mathrm{gDRO}}(\boldsymbol{x}) = \begin{cases} 1, & \eta_g(\boldsymbol{x}) \geq 1/2 \textit{ for all } g \in [G], \\ 0, & \eta_g(\boldsymbol{x}) < 1/2 \textit{ for all } g \in [G], \\ 1/2, & \textit{otherwise}, \end{cases}$$

and the minimal worst-group risk is

$$\min_{h\in\mathcal{H}} \max_{g\in[G]} \mathcal{R}_g(h;P) = \max_{g\in[G]} \mathcal{R}_g(h_{\mathrm{gDRO}};P) = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}\left[\min\left(\max_{g\in[G]}(1-\eta_g(\boldsymbol{x})), \max_{g\in[G]}\eta_g(\boldsymbol{x}), 1/2\right)\right].$$

Proof of Lemma D.2. Fix $\boldsymbol{x} \in \mathcal{X}$ and consider the problem

$$\min_{h\in\mathcal{H}} \max_{g\in[G]} \left[h(\boldsymbol{x})(1-\eta_g(\boldsymbol{x})) + (1-h(\boldsymbol{x}))\eta_g(\boldsymbol{x})\right] = \min_{h\in\mathcal{H}} \max_{g\in[G]} \left[(1-2\eta_g(\boldsymbol{x}))h(\boldsymbol{x}) + \eta_g(\boldsymbol{x})\right]. \quad (6)$$

Let $t_1 = h(\boldsymbol{x}) \in [0,1]$ and $t_2 = \max_{g\in[G]}\left[(1-2\eta_g(\boldsymbol{x}))h(\boldsymbol{x})+\eta_g(\boldsymbol{x})\right] = \max_{g\in[G]}\left[(1-2\eta_g(\boldsymbol{x}))t_1+\eta_g(\boldsymbol{x})\right]$. It is easy to verify that (6) is equivalent to the following linear programming problem

$$\min_{t_1,t_2} t_2 \quad \text{s.t.} \quad 0 \le t_1 \le 1, t_2 \ge (1-2\eta_g(\boldsymbol{x}))t_1 + \eta_g(\boldsymbol{x}) \quad \text{for all } g \in [G]. \tag{7}$$

We divide the problem into the following three cases.

Case 1: When $\eta_g(\boldsymbol{x}) \ge 1/2$ for all $g \in [G]$.
For all $g \in [G]$, since $(1-2\eta_g(\boldsymbol{x})) \le 0$, t_2 is minimized when $t_1 = 1$, where the minimum is $\max_{g\in[G]}(1-\eta_g(\boldsymbol{x})) \le 1/2$.

Case 2: When $\eta_g(\boldsymbol{x}) < 1/2$ for all $g \in [G]$.
For all $g \in [G]$, since $(1-2\eta_g(\boldsymbol{x})) > 0$, t_2 is minimized when $t_1 = 0$, where the minimum is $\max_{g\in[G]}\eta_g(\boldsymbol{x}) < 1/2$.

Case 3: Other, i.e., there exists $g_1, g_2 \in [G]$ such that $\eta_{g_1}(\boldsymbol{x}) \ge 1/2$ and $\eta_{g_2}(\boldsymbol{x}) < 1/2$.
Note that lines $t_2 = (1-2\eta_g(\boldsymbol{x}))t_1 + \eta_g(\boldsymbol{x})$ for all $g \in [G]$ intersect at one feasible point $(t_1,t_2) = (1/2, 1/2)$. It is easy to verify that t_2 is minimized when $t_1 = 1/2$, where the minimum is $1/2$.

Therefore, the solution to problem (6) and (7) is

$$h_\star(\boldsymbol{x}) = t_1 = \begin{cases} 1, & \eta_g(\boldsymbol{x}) \ge 1/2 \text{ for all } g \in [G], \\ 0, & \eta_g(\boldsymbol{x}) < 1/2 \text{ for all } g \in [G], \\ 1/2, & \text{otherwise,} \end{cases}$$

with optimal value

$$\min_{h\in\mathcal{H}}\max_{g\in[G]}\mathcal{R}_g(h;P) = t_2 = \min\left(\max_{g\in[G]}(1-\eta_g(\boldsymbol{x})), \max_{g\in[G]}\eta_g(\boldsymbol{x}), 1/2\right),$$

which yields the desired results. □

Now, consider true group parition $P = \{\mathcal{G}_g\}_{g\in[G]}$ and predicted group partition $P' = \left\{\mathcal{G}'_{g'}\right\}_{g'\in[G']}$. Similarly, for the predicted group partition P', let predicted group annotation $\mathrm{g}' \sim \mathcal{P}_{\mathrm{g}'}$, where $\mathcal{P}_{\mathrm{g}'}(\mathrm{g}' = g') = p'_{g'}$ and $\sum_{g'=1}^{G'} p'_{g'} = 1$. Let $\mathrm{y} \mid \mathbf{x} = \boldsymbol{x}, \mathrm{g}' = g' \sim \mathrm{Bern}(\eta'_{g'}(\boldsymbol{x}))$, where $g' \in [G']$. Therefore, the group-wise risk w.r.t. the predicted group partition P' can be written as

$$\mathcal{R}_{g'}(h;P') = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}\left[h(\mathbf{x})(1-\eta'_{g'}(\mathbf{x})) + (1-h(\mathbf{x}))\eta'_{g'}(\mathbf{x})\right], \quad \text{where } g' \in [|P'|] = [G'].$$

Next, we will show that if the predicted group partition P' identifies the minority groups correctly, applying gDRO on P' guarantees a small worst-case risk. The following Lemma gives the decision boundaries of the predicted groups.

Lemma D.3 (Decision boundary of predicted groups)**.** *Denote the gDRO classifier w.r.t. predicted group partition P' as h'_{gDRO}. Assume that predicted group partition P' identifies the minority groups correctly, i.e., there exists an index set $\mathcal{B}' \subset [G']$ such that $\left\{\mathcal{G}'_{g'}\right\}_{g'\in\mathcal{B}'} = \{\mathcal{G}_g\}_{g\in\mathcal{B}}$. Let $\mathcal{A}' = [G']/\mathcal{B}'$. The decision boundaries $\eta'_{g'}$ of the predicted groups satisfies that*

1. for $g' \in \mathcal{A}'$,

$$\eta'_{g'} = \sum_{g\in\mathcal{A}} w_g\eta_g \, \textit{for some } w_g \in [0,1] \textit{ with } \sum_{g\in\mathcal{A}} w_g = 1;$$

2. for $g' \in \mathcal{B}'$,

$$\eta'_{g'} = \eta_g \quad \textit{for some } g \in \mathcal{B}.$$

Proof of Lemma D.3. For all $g' \in \mathcal{B}'$, since $\left\{\mathcal{G}'_{g'}\right\}_{g'\in\mathcal{B}'} = \{\mathcal{G}_g\}_{g\in\mathcal{B}}$, there exists a corresponding $g \in \mathcal{B}$ such that η_g such that $\mathcal{G}_g = \mathcal{G}'_{g'}$ and hence $\eta'_{g'} = \eta_g$.

Now consider $g' \in \mathcal{A}'$. Recall that $\mathcal{D}$ is the dataset. Note that

$$\bigcup_{g'\in\mathcal{A}'}\mathcal{G}'_{g'} = \mathcal{D}/\bigcup_{g'\in\mathcal{B}'}\mathcal{G}'_{g'} = \mathcal{D}/\bigcup_{g\in\mathcal{B}}\mathcal{G}_g = \bigcup_{g\in\mathcal{A}}\mathcal{G}_g,$$

where $\bigcup_{g\in\mathcal{A}}\mathcal{G}_g$ is the union of majority groups. Therefore, data sample from any group $\mathcal{G}'_{g'}$ with $g'\in\mathcal{A}'$ follows the mixture distribution of the majority groups, i.e., for $g'\in\mathcal{A}'$,

$$\mathrm{y}\mid \mathbf{x}=\boldsymbol{x}, \mathrm{g}'=g' \sim \mathrm{Bern}(\eta'_{g'}(\boldsymbol{x})), \quad \text{where } \eta'_{g'}(\boldsymbol{x})=\sum_{g\in\mathcal{A}} w_g\eta_g(\boldsymbol{x}), w_g\in[0,1], \text{and } \sum_{g\in\mathcal{A}} w_g=1.$$

□

D.3 Proof of Theorem C.2

Given the helper results presented in Appendix D.2, we now provide the proof for Theorem C.2.

Proof of Theorem C.2. By Lemma D.2, we have

$$h_{\mathrm{gDRO}}(\boldsymbol{x})=\begin{cases}1, & \eta_g(\boldsymbol{x})\ge 1/2 \text{ for all } g\in[G],\\ 0, & \eta_g(\boldsymbol{x})<1/2 \text{ for all } g\in[G],\\ 1/2, & \text{otherwise},\end{cases}$$

and

$$h'_{\mathrm{gDRO}}(\boldsymbol{x})=\begin{cases}1, & \eta'_{g'}(\boldsymbol{x})\ge 1/2 \text{ for all } g'\in[G'],\\ 0, & \eta'_{g'}(\boldsymbol{x})<1/2 \text{ for all } g'\in[G'],\\ 1/2, & \text{otherwise}.\end{cases}$$

Therefore, $h_{\mathrm{gDRO}}(\mathbf{x})\ne h'_{\mathrm{gDRO}}(\mathbf{x})$ if and only if one of the following three events happen:

$\mathcal{E}_1:$ $\eta_g(\mathbf{x})\ge 1/2$ for all $g\in[G]$ and there exists $g'\in[G']$ such that $\eta'_{g'}(\mathbf{x})<1/2$;
$\mathcal{E}_2:$ $\eta_g(\mathbf{x})<1/2$ for all $g\in[G]$ and there exists $g'\in[G']$ such that $\eta'_{g'}(\mathbf{x})\ge 1/2$;
$\mathcal{E}_3:$ there exists $g_1, g_2\in[G]$ such that $\eta_{g_1}(\mathbf{x})\ge 1/2$ and $\eta_{g_2}(\mathbf{x})<1/2$, while $\eta'_{g'}(\mathbf{x})\ge 1/2$ for all $g'\in[G']$ or $\eta'_{g'}(\mathbf{x})<1/2$ for all $g'\in[G']$.

Consequently,

$$\mathbb{P}(h_{\mathrm{gDRO}}(\mathbf{x})\ne h'_{\mathrm{gDRO}}(\mathbf{x}))=\mathbb{P}(\mathcal{E}_1\cup\mathcal{E}_2\cup\mathcal{E}_3).$$

Note that for simplicity, here we use $\mathbb{P}$ to denote $\mathbb{P}_{\mathbf{x}\sim\mathcal{P}_{rvx}}$ without causing any ambiguity.

By Lemma D.3, it is easy to verify that $\mathcal{E}_1$ and $\mathcal{E}_2$ are impossible, i.e., $\mathbb{P}(\mathcal{E}_1)=\mathbb{P}(\mathcal{E}_2)=0$. Therefore,

$$\mathbb{P}(h_{\mathrm{gDRO}}(\mathbf{x})\ne h'_{\mathrm{gDRO}}(\mathbf{x}))=\mathbb{P}(\mathcal{E}_1\cup\mathcal{E}_2\cup\mathcal{E}_3)=\mathbb{P}(\mathcal{E}_3). \tag{8}$$

Note that

$$\begin{aligned}
&\mathbb{P}(\mathcal{E}_3)\\
&=\mathbb{P}\big(\,\{\mathbf{x}: \text{there exists } g\in[G] \text{ such that } \eta_g(\mathbf{x})<1/2 \text{ and } \eta'_{g'}(\mathbf{x})\ge 1/2 \text{ for all } g'\in[G']\}\\
&\qquad\bigcup\{\mathbf{x}: \text{there exists } g\in[G] \text{ such that } \eta_g(\mathbf{x})\ge 1/2 \text{ and } \eta'_{g'}(\mathbf{x})<1/2 \text{ for all } g'\in[G']\}\,\big)\\
&=\mathbb{P}\big(\,\{\mathbf{x}: \text{there exists } g\in\mathcal{A} \text{ such that } \eta_g(\mathbf{x})<1/2 \text{ and } \eta'_{g'}(\mathbf{x})\ge 1/2 \text{ for all } g'\in[G']\}\\
&\qquad\bigcup\{\mathbf{x}: \text{there exists } g\in\mathcal{A} \text{ such that } \eta_g(\mathbf{x})\ge 1/2 \text{ and } \eta'_{g'}(\mathbf{x})<1/2 \text{ for all } g'\in[G']\}\,\big)\\
&\hspace{20em}\big(\text{By Lemma D.3}\big)\\
&\le\mathbb{P}\big(\,\{\mathbf{x}: \text{there exists } g\in\mathcal{A} \text{ such that } \eta_g(\mathbf{x})<1/2 \text{ and } \eta(\mathbf{x})\ge 1/2\}\\
&\qquad\bigcup\{\mathbf{x}: \text{there exists } g\in\mathcal{A} \text{ such that } \eta_g(\mathbf{x})\ge 1/2 \text{ and } \eta(\mathbf{x})<1/2\}\,\big)\\
&\hspace{14em}\big(\eta \text{ is a linear combination of } \{\eta'_{g'}\}_{g'\in[G']}\big)\\
&\le\mathbb{P}(\text{there exists } g\in\mathcal{A} \text{ such that } (\eta_g(\mathbf{x})-1/2)(\eta(\mathbf{x})-1/2)\le 0)\\
&=1-\mathbb{P}(\text{for all } g\in\mathcal{A} \text{ such that } (\eta_g(\mathbf{x})-1/2)(\eta(\mathbf{x})-1/2)>0)\\
&\le 1-\mathbb{P}\,(\text{for all } g\in\mathcal{A} \text{ such that } |\eta_g(\mathbf{x})-\eta(\mathbf{x})|<\delta \text{ and } |\eta(\mathbf{x})-1/2|\ge\delta)\\
&\hspace{22em}\big(\text{triangle inequality}\big)\\
&\le 1-\mathbb{P}\,(\text{for all } g\in\mathcal{A} \text{ such that } |\eta_g(\mathbf{x})-\eta(\mathbf{x})|<\delta)\,\mathbb{P}\,(|\eta(\mathbf{x})-1/2|\ge\delta)\\
&\le 1-(1-|\mathcal{A}|\epsilon)(1-\alpha)=|\mathcal{A}|\epsilon+\alpha-|\mathcal{A}|\epsilon\alpha.
\end{aligned} \tag{9}$$

Combining (8) and (9) yields

$$\mathbb{P}(h_{\mathrm{gDRO}}(\mathbf{x}) \neq h'_{\mathrm{gDRO}}(\mathbf{x})) \leq |\mathcal{A}|\epsilon + \alpha - |\mathcal{A}|\epsilon\alpha.$$

Moreover, for any $g \in [G]$,

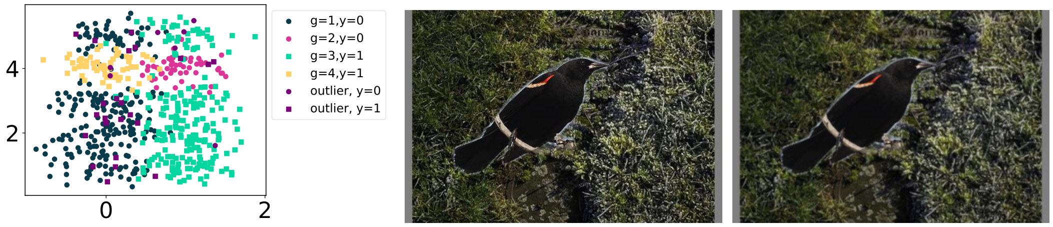
$$\begin{aligned}
\mathcal{R}_g(h'_{\mathrm{gDRO}}; P) - \mathcal{R}_g(h_{\mathrm{gDRO}}; P) &= \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}\left[h'_{\mathrm{gDRO}}(\mathbf{x})(1-\eta_g(\mathbf{x})) + (1-h'_{\mathrm{gDRO}}(\mathbf{x}))\eta_g(\mathbf{x})\right] \\
&\qquad - \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}\left[h_{\mathrm{gDRO}}(\mathbf{x})(1-\eta_g(\mathbf{x})) + (1-h_{\mathrm{gDRO}}(\mathbf{x}))\eta_g(\mathbf{x})\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\mathbf{x}}}\left[(1-2\eta_g(\mathbf{x}))(h'_{\mathrm{gDRO}}(\mathbf{x}) - h_{\mathrm{gDRO}}(\mathbf{x}))\right] \\
&\leq \mathbb{P}(h_{\mathrm{gDRO}}(\mathbf{x}) \neq h'_{\mathrm{gDRO}}(\mathbf{x})) \\
&\leq |\mathcal{A}|\epsilon + \alpha - |\mathcal{A}|\epsilon\alpha.
\end{aligned}$$

Consequently,

$$\max_{g\in[G]} \mathcal{R}_g(h'_{\mathrm{gDRO}}; P) - \max_{g\in[G]} \mathcal{R}_g(h_{\mathrm{gDRO}}; P) \leq |\mathcal{A}|\epsilon + \alpha - |\mathcal{A}|\epsilon\alpha.$$

This directly implies the desired results. □

E Experiment

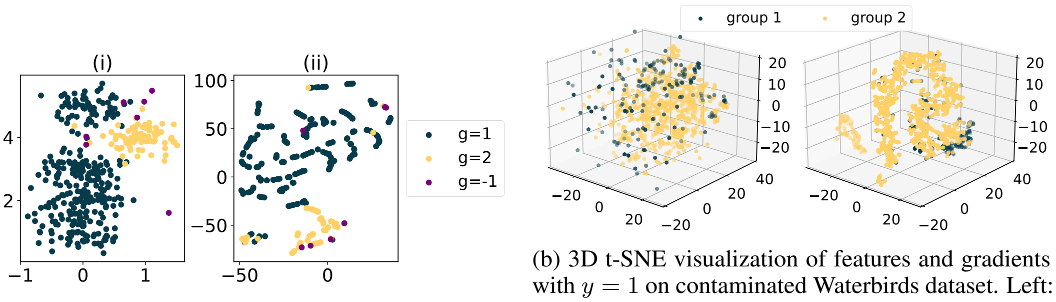


(a) (b) (c)

Figure 4: (a) Scatter plot of contaminated Synthetic dataset. (b) Original image of 010.Red_winged_Blackbird/Red_Winged_Blackbird_0079_4527.jpg. (c) Image 010.Red_winged_Blackbird/Red_Winged_Blackbird_0079_4527.jpg after Gaussian blurring.

E.1 More Details of Experiment Setup

Our experiments are mainly performed on NIVIDA A100 GPUs.



(a) Visualization of input features and 2D t-SNE results of gradients with $y = 0$ on contaminated Synthetic dataset. (i) Input features. (ii) 2D t-SNE results of gradients.

(b) 3D t-SNE visualization of features and gradients with $y = 1$ on contaminated Waterbirds dataset. Left: 3D t-SNE visualization of features extracted from ResNet-50 pretrained on ImageNet (Deng et al., 2009). Right: 3D t-SNE visualization of gradients of the sample's loss w.r.t. the parameters of the last layer.

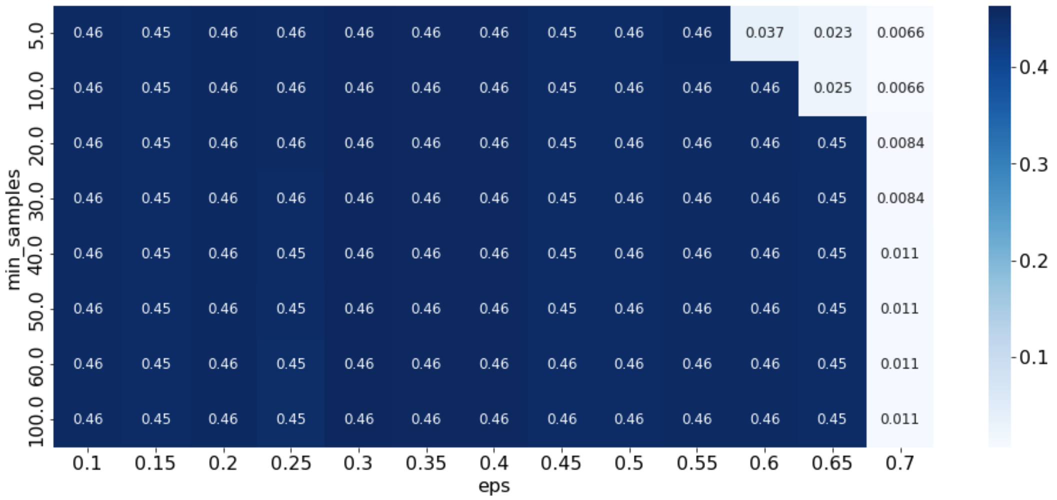


Figure 6: Group identification quality of GRASP v.s. DBSCAN clustering hyperparameters (`eps`: ϵ, `min_samples`: m) measured in Adjusted Rand Index (ARI) on class 0 of Waterbirds dataset.

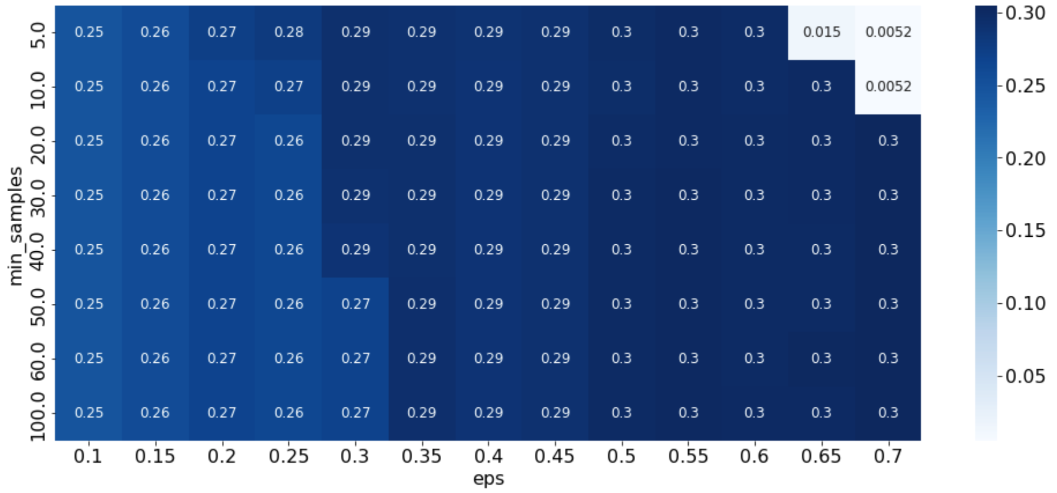


Figure 7: Group identification quality of GRASP v.s. DBSCAN clustering hyperparameters (`eps`: ϵ, `min_samples`: m) measured in Adjusted Rand Index (ARI) on class 1 of Waterbirds dataset.

E.1.1 Datasets and Baselines

Synthetic. We generate a synthetic dataset of 1,000 samples with two features $\mathbf{x} \in \mathbb{R}^2$, a group attribute $\mathrm{g} \in [4]$, and a binary label $\mathrm{y} \in \{0, 1\}$, similar to the motivating example of Figure 1. **(Clean):** The synthetic dataset consists of 10 Gaussian clusters with a variance of 0.01, and each Gaussian cluster contains 100 samples. Class 0 is divided into two groups: group 3 consists of four Gaussian clusters with centers $(1, 5), (1, 3), (1, 2), (1, 1)$; group 2 consists of one Gaussian cluster with center $(0, 4)$. Similarly, Class 1 is divided into two groups: group 1 consists of four Gaussian clusters with centers $(0, 5), (0, 3), (0, 2), (0, 1)$; group 2 consists of one Gaussian cluster with center $(1, 4)$. **(Contaminated):** We contaminate the synthetic dataset by flipping randomly selected 5% of labels. The contaminated synthetic dataset is visualized in Appendix Figure 4a.

Waterbirds. (Clean): Waterbirds (Sagawa et al., 2019; Wah et al., 2011) is a semi-synthetic image dataset of land birds and water birds (Wah et al., 2011) placed on either land or water backgrounds using images from the Places dataset (Zhou et al., 2017). There are 11,788 images of birds on their typical (majority) and atypical (minority) backgrounds. The task is to predict the types of birds and the background type is the group (2 background types per class, a total of 4 groups). We pre-process the dataset as in Idrissi et al. (2022). **(Contaminated):** We contaminate the Waterbirds dataset by introducing outliers in the training and validation datasets by flipping 2% of the class labels, transforming 1% of the images with Gaussian blurring, color dithering 1% of the images, and posterizing 1% of the images (4 bits per color channel). See Appendix Figure 4c for a visual example of a contaminated image.

COMPAS & CivilComments. Both datasets are real and collected by humans, therefore likely to contain outliers. **(COMPAS):** COMPAS (ProPublica, 2021) is a recidivism risk score prediction dataset consisting of 7,214 samples. Each class $y \in [0, 1]$ is divided into six groups: Caucasian males, Caucasian females, African-American males, African-American females, males of other races, and females of other races, making 12 groups in total. **(CivilComments):** CivilComments (Dixon et al., 2018; Koh et al., 2021) is a language dataset containing online forum comments. The task is to predict whether comments are toxic or not. We preprocess the dataset as in Idrissi et al. (2022). We divide comments in each class into two groups according to the presence or absence of identity terms pertaining to protected groups (LGBT, Black, White, Christian, Muslim, other religion).

Experimental baselines. We compare GRASP to four different types of baselines: (1) standard empirical risk minimization (ERM), (2) a group-aware method (gDRO (Sagawa et al., 2019)), (3) a group-oblivious method (DORO, CVaR-DORO variation (Zhai et al., 2021)), and (4) two group-learning methods (EIIL (Creager et al., 2021), George (Sohoni et al., 2020)). We chose DORO among the methods relying on loss values to improve worst-group performance because it is the only method from this group designed to be robust to outliers. Recall that only the group-aware method (gDRO) has access to the true group annotations, thus it should be interpreted as an "oracle" baseline.

For all methods, on a given dataset, we train models with the same architecture and initialization. Recall that these models can be different from the models used in estimating group annotations with any of the group-learning methods. See Appendix E.1 for details.

We also perform an ablation study by considering an additional group-learning baseline, Feature Space Partitioning (FeaSP). It is identical to GraSP except it performs DBSCAN clustering in the feature space. Comparison to FeaSP emphasizes the importance of clustering in the gradient space as opposed to other choices such as the clustering method and distance measure.

The batch size of Synthetic, Waterbirds, COMPAS, and CivilComments datasets are 128, 128, 128, and 32 for both group inference and downstream DRO tasks. We split the Synthetic and COMPAS datasets into training, validation, and test datasets at the ratio of 0.6:0.2:0.2. We follow an identical procedure to Idrissi et al. (2022) to pre-process the Waterbirds and Civilcomments dataset. Fig. 4a visualizes the contaminated Synthetic dataset. We provide an example of a contaminated sample in Fig. 4b and Fig. 4c, which present an image before and after Gaussian blurring.

E.1.2 Group annotations quality.

To collect the gradients of the corresponding datum's loss w.r.t. the model parameters, we train a logistic regression model on 50 epochs on the Synthetic dataset, a three-layer ReLU neural network with 50 hidden neurons for 300 epochs on the COMPAS dataset, and a BERT (Devlin et al., 2018) model for 10 epochs on the CivilComments dataset. For the Waterbirds dataset, we first featurize the images using a ResNet50 pre-trained on ImageNet (Deng et al., 2009), and then train a logistic regression for 360 epochs We tune the DBSCAN clustering hyperparameters $\epsilon \in \{.1, .2, .3, .5, .7\}, m \in \{10, 20, 30, 50, 70, 100\}$ for each $y \in \mathcal{Y}$, for both FeaSP and GRASP. We tune the learning rate of EIIL in $\left\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\right\}$, run EIIL for 50 epochs on Synthetic, Waterbirds, and COMPAS datasets, three epochs on Civilcomments dataset. We tune the overcluster factor of George in $\{1, 2, 5, 10\}$, and employ the over-cluster Gaussian Mixture Model clustering for George. Lastly, we select the best EIIL epoch and all hyperparameters mentioned above based on *Silhouette Coefficient*, a measure assessing the clustering quality in terms of the degree to which a sample clusters with other similar samples.

E.1.3 Worst-group performance.

We use Adam optimizer for all trainings. We tune outlier fraction $\epsilon \in \{.005, .01, .02, .1, .2\}$ and minimal group fraction $\in \{.1, .2, .5\}$ for (CvAR-)DORO on all datasets. We tune the learning rate $\in \left\{10^{-5}, 10^{-4}, 10^{-3}\right\}$ and weight decay $\in \left\{10^{-4}, 10^{-3}, 10^{-2}\right\}$ for all methods. We select the step size of the group weights q in gDRO (Sagawa et al., 2019) $\in \{.001, .01, .1\}$. We train a three-layer

ReLU neural network with 50 hidden neurons per layer for 50 and 300 epochs on the Synthetic and COMPAS datasets, respectively. We train a logistic regression model with 360 epochs on the Waterbirds dataset, and a BERT model for 10 epochs on the Civilcomments dataset.

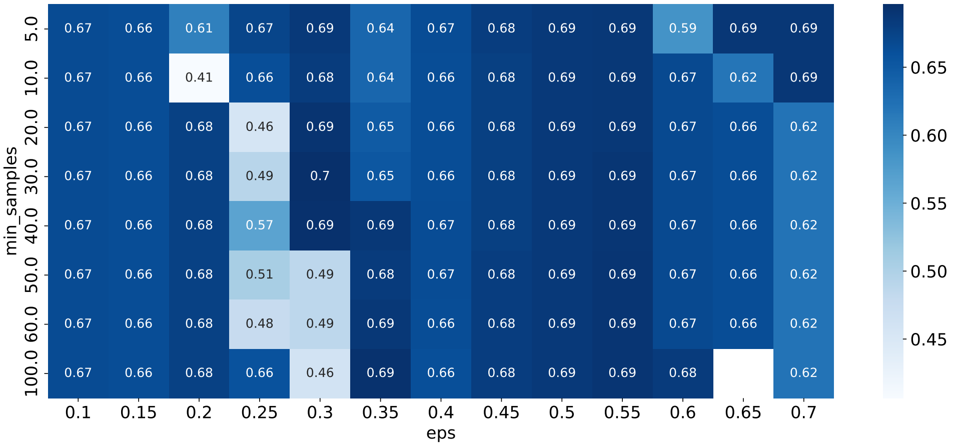


Figure 8: Worst-group accuracy of GRASP v.s. DBSCAN clustering hyperparameters (`eps`: ϵ, `min_samples`: m) on Waterbirds dataset.

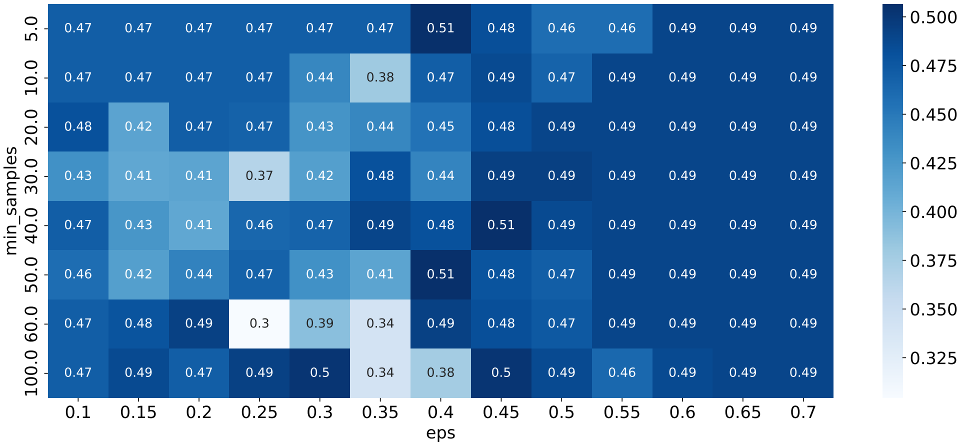


Figure 9: Worst-group accuracy of GRASP v.s. DBSCAN clustering hyperparameters (`eps`: ϵ, `min_samples`: m) on COMPAS dataset.

E.2 Additional Experiments

E.2.1 Visualization of Gradient Space and Feature Space

In this section, we visualize the gradient space and feature space of contaminated Synthetic (see Fig. 5a) and Waterbirds dataset (see Fig. 5b).

E.2.2 Robustness to DBSCAN Clustering Hyperparameters

In this experiment, we investigate the effect of clustering hyperparameters on group inference and downstream DRO task performances. In doing so, we let $\epsilon \in \{.1, .15, .2, .25, .3, .35, .4, .45, .5, .55, .6, .65, .7\}$ and $m \in \{5, 10, 20, 30, 40, 50, 60, 100\}$ and visualize how ARI varies with different choice of clustering hyperparameters on different classes of Waterbirds dataset in Fig. 6 and Fig. 7. We observe that the group identification performance is robust to clustering hyperparameters. For worst-group performance, we set the ϵ and m to be the same for different classes on the datasets. We visualize how it varies with clustering hyperparameters on Waterbirds and COMPAS dataset in Fig. 8 and Fig. 9. A similar phenomenon is observed for worst-group performance — we find that worst-group performance is fairly robust to DBSCAN clustering hyperparameters.

F Discussion and Conclusion

We considered the problem of learning in the presence of outliers and minority groups. Our method facilitates the training of models that exhibit comparable performance across groups, even without group annotations and in the presence of outliers, by leveraging gradient space clustering to predict group annotations and identify outliers. In general, a limitation of the gradient space is the simplification of the structure of correctly classified points, typically found in the majority group. While perhaps not ideal for data exploration, it does not impact downstream worst-group performance.

As a next step, when training models with GRASP group annotations, it would be interesting to consider alternatives to gDRO that are accounting for noise in group annotations (Lamy et al., 2019; Mozannar et al., 2020; Celis et al., 2021) to counteract potential errors in GRASP annotations. Alternatively, one can consider training with group-oblivious methods such as DORO (Zhai et al., 2021) and performing model selection on the validation data with GRASP group annotations.