Auxiliary Variables help Improving Group Robustness Through Bias Amplification

Jiarui Liu
Department of Computer Science
University of Michigan
ljrjerry@umich.edu

Muzhe Wu
Department of Computer Science
University of Michigan
henrw@umich.edu

Abstract

Neural networks produced by standard training tend to suffer from poor accuracy on rare subgroups despite achieving high accuracy on average due to the *spurious correlation*. Previously we proposed BAM, a novel two-stage training algorithm, comprising a *bias amplification* stage with the learnable *auxiliary variable* and rebalanced training stage using upweighted error samples, which empirically improves the worst-group accuracy in various benchmarks with state-of-the-art performance. To investigate the roles of different elements in BAM, including the *auxiliary variables* and *Class Difference* stopping criterion, in this paper, we adopt two new datasets and conduct extensive experiments with carefully controlled parameters. With these experiments, we not only justify the assumption on the effectiveness of BAM but also shows the ubiquity of ClassDiff, for which we provide detailed discussions.

1 Introduction

This is a followup work based on our prior work [22] [1].

The *spurious correlation* is a phenomenon in machine learning where a model tends to "learn" certain decision rules based on spurious features such as backgrounds, thus likely to have unintended behaviors for subgroups from a data distribution. It prevails in various fields, including computer vision [1], natural language processing [9], and reinforcement learning [20], and due to the ubiquity and gravity of this problem, extensive efforts have been made to address it.

Prior works mainly include methods that require group annotations for the whole training set or only the validation set. For the latter, they usually train two models, where the first is used to identify minority samples, based on which a second model is trained focusing on improving worst-group classification performance [23, 26, 27, 40]. In contrast, in our prior work, we proposed the BAM algorithm, which only requires training one model. BAM also contributes some interesting insights, such as the two-stage training process comprising bias amplification with squared loss and rebalanced training. With experiments on some commonly used benchmarks [35, 30, 24, 30, 36, 30, 2, 18], we observed that BAM leads to consistent improvement in worst-group accuracy and achieves state-of-the-art performance. However, to better understand why this procedure works well and how this approach can be applied to more generalized settings, we need to look deeper into different elements in BAM and perform experiments on more carefully constructed datasets.

In this work, we adopt two additional datasets: Controlled-Waterbirds and Colored-MNIST. By carefully controlling their specifications and conducting experiments with various settings, such as

[1] We only provide abridged version of Related Works and Methodology Sections to avoid too much repetition. Please refer to the original paper for more details.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

auxiliary variables, upweight factors, stopping epochs, and losses, we gain deeper insights into each of them and validate the robustness of BAM.

2 Related Works

A variety of recent work discussed different realms of robustness, for instance, class imbalance [10, 12, 15, 14, 33], and robustness in distribution shift, where the target data distribution is different from the source data distribution [5, 39, 25, 19, 38]. In this paper, we mainly focus on improving group robustness. Categorized by the amount of information we have for training and validation, we discuss three directions below:

Improving Group Robustness with Training Group Annotations. Multiple works have used training group annotations to improve worst-group accuracy [3, 16, 8, 4, 31]. Other works include minimizing the worst-group training loss using distributionally robust optimization (Group-DRO) [30], simple training data balancing (SUBG) [13], and retraining the last layer of the model on the group-balanced dataset (DFR) [17]. These methods achieve state-of-the-art performance on all benchmark datasets. However, acquiring spurious attributes of the entire training set is costly and unrealistic in real-world datasets.

Improving Group Robustness with Validation Group Annotations Only. Acknowledging the cost of obtaining group annotations, many recent works focus on the setting where training group annotations are not available [7, 28, 21, 29]. Approaches closely related to our method usually use the first model to identify minority samples and then train a separate model based on the results predicted by the first model [37, 34]. For example, JTT [23] first trains an ERM model to identify minority groups in the training set (similar to EIIL [6]), and then trains a second ERM model with these selected samples to be upweighted. However, these approaches all focus on the robust training of the second model and fail to consider the potential of accumulating biased knowledge from the first model.

Improving Group Robustness without any Group Annotations. Relatively little work has been done under the condition that no group information is provided for both training and validation. [13, 23] observe a significant drop (10% - 25%) in worst-group test accuracy if using the highest *average* validation accuracy as the stopping criterion without any group annotations. A recent work, GEORGE [32], tries to separate unlabeled classes in deep model feature spaces and then use the generated pseudo labels to train the model via the distributionally robust optimization objective. However, there is a considerable performance gap between GEORGE and the supervised methods.

3 Methodology

In the previous work, we designed the BAM training algorithm consisting of a two stages.

In Stage 1, we train a bias-amplified model to bias toward majority group samples unintentionally. Inspired by the work done by [11], we introduce trainable auxiliary variables b_i for each data sample and add it to the network's output. The objective function is formally defined as:

$$R_1(\theta, B) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i) + \lambda b_i, y_i). \tag{1}$$

where $f_\theta : \mathcal{X} \to \mathbb{R}^C$ is a neural network with parameters θ and the class number $C = |\mathcal{Y}|$, and λ is a hyperparameter that controls the strength of the auxiliary variables. We also further adopt squared loss $\ell(z, y) = \|z - e_y\|_2^2$ where $e_y \in \mathbb{R}^C$ is the one-hot encoding for the label y.

In stage 2, we continue to train the (same) model using the error set we obtained in Stage 1, i.e., data points the model misclassifies. We adopt the rebalanced loss that upweights the examples in the error set E:

$$R_2(\theta) = \mu \sum_{(x,y)\in E} \ell_{\text{CE}}(f_\theta(x), y) + \sum_{(x,y)\in D \setminus E} \ell_{\text{CE}}(f_\theta(x), y), \tag{2}$$

where ℓ_{CE} is the cross-entropy loss and μ is a hyperparameter (upweight factor).

In particular, we have a special stopping criterion for Stage 2 training. Note that the group annotations may or may not be available in our setting. We will stop at the highest worst-group validation accuracy if it is available. However, if there are no group annotations presented, we will instead use the *minimum class difference*, i.e., stop when the sum of pairwise validation accuracy differences between classes is at the minimum. The class difference can be formally defined as

$$\text{ClassDiff} = \sum_{i,j=1}^{C} |\text{Acc}(\text{class } i) - \text{Acc}(\text{class } j)|. \tag{3}$$

In this work, we use the same method while focusing more on the experiment side.

4 Experiments

4.1 Datasets

The previous method we proposed, BAM, has been tested on datasets including Waterbirds, CelebA, CivilComments-WILDS, and MultiNLI (see Table 3 for specifications) and outperformed state of the art on nearly all of them. Some ablation studies about loss functions and auxiliary variables have also been done to prove that all the model components are useful. However, as we see in Table 3, the sizes of different classes and groups are not balanced on all four datasets. As the training process becomes undetermined and unpredictable because of the nature of these datasets, we could not perform a fine-grained analysis of the role of each model component. In addition, Waterbirds and CelebA have significant distribution shifts in the training and test data, and two NLP datasets have too much noise that impedes deep learning models from learning well on the classification task, even without spurious correlation.

To further investigate how auxiliary variables, upweight factors, stopping epochs in Stage 1, and losses play their roles in BAM and address the problems above, we consider two new CV datasets: Controlled-Waterbirds and Colored-MNIST, and manually control their class and group counts. In Colored-MNIST, we set class 0 for numbers from 0 to 4 and class 1 for numbers from 5 to 9. We make the class labels spuriously correlated with colors and randomly flip the color of a small subset to generate minority groups. Figure 1 shows what their majority and minority groups are. Table 1 shows some parameters of the two datasets.

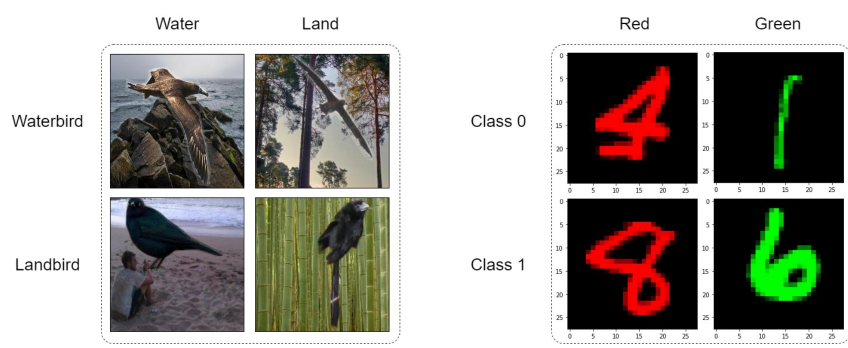


Figure 1: Based on their group sizes, for Controlled-Waterbirds on the left, WW and LL are majority groups. Similarly, for Colored-MNIST on the right, Class 0 in red and Class 1 in Green are majority groups.

Table 1: Important parameters to construct Controlled-Waterbirds and Colored-MNIST.

Dataset	Parameter	Description
Controlled-Waterbirds	d	group sizes of WW, WL, LW, LL in order
	s	total dataset size
Colored-MNIST	fr	minority group ratios in class 0 & 1
	cr	class ratios of class 0 & 1
	s	total dataset size

4.2 The Role of Auxiliary Variables

To validate that auxiliary variables help improve the model performance in a non-trivial way, we did additional ablation studies on new datasets. Section 4.2

Figure 2 shows the change of model performance with different auxiliary variables. Note that $\lambda = 0$ will disable the auxiliary variables to fit training samples. With a moderately large $\lambda \cdot b$, it is easy to observe that the model achieves higher robust worst-group test accuracy than without b on both datasets. It indicates the important role of auxiliary variables.

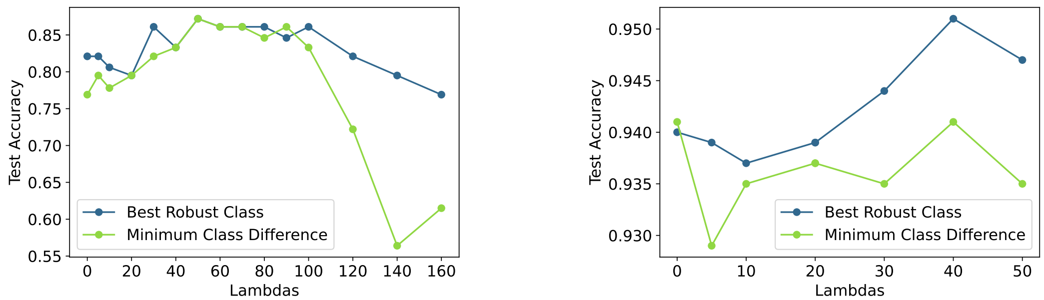


Figure 2: Worst-group accuracy with different lambdas on Controlled-Waterbirds (left) and Colored-MNIST (right). For Controlled-Waterbirds, we fix dataset parameters $d = \{1800, 200, 200, 1800\}$ and hyperparameters $\mu = 100, T = 100$. For Colored-MNIST, we fix dataset parameters $fr = \{0.1, 0.1\}, cr = 1.0, s = 50000$ and hyperparameters $\mu = 10, T = 50$. For simplicity, we choose the same hyperparameters of Controlled-Waterbirds as in Waterbirds in the previous work.

4.3 Negative Correlations of ClassDiff and Worst-Group Accuracy

Section 3 It explains the definition of ClassDiff, which is mainly used to evaluate when to stop Stage 2 without the help of validation group annotations. In Figure 3 of the previous work BAM, we show the negative correlation between absolute validation class difference and worst-group accuracy on all four benchmarks (Waterbirds, CelebA, CivilComments-WILDS, MultiNLI). In this project, we did abundant experiments to verify its success on new datasets by varying the dataset sizes, class size ratios, and group size ratios over a large range. In every single setting we tested, we observed similar negative correlations. These findings suggest that ClassDiff could be useful in general when no group annotation is available, as it is robust across different datasets and parameters. Here we show one example per dataset each in Figure 3. Appendix A.3 displays a larger and random subset of our experiments without cherry picking, and explains the subset generation details.

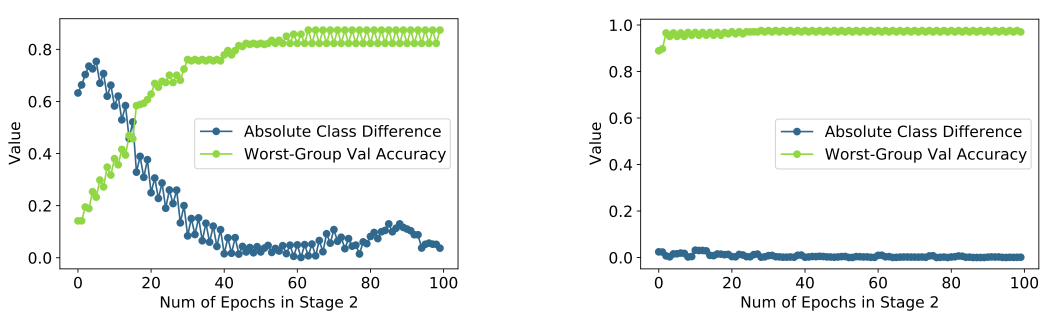


Figure 3: Visualization on relations between ClassDiff and worst-group validation accuracy. Left: Controlled-Waterbirds. $d = \{1400, 200, 600, 4200\}, \lambda = 50, T = 100, \mu = 100$. Right: Colored-MNIST. $s = 50000, fr = \{0.2, 0.2\}, cr = 1.0, \lambda = 20, T = 11, \mu = 5$.

5 Discussions

Different Dataset Sizes Since the total size of Waterbirds is limited, varying the total size could cause a very small size of minority groups. Figure 4 shows the performance comparison of $\lambda = 0$ vs. $\lambda = 50$. On Colored-MNIST, we see a tiny but persistent difference of different λ's on robust worst-group test accuracy from the left plot. The relation between ClassDiff and λ seems not observable, while ClassDiff achieves good performance on all the total dataset size s listed above. The robust test accuracy is comparably robust to s.

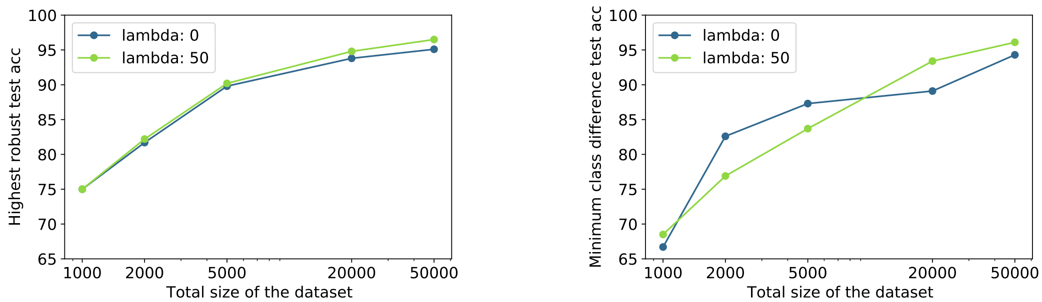


Figure 4: Robust worst-group accuracy and minimum ClassDiff accuracy for different dataset sizes on Colored-MNIST. We consider $s \in \{1000, 2000, 5000, 20000, 50000\}$, fix $fr = \{0.1, 0.1\}, cr = 1.0$, and tune $\mu \in \{5, 10, 20, 50\}, T \in \{20, 50\}$.

Different Class Imbalance Ratios Figure 5 shows the robustness of BAM w.r.t. class imbalance ratios.

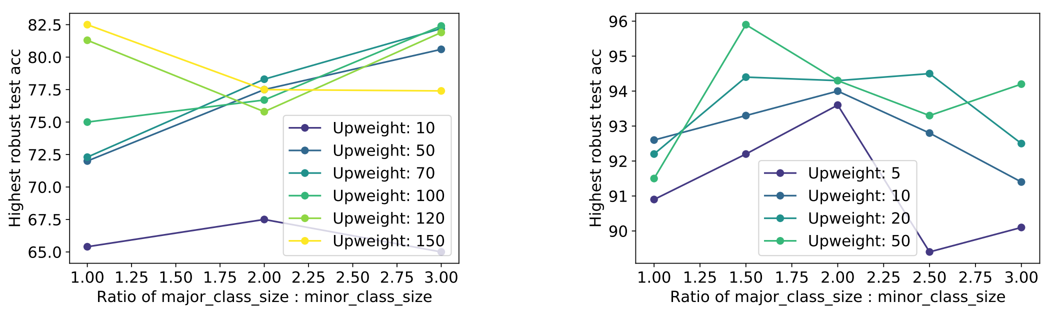


Figure 5: Robust worst-group accuracy for different class imbalance ratios. For Controlled-Waterbirds (left), we choose WW $= 1400$, WL $= 200$, $\lambda = 50, T = 100$. For Colored-MNIST (right), we choose $s = 20000, fr = \{0.1, 0.1\}, \lambda = 50, T = 20$. Each point is averaged over 3 random experiments.

Different Group Imbalance Ratios Figure 6 shows the robustness of BAM w.r.t. group imbalance ratios.

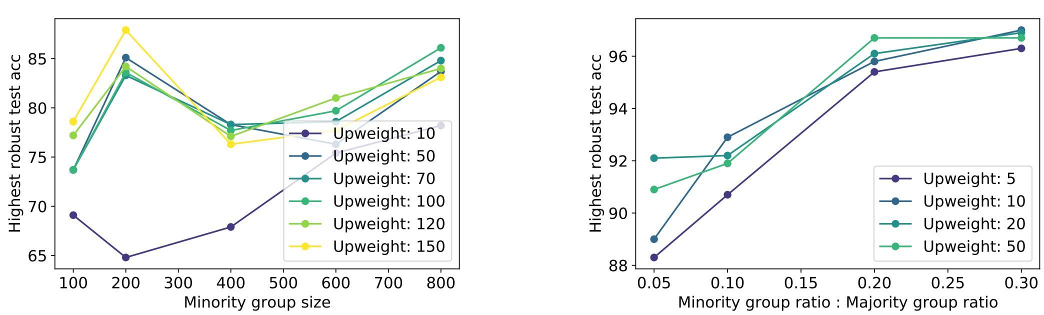


Figure 6: Robust worst-group accuracy for different class imbalance ratios. For Controlled-Waterbirds (left), we choose WW $= 1800$, LL $= 1800$, $\lambda = 50, T = 100$. For Colored-MNIST (right), we choose $s = 20000, cr = 1.0, \lambda = 50, T = 50$. Each point is averaged over 3 random experiments.

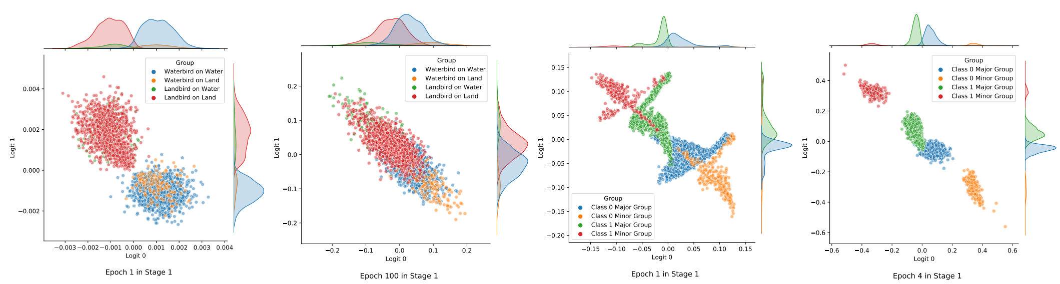
Different Upweight Factors Since Colored-MNIST is a relatively simple dataset for addressing spurious correlation problems, the change of worst-group performance is fairly tiny by varying the upweight factor $\mu \in \{2, 5, 10, 20, 50\}$. Appendix A.4 shows one example of the corresponding accuracy with all the μ's in the above set. On Controlled-Waterbirds, the trend is more obvious. In general, as μ increases, the worst-group test accuracy first rises and then drops. However, over the set $\mu \in \{10, 50, 70, 100, 120, 150\}$ that we tested, it is robust to μ within a large range from 50 to 150. Figure 5, Figure 6 and Table 2 show such robustness.

Table 2: Experiments on Controlled-Waterbirds dataset. Parameters and hyperparameters are $d = \{1800, 200, 200, 1800\}$, $\lambda = 50$.

T	Upweight factor μ					
	10	50	70	100	120	150
10	74.4	84.6	86.1	84.6	84.6	84.6
20	72.2	84.6	87.2	86.1	87.2	84.6
50	63.9	86.1	86.1	88.9	86.1	86.1

Different Stage 1 Epochs In Figure 10 and Figure 11, KDE-plot shows how the values of auxiliary variables changes as Stage 1 proceeds. Carefully controlling all other parameters and hyperparameters, we observe some clear patterns below on these two datasets: First, as T grows, the logits become larger and increasingly influence the model prediction outputs. Second, minority and majority groups are easier to be separated from each other, resulting in better error sets.

In summary, a larger T will generally lead to a better error set, while the model prediction will be wrong on majority groups if T is too large.



(a) (b) (c) (d)

Figure 7: (a)(b) Epoch 1 and 100 in Stage 1 on Waterbirds. Logit 0 corresponds to the prediction on the waterbird class, and logit 1 corresponds to landbird. The group sizes are 1800, 200, 200, 1800 in order. (c)(d) Epoch 1 and 4 in stage 1 on Colored-MNIST. Logit 0 corresponds to the prediction on class 0, and logit 1 corresponds to class 1.

6 Conclusion

In this paper, we tested and analyzed BAM on two new datasets and verified the robustness of BAM varying a large range of parameters and hyperparameters (e.g., the total data size, class imbalance ratio, and group imbalance ratio, μ, T). In addition, we demonstrate the effectiveness of auxiliary variables by varying values of λ. We also validate the negative correlation between ClassDiff and worst-group validation accuracy, which indicates its potential to supplant validation group annotations for less burden.

For future work, we will

- Look deeper into the effects of different losses in Stage 1.
- Validate BAM on more complex datasets such as CIFAR10-MNSIT, and test ClassDiff on multi-classification datasets.
- Aalternate the second stage of BAM with the method proposed by [40], and analyze whether this could improve the performance of the second stage.

References

[1] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[2] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[3] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019.

[4] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. *arXiv preprint arXiv:2006.15766*, 2020.

[5] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.

[6] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

[7] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

[8] Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

[9] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[11] Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. *arXiv: Learning*, 2020.

[12] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[13] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022.

[14] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[15] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

[16] Fereshte Khani, Aditi Raghunathan, and Percy Liang. Maximum weighted loss discrepancy. *arXiv preprint arXiv:1906.03518*, 2019.

[17] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

[18] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021.

[19] Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. *arXiv preprint arXiv:2202.03418*, 2022.

[20] Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, 26(2):274–306, 2020.

[21] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

[22] Gaotang Li, Jiarui Liu, and Wei Hu. Bias amplification improves worst-group accuracy without group information. *Artificial life*, 26(2):274–306, 2020.

[23] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR, 18–24 Jul 2021.

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[25] Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

[26] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

[27] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022.

[28] Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

[29] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

[30] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[31] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 13–18 Jul 2020.

[32] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

[33] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, 2020.

[34] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Towards debiasing nlu models from unknown biases. *arXiv preprint arXiv:2009.12303*, 2020.

[35] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[36] Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[37] Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet, Timothy J Hazen, and Alessandro Sordoni. Increasing robustness to spurious correlations using forgettable examples. *arXiv preprint arXiv:1911.03861*, 2019.

[38] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. *arXiv preprint arXiv:2201.00299*, 2022.

[39] Jingzhao Zhang, Aditya Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. *arXiv preprint arXiv:2010.12230*, 2020.

[40] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

A Appendix

A.1 Team contributions

Name	Contributions
Jiarui Liu	Designed, ran experiments; analyzed results; Visualized figures; Wrote Experiments and Discussions Sections.
Muzhe Wu	Visualized figures; wrote Abstract, Introduction, Related Works, Methodology (adaption), and Conclusion Sections.

A.2 Prior Dataset Specifications

Table 3: Group counts for training sets of Waterbirds, CelebA, CivilComments-WILDS, and MultiNLI. We can observe that these datasets exhibit clear class and group imbalance.

Dataset	Group Description	Group Size	Dataset	Group Description	Group Size
Waterbirds	Waterbird in water (WW)	1057	CelebA	Blond and female	22880
	Waterbird in land (WL)	56		Blond and male	1387
	Landbird in water (LW)	184		Not blond and female	71629
	Landbird in land (LL)	3498		Not blond and male	66874
CivilComments-WILDS			MultiNLI	Contradiction with negation	57498
	Toxic with identity	17784		Contradiction without negation	11158
	Toxic without identity	12731		Entailment with negation	1521
	Non toxic with identity	90337		Entailment without negation	67376
	Non toxic without identity	148186		Neutral with negation	1992
				Neutral without negation	66630

A.3 ClassDiff Visualization

We generate a random subset of experiments with size 10 and visualize their relations between ClassDiff and worst-group validation accuracy. First, we sort all the experiment logs on Greatlakes by their starting training time. Then, fix 42 as the random seed of Numpy and generate a random indices list with size 10. We select the experiments by these indices and plot the ClassDiff relations to avoid cherry picking.

A.4 Upweight Factors Visualization

Table 4 shows that Colored-MNIST does not have clear differences when vary upweight factors.

Table 4: Experiments on Colored-MNIST dataset. Parameters and hyperparameters are $s = 50000, fr = \{0.1, 0.1\}, cr = 1.0, \lambda = 50$.

		Upweight factor μ				
		2	5	10	20	50
T	10	92.9	95.3	95.9	95.5	96.3
	20	93.1	95.1	96.3	95.9	96.3
	30	92.7	94.9	95.9	95.9	96.3
	40	93.1	94.7	95.9	95.9	96.7
	50	93.4	94.9	94.7	95.7	96.5

A.5 Different Epochs visualization

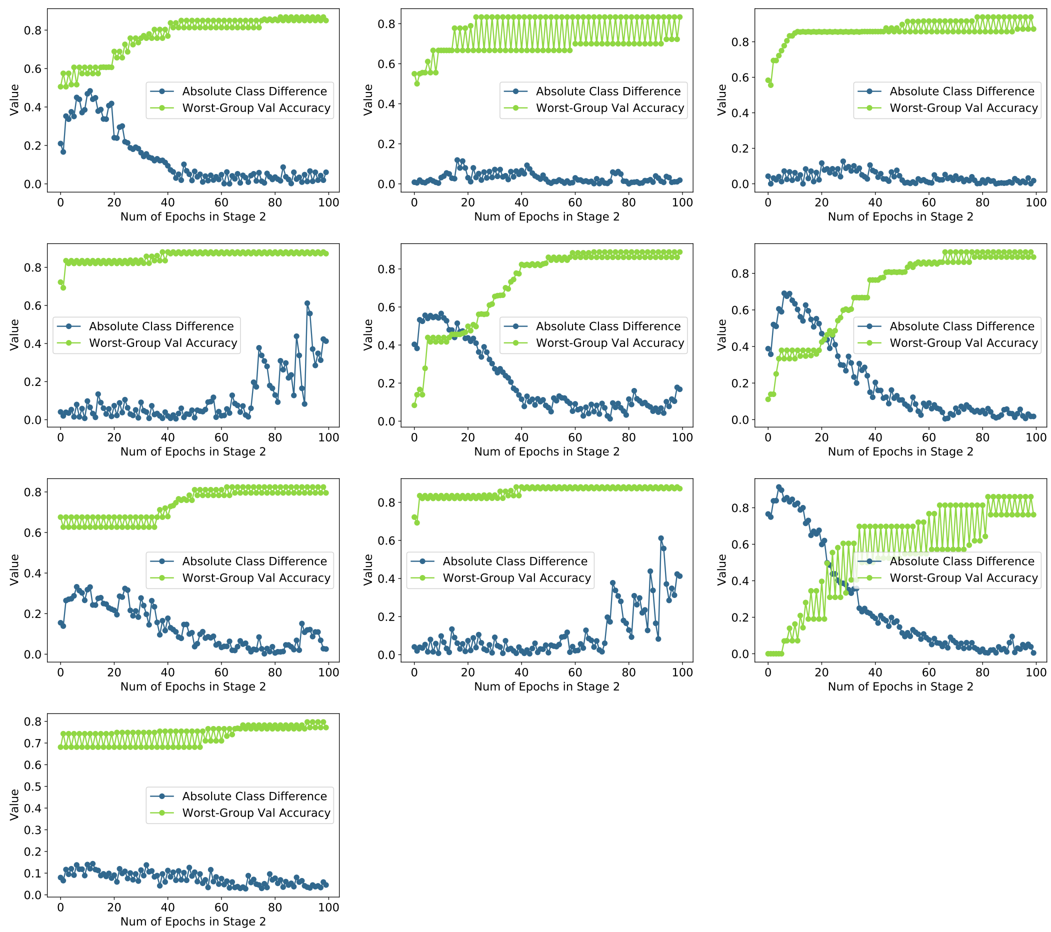


Figure 8: 10 random experiments of the relations between ClassDiff and worst-group validation accuracy on Controlled-Waterbirds.

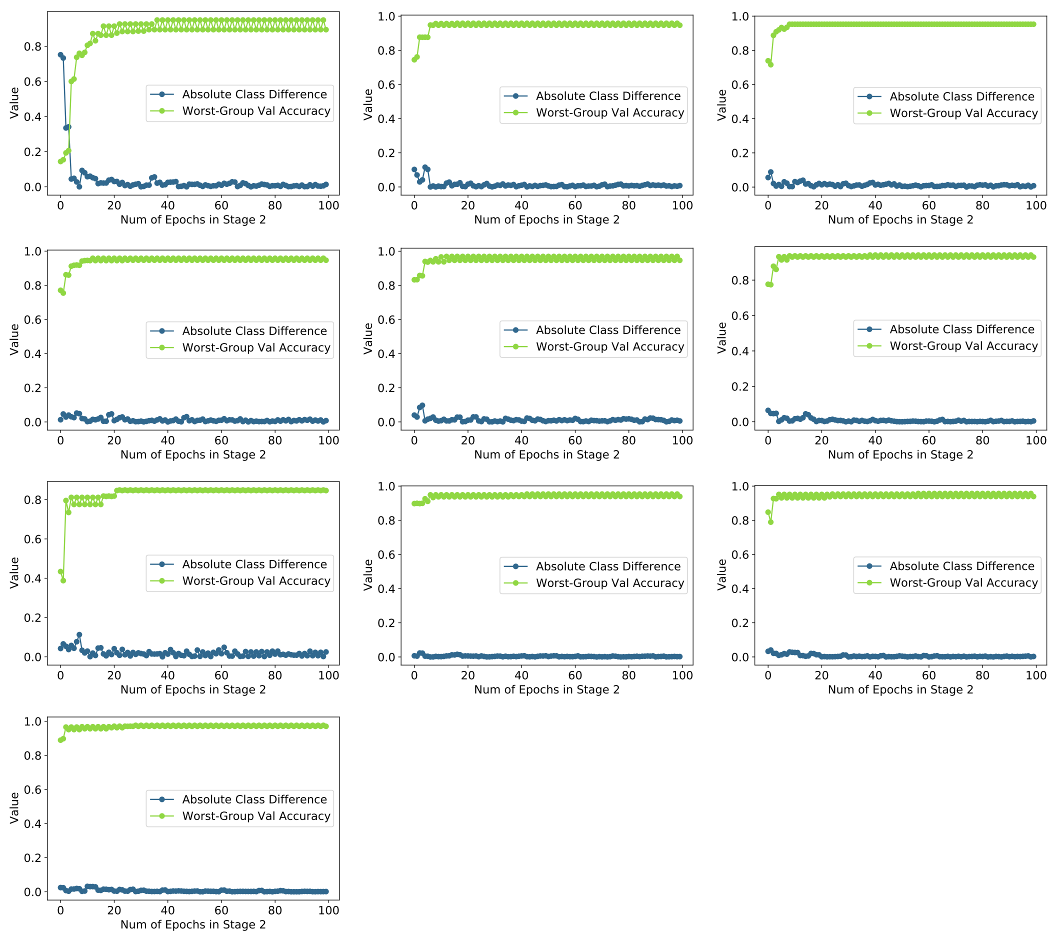


Figure 9: 10 random experiments of the relations between ClassDiff and worst-group validation accuracy on Colored-MNIST.

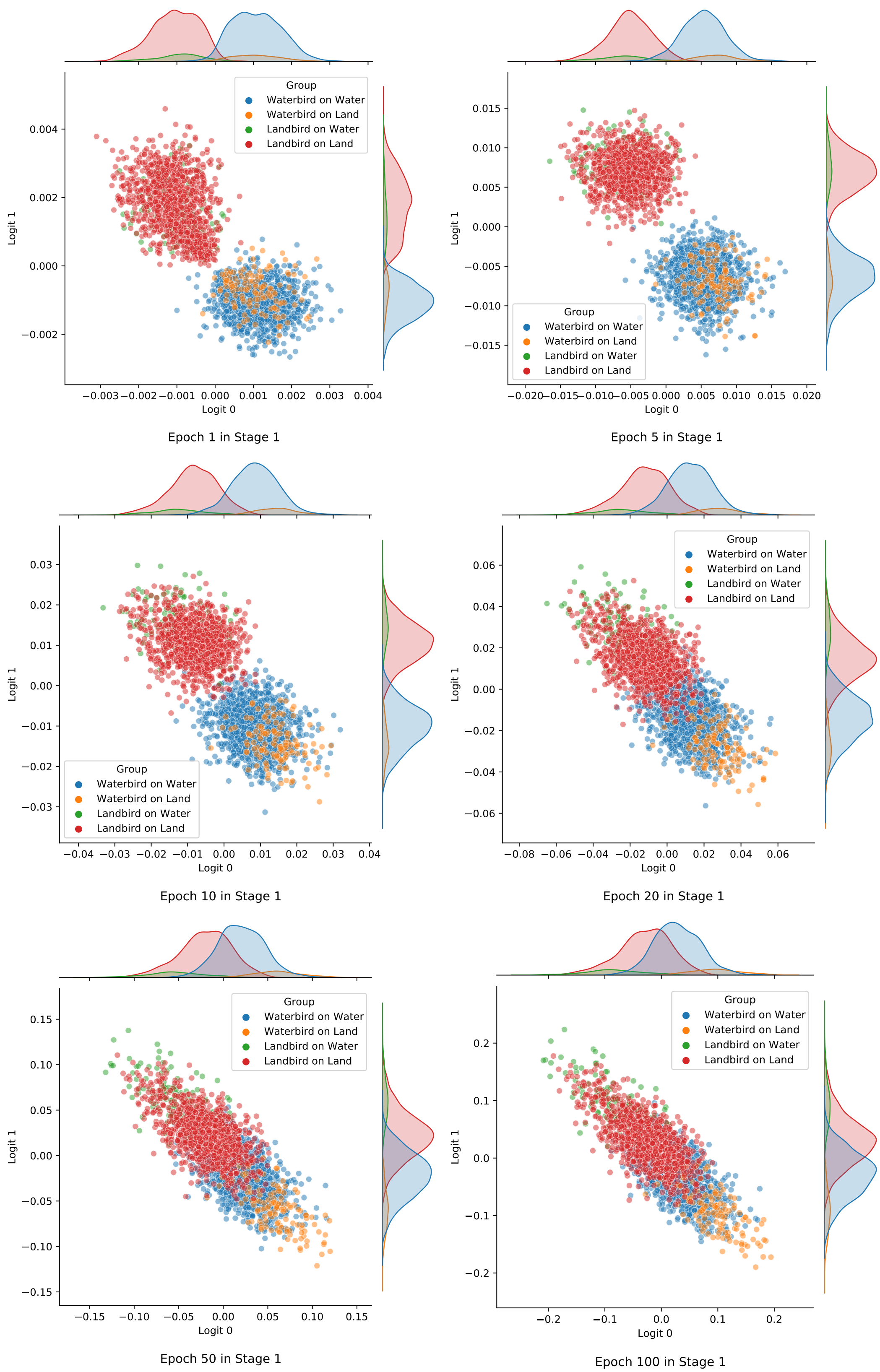


Figure 10: Different epochs in Stage 1 on Waterbirds. Logit 0 corresponds to the prediction on the waterbird class, and logit 1 corresponds to landbird. The group sizes are 1800, 200, 200, 1800 in order.

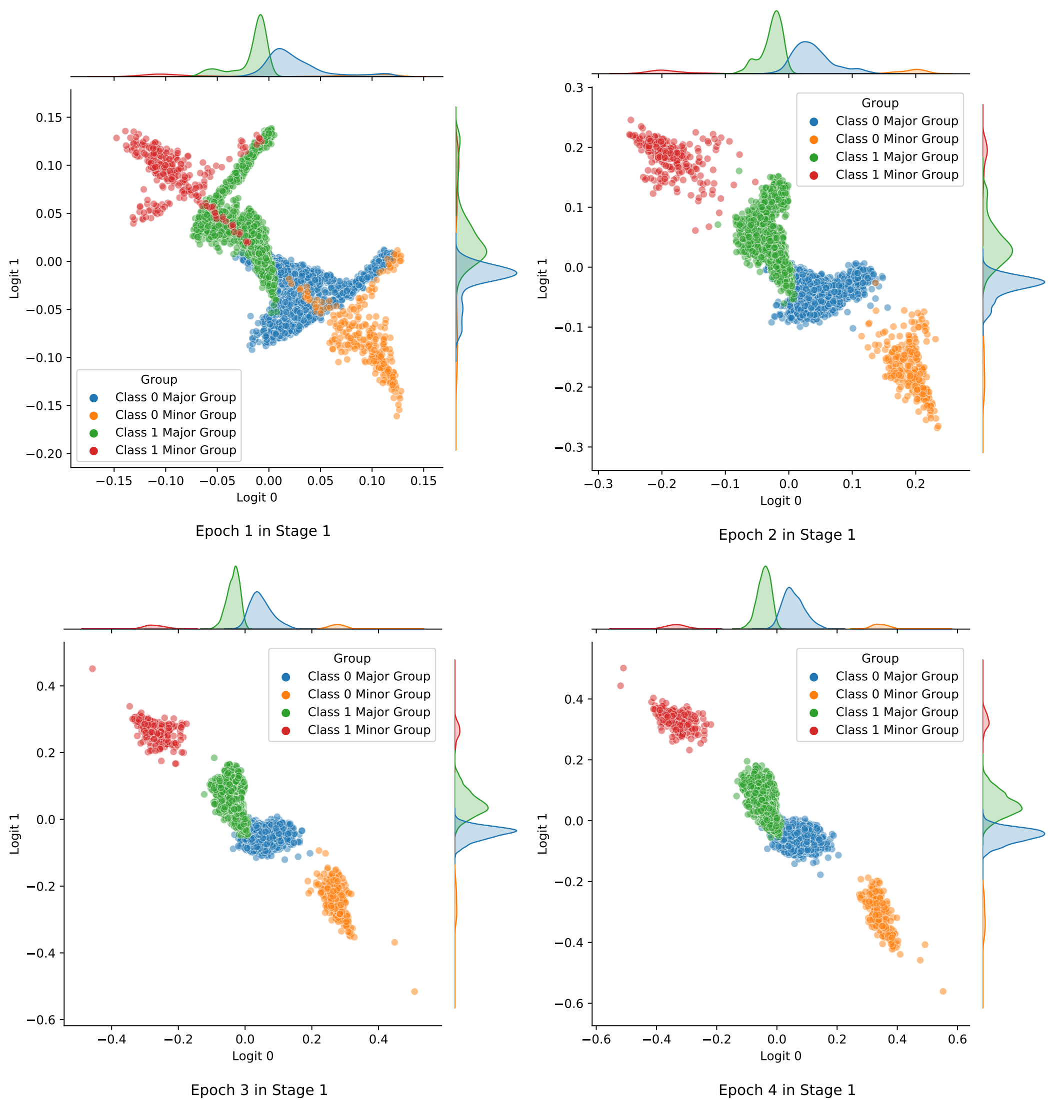


Figure 11: Different epochs in Stage 1 on Colored-MNIST. Logit 0 corresponds to the prediction on class 0, and logit 1 corresponds to class 1.