M.S. THESIS

Rethinking a Decision Boundary and Penalizing Majority Groups in Imbalanced Learning

예측 경계에 대한 고찰과 불균형 학습에서의 과반수 그룹 불이익을 통한 성능 향상

Youngseok Yoon

February 2023

Department of Electrical and Computer Engineering
College of Engineering
Seoul National University

Rethinking a Decision Boundary and Penalizing Majority Groups in Imbalanced Learning

예측 경계에 대한 고찰과 불균형 학습에서의
과반수 그룹 불이익을 통한 성능 향상

지도교수 이 정 우

이 논문을 공학석사 학위논문으로 제출함

2023년 2월

서울대학교 대학원

전기 정보 공학부

윤 영 석

윤영석의 공학석사 학위 논문을 인준함

2023년 2월

위 원 장 _____

부위원장 _____

위 원 _____

Rethinking a Decision Boundary and Penalizing Majority Groups in Imbalanced Learning

Advisior: Jungwoo Lee

Presented to the Graduate School of Seoul National University in Partial Fulfillment of the Requirements for

The Degree of Master

February 2023

Youngseok Yoon

Department of Electrical and Computer Engineering

College of Engineering

Seoul National University

Confirming the master's thesis written by

Youngseok Yoon

February 2023

Chair _____

Vice Chair _____

Examiner _____

Abstract

Establishing a class-balanced dataset on a large scale for stable learning is impractical for real-world problems, as collecting samples is extremely hard for certain categories or groups. This imbalance results mainly from the natural characteristics of these minorities and the hierarchical structure of underlying attributes. It causes a disparity or unfairness in performance among groups. Several existing approaches encourage a model to pay equal attention to all groups by resampling or reweighting minority groups. Also, data augmentation or generative methods have been used to resolve this problem and improve generalization performance. However, all these methods fail to eliminate the negative impacts of overfitting caused by the lack of diversity in minorities. In this paper, we first demonstrate the classifier's tendency to be over-confident in its predictions. Then we propose a novel post-processing method called Prediction Penalty that places a penalty on majorities to enhance the performance of minority groups in terms of accuracy. It is compatible with other methods, and we introduce an adaptive algorithm to find the best-performing penalty function. Our approach suggests a novel perspective on making a decision boundary robust to data imbalance and bias. Experimental results on various datasets and imbalance settings show significant performance enhancement in both average and robustness and demonstrate the benefit of the new robust decision boundary for imbalanced learning.

Keyword: imbalanced learning, post-processing, image classification

Student Number: 2021-27158

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Recent advances in machine learning and deep learning require a vast amount of training data to enable reasonable performance in the sense of generalization by avoiding overfitting. It heavily depends on emerging techniques to collect and manage many training samples in fields like computer vision, natural language processing, and reinforcement learning. However, collecting a balanced dataset in the real world is not practical, and a balance between populations of categories usually collapses during collecting samples. This problem often limits the utilization of recent algorithms that assume the balance among categories. Various works revealed the necessity of balancing categories for fair and robust model performance [2, 3, 13, 17, 24, 25, 33, 35]. Recent approaches have begun to consider this imbalance to deal with real-world problems in practical domains [4, 8, 30, 40].

The category imbalance occurs due to several reasons in the real world. First, some categories are naturally rare, and it is challenging to secure enough samples belonging to these minority groups. For example, pictures of some marine animals, such as whales and dolphins, are more brutal to be obtained than cats and dogs, and there are extreme cases

in reinforcement learning scenarios that rarely occur with fewer populations. Furthermore, as the number of categories increases to build a more challenging classification task such as ImageNet-1K [6], the appearance of so-called tail categories is inevitable [33]. Recent cases are more extreme, like the ImageNet-21K [26] task containing a total of 21,841 categories. Second, the hierarchical structure among attributes causes an imbalance between categories when these attributes are correlated. This problem has been discussed in various approaches concerning fairness, where specific attributes have spurious correlations [8, 30, 38]. Although we wish the model not to consider race when it predicts hair color, correlation in natural data tends to deceive the model. This imbalance in training data often leads to unfair predictions.

The lack of training samples in minority groups usually results in inferior performance in two ways, depending on the degree of imbalance. In the extreme case, the model ignores minority groups during training and fails to predict minority samples. On the other hand, for relaxed imbalance, it overfits the minority samples in the training dataset and cannot provide reasonable decision boundaries for test samples. Although these different cases are distinguishable by comparing training errors in the minority groups, the overall performance in the training set is superior while the generalization performance is poor for both cases. Consequently, the performance of minority groups tends to be disastrous, and it is impossible to utilize these models for minority groups.

1.1 Related Work

1.1.1 Imbalanced Learning and Robust Training

Beyond the superior performance of recent deep neural networks, reliability receives broad attention for practical applications. Much research dedicates to robust machine learning techniques in an imbalanced or polluted dataset. This paper mainly focuses on an imbalanced training dataset and releasing its effects on the classifiers.

Typically there are three lines of approaches to dealing with an imbalanced dataset: pre-processing, in-processing, and post-processing. Pre-processing methods balance the number

of samples among categories by adjusting the sampling rate of each category [3, 17, 28, 29, 35] or generating samples from minority distribution [4, 36, 40]. In-processing approaches use a regularization objective to modify the decision boundary of classifiers for a better generalization to unseen samples in minority groups [5, 8, 23, 27, 30, 37, 38]. Few methods post-process the classifier predictions to take account of the category imbalance [10, 11] or unfairness [15, 20, 21, 34]. They modify the output of a model to achieve balance or fairness, and common methodologies adjust decision boundaries for each subgroup based on the fairness objective.

The most commonly used methods for dealing with imbalance are based on data resampling. Simple methods that have performed well include oversampling from the minority groups [18] and undersampling from the majority groups [14]. A similar approach is to reweight the loss of samples from different subgroups [39]. Although we categorized resampling into a pre-precessing method and reweighting into an in-processing method, they share a similar idea of repopulating a training dataset. Among these resampling/reweighting methods, sampling from each subgroup with equal probability is known to outperform others [2].

Recent works introduce optimization methods that are robust to class imbalance. GDRO replaces the ERM objective with a distributionally robust optimization objective so that the model maximizes the worst-case subgroup accuracy instead of the average [30]. SGDRO maximizes the worst-case class-conditional subgroup accuracy to consider cases with class-subgroup hierarchy [8]. SGDRO becomes reweighting each subgroup equally when there is no class-subgroup hierarchy and subgroups serve as categories.

1.1.2 Fairness in Machine Learning

Fair prediction is also required for reliable applications of machine learning techniques. Since empirical data frequently contain unfair bias and can impair performance and cause social problems, algorithms for training a fair classifier have developed increasingly. While there are many notations and metrics to measure fairness in machine learning, disparate impact [7], equalized odds [9], and equal opportunity [1] are mainly used to estimate the fairness

of classifiers. Fairness in classification tasks aims to balance errors across subgroups, i.e., genders or racial groups. Definitions of other forms of fairness can be found in [31]. In this paper, we do not consider fairness measures as major metrics. However, we use datasets with correlated attributes to consider spurious correlation and test our algorithms in an attribute-based classification task.

More recently, [32] pointed out that most methods improve fairness measures by worsening the classification accuracy for both majority and minority groups. [40] claimed that using the accuracy of minority groups is a more reliable measurement of fairness than previous definitions. We compare both average accuracy and minority group accuracy (Robust accuracy) to demonstrate the robustness of algorithms. Moreover, a discrepancy between maximum and minimum category accuracy is also used instead of previous fairness measures.

1.2 Contributions

Besides manipulating the training procedure, we propose adjusting classifiers' decision boundaries. The tendency of over-parameterized neural networks to be over-confident in their predictions provides a rationale for this adjustment. Our method expands the region occupied by minority groups in the decision space and reduces that of majority groups, encouraging the classifier to generalize better to minorities. It post-process the model predictions and does not affect the training dynamics. Our experimental results demonstrate that this modification does not harm the performance in overall groups.

This paper proposes a novel paradigm to adjust a decision boundary to penalize the predictions on majority groups. While adjusting decision boundaries is not unique, our method exploits the population of a training dataset to compute a statistic vector. It penalizes the model prediction using it as shown in Fig. 1.1. Ours is compatible with other methods and does not affect the stability of training algorithms, as the penalty quantity is independent of the classifier and training dynamics. Also, our adaptive penalizing algorithm efficiently determines the most effective penalty function without additional hyperparameter search. It only requires a small set of validation samples which is possibly imbalanced. Our

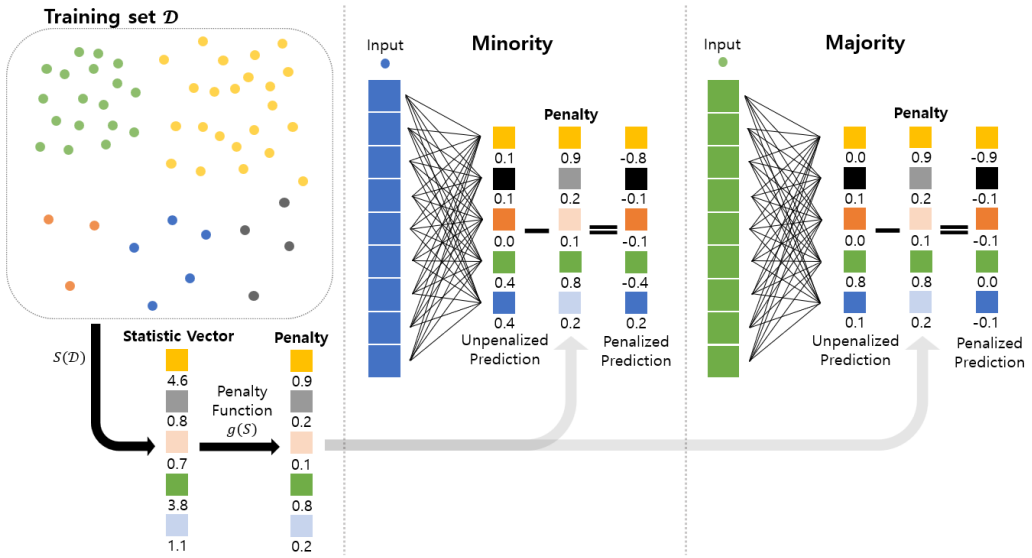


Figure 1.1. The Diagram of our Prediction Penalty method. After the training ends, the model prediction is penalized using the statistic vector computed from a training set. Once the statistic vector and the penalty quantity are computed, they are used with little computation overhead, requiring no additional computation for each sample. The model should be strongly sure to classify as majorities since these categories have a massive penalty. On the other hand, a little penalty is exerted on minority categories.

method demonstrates an impressive performance gain in the average and robust sense for various imbalanced environments. Our work is different from previous works in that it adaptively adjusts the decision boundaries based on the populations of the training set rather than using a fixed modification method from a given fairness measure.

Our contributions are as follows:

- We demonstrate the tendency of over-parameterized networks to be over-confident in their predictions.
- We propose a novel post-processing method called Prediction Penalty that works independently of the training process and requires no additional resources.
- Our extensive experiments on several imbalanced settings and datasets reveal the effect of our adaptive algorithm for the Prediction Penalty method in real-world problems.

Chapter 2

Background

We explain the notations and baseline methods in this chapter. We compare Empirical Risk Minimization (ERM) as a default baseline method. Based on ERM, we suggest Subgroup Resampling (SR) and Distributionally Robust Optimization (DRO) as representatives for pre-processing and in-processing methods, respectively. They differ in the underlying data distribution assumed and optimization strategy. A brief explanation and our realization of these algorithms are provided in the following.

2.1 Notations and Setting

This paper mainly considers a C-way classification task. The goal is to find a parameter θ in a family of parameters Θ that a classifier $f(x;\theta)$ predicts a label $y \in \mathcal{Y}$ provided an input $x \in \mathcal{X}$. The training algorithms aim to optimize an objective under some target distribution P, which is their assumption of the test distribution. Given a loss function $l(y, f(x;\theta))$ that evaluates a model prediction, an objective function is the expectation of this loss function under P.

A set of sampled data points is required to optimize the expectation under P. We denote a dataset $\mathcal{D} = \{(x_i, y_i)|i \in [1:N]\}$ as a collection of data sampled from an empirical distribution $\hat{P}$. The algorithm can not manipulate the empirical distribution $\hat{P}$, but it models P to optimize the objective function.

Both the natural characteristics (true distribution) and the sampled dataset (empirical distribution) can cause a data imbalance. No matter what the reason is, imbalanced learning or training is required when disjoint sub-datasets $\mathcal{D}_c = \{(x_i, y_i)|i \in [1:N], y_i = c\}$ are not balanced where $\bigcup \mathcal{D}_c = \mathcal{D}$. We denote the size of each sub-dataset $\mathcal{D}_c$ as N_c. The distribution of these sizes represents an imbalance in a training dataset.

2.2 Distributionally Robust Optimization

2.2.1 ERM and limitations

ERM assumes that $\hat{P}$ is sampled from P without any bias and directly optimizes a classifier under the empirical distribution. Therefore it samples data points in $\mathcal{D}$ uniformly without any considerations of imbalance. The objective of ERM is

$$\theta_{ERM} = \arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\hat{P}}[l(y, f(x;\theta))], \tag{2.1}$$

where the model is optimized using a naive mean of loss values computed on the dataset $\mathcal{D}$ sampled from $\hat{P}$. As a result, ERM optimizes a classifier to work well on a test distribution identical to an empirical one where data samples for training are obtained.

The overfitting occurs even when train and test distributions are identical because $\mathcal{D}$ cannot represent $\hat{P}$ perfectly. ERM suffers from additional overfitting caused by any possible disparity between the train and test distributions, which usually take the form of an imbalance in a training set. It makes ERM inapplicable when robust performance on minorities is required or minority groups in the training set are no longer minor in the test samples.

There are two problems to consider under the imbalance among categories. First, the

limited samples of minority groups cannot demonstrate the empirical distribution. ERM usually exhibits serious overfitting on the minority groups as it assumes nothing about this imbalance. Next, the test distribution is usually i.i.d. or less imbalanced than the train distribution. It causes a mismatch between ERM's assumption and the actual distribution for test or application.

2.2.2 Group DROs

A training dataset often contains biases and spurious correlations, which lead a model to learn heuristics that work well on most training samples but perform poorly on test data. DRO [8, 30, 38] was introduced to improve the test performance by minimizing the worst-case objective over all potential test distributions. Empirically, the training set is divided into groups, and the worst-case objective is optimized.

Group DRO (GDRO) [30] is an instantiation of DRO on groups which has shown notable performance in generalization for minority groups when used with adequate regularization terms. Groups are organized to share some attributes or belong to the same class. The objective of GDRO is defined as the worst-case group training loss for all predefined groups and is estimated with

$$\theta_{GDRO} \quad = \arg\min_{\theta \in \Theta} \max_{c \in [1,C]} \mathbb{E}_{(x,y)\sim \hat{P}_c}[l(y, f(x;\theta))]. \tag{2.2}$$

More recent work proposes SGDRO [8], which introduces class-subgroup hierarchy to obtain class-conditional subgroup robustness. It encourages the model not to overlook certain classes when subgroups are organized within categories. The objective of SGDRO is

$$\theta_{SGDRO} = \arg\min_{\theta \in \Theta} \mathbb{E}_{y \in Y}\{\max_{c \in C_y} \mathbb{E}_{(x,y)\sim \hat{P}_c}[l(y, f(x;\theta))]\}, \tag{2.3}$$

where C_y is a set of subgroups that samples to class y can belong. When there is no hierarchy, the objective of SGDRO is a mean value of each category's objective, and it is similar to that of subgroup resampling in the next section.

2.3 Subgroup Resampling

Subgroup Resampling (SR) assumes the uniform distribution over the prior distribution of groups or categories, although the empirical distribution $\hat{P}$ does not seem to be. Accordingly, it requires adjustment in sampling data in the training set. It samples a category following a uniform distribution, then data points according to conditional distribution on it. This strategy stresses all categories equally by compelling the number of samples per category to remain the same. Accordingly, the objective of SR is

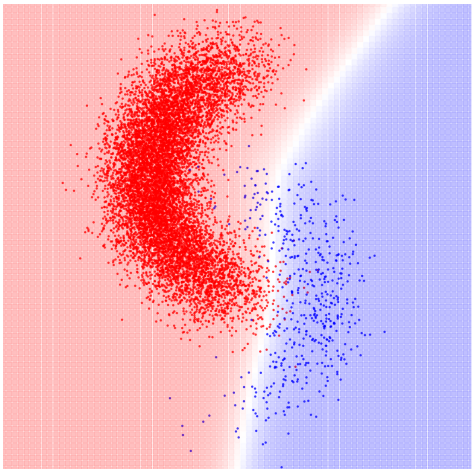
$$\theta_{SR} \quad = \arg\min_{\theta \in \Theta} \mathbb{E}_{P_c \sim U_{\{1,C\}}} \mathbb{E}_{(x,y)\sim \hat{P}_c}[l(y, f(x;\theta))], \tag{2.4}$$

where P_c is a prior over category c and $U_{\{1,C\}}$ is a discreet uniform distribution over categories. Note that SR samples the same number of data points from each sub-dataset. It causes some samples from majority groups not to be selected, while samples from minorities can be sampled multiple times. Sample complexity differs from that of SGDRO when there is no category-subgroup hierarchy while they have similar objective forms.

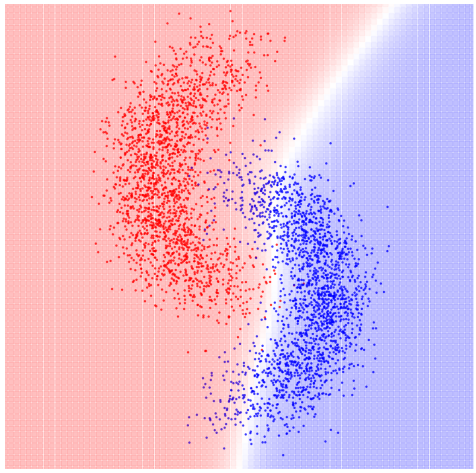
We implement SR by adjusting each sample's sampling rate according to the training set's group population. The sampling ratio of a data point in $\mathcal{D}_c$ is $\frac{1}{CN_c}$ while it is sampled by $\frac{1}{N}$ for ERM. More general methods [3, 17, 28, 29] adjust the sampling rate sample-wise. However, we use this simplified resampling strategy for comparison as sample-wise resampling or reweighting methods usually require additional computational costs and resources.

2.4 Over-Confidence of Neural Networks

While over-parameterized neural networks represent impressive performance on various classification tasks, the problem of over-confidence has arisen for their applications. Besides their success in overall performance, the prediction for each sample does not stand for the likelihood that it belongs to the categories. Fundamentally, these models are trained to output a one-hot vector when the cross-entropy loss is used as a criterion, and this strategy fails to consider the ambiguity of underlying data distribution. Accordingly, a user cannot



(a) An imbalanced train dataset. There are 10000 and 500 points in majority and minority, respectively.



(b) A balanced test dataset. The classifier fails to demonstrate robust performance on balanced datasets due to biased decision boundaries.

Figure 2.1. The decision surface of the classifier on an imbalanced synthetic dataset. The red and blue points demonstrate the majority and minority, respectively. The thickness of the color in the background represents the prediction confidence of the model. The model is almost perfectly sure about its predictions except for the vicinity of the decision boundary. While the classifier fails to classify many minority samples, it is over-confident in the wrong predictions.

estimate the danger of utilizing the model. Both the nature of the cross-entropy loss and the massive model capacity are responsible for this problem, and it becomes more severe for the prediction of minority groups since it affects the decision boundaries of the models.

Fig. 2.1 illustrates the behavior of a classifier and the decision boundary on a synthetic dataset when the imbalance in a training set is present. The ratio between the majority and minority categories is 20, and the simple model exhibits superior average performance. The plane's thickness represents the classifier's confidence in its prediction. The classifier is very confident in its prediction except for the vicinity of the boundary, even though it failed to work on the training set as shown in Fig. 2.1a. Moreover, the model provides a more favorable decision boundary to the majority group that divides the region of the minority group. Accordingly, Fig. 2.1b shows the failure of generalization on balanced test distribution. The model should consider the imbalance in a training dataset and adjust its boundary for safer prediction. We introduce a novel method to eliminate this bias by penalizing the model predictions when the imbalance is present in a training dataset.

Chapter 3

Prediction Penalty

This section introduces our Prediction Penalty (P.P.) method to exert penalties on model predictions. Penalties are computed using a statistic vector and a penalty function. This section demonstrates our method's process, components, and adaptive algorithm.

3.1 Process of Prediction Penalty

To mitigate the negative impacts of data imbalance, we compute the statistic vector $S(\mathcal{D})$ once the training is over or before the training begins. Although it should be computed after the training for each epoch or batch ends when it is a function of training dynamics such as validation error or performance, we propose to use a simpler value such as group population and demonstrate that it works well. In that case, the timing to compute the statistic vector does not matter as it is neither used nor adjusted during training. It is a characteristic of the training dataset that does not change. Also, this vector is computed category- or group-wise rather than sample-wise in our instantiation, making the computation trivial.

When the model infers, an unprocessed prediction is penalized using this M-dimensional

statistic vector $S(\mathcal{D})$ and a penalty function $g : \mathbb{R}^M \to \mathbb{R}^C$ as

$$PP[\hat{y} = c|x;\theta] = P[\hat{y} = c|x;\theta] - g_c(S(\mathcal{D})) \tag{3.1}$$

$$= \frac{e^{f_c(x;\theta)}}{\sum_{i=1}^{C} e^{f_i(x;\theta)}} - g_c(S(\mathcal{D})) \tag{3.2}$$

where the unprocessed prediction is a probability vector computed using a softmax function. The penalty function outputs C-dimensional vector that each dimension represents a category.

Prediction Penalty is a post-processing method with several advantages compared to other pre- or in-processing methods. First, this method applies to other debiasing methods where a training dataset is provided, and a model prediction is generated in the form of probability. Second, there is little computation overhead for computing penalties and penalized prediction. The penalties are computed in advance like the statistic vector, or various penalty functions are compared simultaneously to find the best one for validation.

3.2 Statistic Vector and Prediction Function

The statistic vector of a dataset is designed to measure the generalization capacity of each category. Although formal generalization depends on a model's prediction, the generalization (or overfitting) capacity partly depends on a dataset's innate character, such as imbalance, after the model converges. Thus our implementation excludes the model and entirely depends on the training dataset. Many design choices are possible if they can represent the category- or group-wise generalization capacity. Penalizing each category using this vector results in equalized generalization level, removing the imbalance effects from a training dataset. In order words, a penalty measures the superiority of major groups compared to minorities in the training procedure.

As discussed above, the statistic vector can be any measure of the training dataset that quantifies the portion of each sub-dataset $\mathcal{D}_c$. Accordingly, we design the statistic vector to

be computed component-wise on the corresponding sub-dataset $\mathcal{D}_c$ represented as

$$S_c(\mathcal{D}) = s(\mathcal{D}_c), \tag{3.3}$$

where $s(\cdot)$ is a scalar value function of each sub-dataset $\mathcal{D}_c$ and $M = C$. However, the form of the statistic vector is not limited to C-dimensional vector in general. Vector representation can be used to express each category or a single scalar value can be used to represent all categories.

Although we provide the most straightforward method, where each component of the statistic vector is a function of corresponding sub-dataset $\mathcal{D}_c$, other variants to use similarities among samples or pre-trained features are also applicable. However, we demonstrate that our simple method constantly improves classification performance with minimum effort. Also, because of its instability, we do not use real-time generalization measures such as validation loss or accuracy, although they are conceptually applicable.

The little computation overhead required for penalizing enables us to simultaneously assess many variants of penalty functions (in terms of penalty strength). Normalization is the only condition of these penalty functions, as the penalties are exerted to the probability value. Accordingly, we provide an adaptive algorithm that finds the best penalty function on a validation set and then uses it for inference.

3.3 Adaptive Prediction Penalty

We propose an adaptive algorithm for Prediction Penalty that determines the most promising penalty functions based on validation performances, which requires complete separation of train and validation procedure. Our proposal on unconstrained design choices of the statistic vector and the penalty function enables this modified algorithm to be effective. The only requirement for this adaptive algorithm is to track the best penalty function among all possibilities, and it can be seen as an application of validation monitoring.

The adaptive characteristics relieve any additional hyperparameter search, which is one of the most powerful advantages of our work. Algorithm 1 represents the detailed process

Algorithm 1: Adaptive Prediction Penalty

Input: Model $\mathcal{M}$, a set of prediction functions $\mathcal{G}$
Train dataset $\mathcal{D}$, valid dataset $\mathcal{D}_v$

Prepare:
Compute a statistic vector $S(\mathcal{D})$
A set of penalties $\mathcal{P} \leftarrow []$
for g *in* $\mathcal{G}$ **do**
    Compute the penalty $g(S(\mathcal{D}))$
    Put $g(S(\mathcal{D}))$ in P
Train:
while $\mathcal{M}$ *converges* **do**
    $\mathcal{M} \leftarrow \text{Train}(\mathcal{M}, \mathcal{D})$
Find Penalty:
$acc \leftarrow 0$
$penalty \leftarrow$ None
for p *in* $\mathcal{P}$ **do**
    if $acc <$ *Evaluate(*$\mathcal{M}$*,* $\mathcal{D}_v$*,* p*)* **then**
        $acc \leftarrow \text{Evaluate}(\mathcal{M}, \mathcal{D}_v, p)$
        $penalty \leftarrow p$

Output: $\mathcal{M}, penalty$

of our adaptive algorithm, and the example of penalty functions can be found in Sec. 6.1. Algorithm 1 can be modified further to validate more than one statistic vector while it now uses a single statistic vector. As mentioned above, a more complex version may use a statistic vector concerning real-time training dynamics such as validation error or performance and/or sample-wise vectors. In that case, prepare part will be aggregated in find prediction part in Algorithm 1.

Chapter 4

Experiments

As discussed in Chapter 1, two types of imbalance can occur in the real world. First, the number of samples is limited due to the characteristic of the category itself, while there is no or less correlation with other categories. Standard image classification tasks fall into this type. We refer to it as category-based classification and use imbalanced versions of MNIST, FashionMNIST, and CIFAR10 datasets to demonstrate the effect of Prediction Penalty on this setting. We refer to the other type as attribute-based classification, which occurs when underlying attributes for categorization have correlations. In this case, some attributes are highly correlated to specific categories, fooling the classifiers into basing these attributes rather than actual features for their inference. We use Waterbird and CelebA datasets for attribute-based classification.

4.0.1 Baselines

We compare the Prediction Penalty method to SR and GDROs. However, SR and GDROs can be combined since they are pre-processing and in-processing methods. Accordingly, we

categorize methods into ERM, GDRO, and SGDRO, then compare performance with/without SR and Prediction Penalty. We test each optimization strategy on these four settings and demonstrate how our Prediction Penalty method enhances previous methods.

The performance of GDRO depends heavily on the group adjustment parameter, which benefits the minority group in the optimization procedure introduced in [30]. We use the best result from various group adjustment parameter choices. The detailed performance on each value that is not provided in the main text due to the page limit can be found in Sec. 6.2.1.

While SGDRO finds the worst group for optimization class-wise, there is only one subgroup per category in our setting. Accordingly, SGDRO works by averaging group-wise loss, and the form of the objective function is similar to the subgroup reweighting strategy, where the loss of each sample is reweighted to balance the population. We note the difference from SR as there is no random sampling, and the performance of ERM + SR is usually worse than that of SGDRO.

The most simple version of the Prediction Penalty method is used for comparison. We use the sub-dataset population as the statistic vector and the adaptive Prediction Penalty algorithm with penalty functions introduced in Sec. 6.1.

4.0.2 Metrics

Comparing average accuracy for classification has a limitation when a huge discrepancy between the majority and minority performance due to a severe imbalance in a training dataset exists. Well-trained classifiers should be able to classify an input fairly among categories without bias. We demonstrate robust accuracy as the minimum accuracy among categories and the accuracy gap as the discrepancy between the maximum and minimum category accuracy. These metrics represent the model's ability to operate robustly under the imbalanced training set.

We report the averaged performance over three runs per case, and a 95% confidence interval is also reported. The best value for each metric is highlighted in bold, and the values contained in the most narrow confidence from the best values are also highlighted.

Datasets	Imbalance Type	Imbalance Population
MNIST	Tailed Minority	$[10, 20, 40, 80, 150, 250, 500, 1000, 2000, 5000]$
	Group Imbalance	$[25, 25, 25, 25, 25, 5000, 5000, 5000, 5000, 5000]$
CIFAR10	Tailed Minority	$[225, 225, 450, 450, 900, 900, 1800, 2700, 3600, 4500]$
	Group Imbalance	$[225, 225, 225, 225, 225, 4500, 4500, 4500, 4500, 4500]$

Table 4.1. Imbalance populations for category-based classification. The figures are artificially determined to represent a strong imbalance. Results for these imbalances are provided in Tab. 4.2 and Tab. 4.3. Comparisons for relaxed settings are provided in Sec. 4.3.

Higher average and robust accuracy are promising, while a lower accuracy gap is better.

4.1 Category-based Classification

Category-based classification is a standard setting where each category is independent. No correlation among categories is assumed for this problem. In this case, an imbalance in the training dataset occurs independently in each category.

4.1.1 Datasets

We use imbalanced versions of MNIST, FashionMNIST, and CIFAR10 datasets for category-based classification. Each task is a 10-way image classification following the setting of the original datasets.

Two types of imbalance settings are considered for category-based classification: tailed minority and group imbalance. For the tailed minority, the number of samples per category is reduced gradually. It is a general scenario where the number of possible samples per category is independent and depends on the category's characteristics. On the other hand, half of the categories belong to the majority group, while others belong to the minority group in the group imbalance setting. It is also an applicable situation where different datasets are combined for an augmented classification task. The detailed population is demonstrated in Tab. 4.1 for each dataset. We also provide results on the degree of the imbalance for each case in Sec. 4.3.

			Average Acc.	Robust Acc.	Accuracy Gap.
Tailed Minority	ERM		0.827 ± 0.018	0.556 ± 0.042	0.429 ± 0.041
		+ P.P.	0.871 ± 0.006	0.704 ± 0.028	0.269 ± 0.023
		+ SR	0.877 ± 0.021	0.791 ± 0.026	0.167 ± 0.039
		+ SR + P.P.	0.885 ± 0.025	0.814 ± 0.016	0.140 ± 0.023
	GDRO		**0.909 ± 0.006**	0.847 ± 0.009	0.105 ± 0.010
		+ P.P.	**0.906 ± 0.011**	**0.858 ± 0.013**	**0.088 ± 0.007**
		+ SR	0.900 ± 0.010	0.811 ± 0.031	0.151 ± 0.041
		+ SR + P.P.	**0.909 ± 0.008**	**0.863 ± 0.008**	**0.088 ± 0.012**
	SGDRO		0.881 ± 0.023	0.755 ± 0.072	0.199 ± 0.064
		+ P.P.	0.894 ± 0.021	0.798 ± 0.050	0.166 ± 0.048
		+ SR	0.884 ± 0.018	0.776 ± 0.034	0.195 ± 0.033
		+ SR + P.P.	0.893 ± 0.021	0.805 ± 0.031	0.162 ± 0.026
Group Imbalance	ERM		0.777 ± 0.004	0.465 ± 0.018	0.527 ± 0.015
		+ P.P.	0.901 ± 0.002	0.829 ± 0.026	0.126 ± 0.020
		+ SR	0.860 ± 0.005	0.715 ± 0.030	0.234 ± 0.022
		+ SR + P.P.	**0.913 ± 0.004**	**0.862 ± 0.005**	**0.116 ± 0.004**
	GDRO		0.891 ± 0.009	0.818 ± 0.028	0.157 ± 0.022
		+ P.P.	0.896 ± 0.009	0.844 ± 0.020	0.123 ± 0.045
		+ SR	0.898 ± 0.006	0.808 ± 0.020	0.170 ± 0.030
		+ SR + P.P.	0.892 ± 0.016	0.836 ± 0.025	0.125 ± 0.009
	SGDRO		0.877 ± 0.006	0.711 ± 0.068	0.271 ± 0.071
		+ P.P.	0.904 ± 0.003	0.831 ± 0.010	0.139 ± 0.010
		+ SR	0.853 ± 0.009	0.640 ± 0.051	0.343 ± 0.063
		+ SR + P.P.	0.906 ± 0.008	0.848 ± 0.025	0.122 ± 0.037

Table 4.2. Category-based classification results on the imbalanced MNIST dataset. The imbalance ratio is 500 and 200 for the tailed minority and group imbalance, respectively, as provided in Tab. 4.1.

4.1.2 Architecture

We use fully-connected neural networks with two hidden layers, each containing 100 neurons for MNIST variants following [22]. To deal with higher data complexity, ResNet 18 architecture for 32×32 resolution is used for the CIFAR10 dataset [12]. The ReLU activation function is used for all cases. We use Adam optimizer [16] for MNIST and SGD optimizer for CIFAR10, respectively, with a learning rate of 0.001. These settings on model architecture are chosen where it fits on the balanced version and fails to generalize on its imbalanced counterpart.

			Average Acc.	Robust Acc.	Accuracy Gap.
Tailed Minority	ERM		0.496 ± 0.006	0.139 ± 0.032	0.744 ± 0.030
		+ P.P.	0.525 ± 0.007	0.319 ± 0.014	0.450 ± 0.065
		+ SR	0.424 ± 0.028	0.179 ± 0.040	0.513 ± 0.055
		+ SR + P.P.	0.443 ± 0.026	0.272 ± 0.032	0.474 ± 0.063
	GDRO		**0.565 ± 0.023**	0.380 ± 0.038	0.445 ± 0.105
		+ P.P.	**0.561 ± 0.012**	**0.418 ± 0.025**	**0.320 ± 0.062**
		+ SR	0.473 ± 0.040	0.254 ± 0.047	0.502 ± 0.073
		+ SR + P.P.	0.486 ± 0.040	0.323 ± 0.023	0.436 ± 0.047
	SGDRO		0.431 ± 0.047	0.258 ± 0.064	0.405 ± 0.033
		+ P.P.	0.444 ± 0.038	0.270 ± 0.029	0.360 ± 0.067
		+ SR	0.426 ± 0.031	0.225 ± 0.047	0.433 ± 0.129
		+ SR + P.P.	0.434 ± 0.041	0.304 ± 0.026	**0.279 ± 0.059**
Group Imbalance	ERM		0.499 ± 0.007	0.080 ± 0.015	0.815 ± 0.005
		+ P.P.	0.528 ± 0.008	0.386 ± 0.019	0.302 ± 0.041
		+ SR	0.421 ± 0.027	0.165 ± 0.025	0.568 ± 0.092
		+ SR + P.P.	0.456 ± 0.012	0.328 ± 0.006	**0.293 ± 0.006**
	GDRO		0.509 ± 0.052	0.339 ± 0.072	0.341 ± 0.103
		+ P.P.	**0.580 ± 0.011**	**0.447 ± 0.001**	0.351 ± 0.060
		+ SR	0.390 ± 0.028	0.171 ± 0.012	0.501 ± 0.088
		+ SR + P.P.	0.536 ± 0.022	0.388 ± 0.013	0.378 ± 0.006
	SGDRO		0.475 ± 0.014	0.235 ± 0.100	0.535 ± 0.122
		+ P.P.	0.502 ± 0.002	0.352 ± 0.003	0.307 ± 0.020
		+ SR	0.410 ± 0.029	0.149 ± 0.053	0.530 ± 0.082
		+ SR + P.P.	0.454 ± 0.012	0.320 ± 0.019	0.304 ± 0.038

Table 4.3. Category-based classification results on imbalanced CIFAR10 dataset. The imbalance ratio is 20 for both types, as provided in Tab. 4.1.

4.1.3 Results

Tab. 4.2 and Tab. 4.3 show the results of category-based classification on imbalanced MNIST and CIFAR10 datasets. Results on the FashionMNIST dataset and more relaxed imbalance settings are provided in Sec. 6.2. The tables demonstrate our Prediction Penalty method's effects and compatibility with existing robust learning methods.

On the MNIST dataset, our Prediction Penalty method always enhances robust accuracy and reduces the accuracy gap while preserving the average accuracy. While it is widely believed that there is a definite trade-off between average and robust accuracy, it improves both measures in most cases for these multi-way classification problems. The results suggest that ours adjusts the generalization level of majority categories effectively. Moreover, it is

Datasets	Categorization	Population
WaterBird	Land bird in the land, Land bird in the water	3498, 184
	Water bird in the land, Water bird in the land	56, 1057
CelebA	Black hair female, Black hair male	4054, 66874
	Blond hair female, Blond hair male	22880, 1387

Table 4.4. Categorization and imbalance populations for attribute-based classification. The populations are decided in advance, and we undersample "black hair female" category to reveal a more apparent hierarchy.

notable that a simple adjustment to model predictions significantly improves the performance of ERM, and the previous way to determine the model prediction is unsuitable for imbalanced classification.

Our method also improves the robust performance in the imbalanced version of the CIFAR10 dataset. It provides evidence that Prediction Penalty works on more complicated model architecture and data. In fact, it works better for a more complicated classification task. Surprisingly, SR fails to improve the performance of ERM. Although ERM + SR + P.P. shows poor performance on the tailed minority, P.P. works well without SR.

4.2 Attribute-based Classification

Some explicit attributes may be highly correlated to the target value in real-world classification tasks. However, these spurious correlations provide incorrect bias to the classifier and generate unfairness in the prediction. For example, it is unfair to decide one's gender according to their hair color, even if men usually have black hair while women are blond. The bias gets even worse when an imbalance in the populations of women and men presents in a training set. Also, the background of the images may provide some evidence to the classifier, but it should not consider the background and bases its prediction on exact features. Attribute-based classification task considers such problems where attributes are highly but incorrectly correlated to the targets.

			Average Acc.	Robust Acc.	Accuracy Gap.
WaterBird	ERM		0.729 ± 0.011	0.289 ± 0.051	0.700 ± 0.051
		+ P.P.	0.779 ± 0.016	0.693 ± 0.012	0.228 ± 0.020
		+ SR	0.757 ± 0.085	0.654 ± 0.037	0.214 ± 0.064
		+ SR + P.P.	0.815 ± 0.009	0.788 ± 0.023	0.070 ± 0.020
	GDRO		**0.849 ± 0.010**	**0.820 ± 0.002**	0.051 ± 0.014
		+ P.P.	**0.854 ± 0.005**	**0.826 ± 0.010**	0.054 ± 0.022
		+ SR	0.819 ± 0.031	0.786 ± 0.017	0.083 ± 0.033
		+ SR + P.P.	0.829 ± 0.007	**0.820 ± 0.007**	**0.027 ± 0.020**
	SGDRO		0.830 ± 0.009	0.685 ± 0.061	0.253 ± 0.076
		+ P.P.	0.824 ± 0.005	0.807 ± 0.010	0.051 ± 0.010
		+ SR	0.798 ± 0.007	0.655 ± 0.023	0.245 ± 0.028
		+ SR + P.P.	0.812 ± 0.004	0.779 ± 0.019	0.085 ± 0.028
CelebA	ERM		0.850 ± 0.002	0.215 ± 0.013	0.771 ± 0.013
		+ P.P.	0.843 ± 0.005	**0.713 ± 0.013**	0.244 ± 0.014
		+ SR	0.870 ± 0.003	0.644 ± 0.013	0.266 ± 0.021
		+ SR + P.P.	0.808 ± 0.005	0.691 ± 0.033	0.204 ± 0.038
	GDRO		0.875 ± 0.010	**0.742 ± 0.084**	**0.143 ± 0.103**
		+ P.P.	**0.888 ± 0.001**	**0.730 ± 0.019**	**0.171 ± 0.021**
		+ SR	0.860 ± 0.013	**0.717 ± 0.059**	**0.158 ± 0.080**
		+ SR + P.P.	0.879 ± 0.006	**0.738 ± 0.026**	**0.151 ± 0.027**
	SGDRO		0.879 ± 0.003	0.682 ± 0.016	0.231 ± 0.020
		+ P.P.	0.884 ± 0.002	**0.736 ± 0.028**	**0.162 ± 0.030**
		+ SR	0.865 ± 0.002	0.684 ± 0.024	0.222 ± 0.029
		+ SR + P.P.	0.878 ± 0.007	**0.734 ± 0.051**	**0.159 ± 0.055**

Table 4.5. Attribute-based classification results on WaterBird and CelebA dataset. Pre-trained ResNet 50 is used.

4.2.1 Datasets

WaterBird [30] and CelebA [19] datasets are widely used in fairness literature that considers the spurious correlations among attributes. WaterBird is constructed by generating an artificial correlation between an image's object and background. For CelebA, we use "Blond Hair" and "Male" features to decide the dataset's category. We downsample it to stress the spurious correlations among these features. We use these two attributes per dataset to form a 4-way classification task on correlated attributes, while other works usually use a 2-way classification. The populations and detailed categorization can be found in Tab. 4.4.

4.2.2 Architecture

We use ResNet 50 and ResNet 18 architectures [12] for the attribute-based classification task. The detailed configuration differs from that was used for the CIFAR10 since this task's image resolution is 224×224. We use pre-trained weights from PyTorch and torchvision. We use the SGD optimizer with a learning rate of 0.001 and 0.0001 for Waterbird and CelebA, respectively.

4.2.3 Results

Tab. 4.5 demonstrates the results of the attribute-based classification on ResNet 50 architecture. Results on ResNet 18 architecture can be found in Tab. 6.11. Our Prediction Penalty method improves the performance of existing methods in most cases. It improves both average and robust performance, proving compatibility with other methods. ERM's remarkably low robust accuracy stands for the spurious correlations in these datasets, and improved performance demonstrates that ours effectively removes them. It shows that our method is capable of eliminating various imbalances with no modification or assumption.

4.3 More Imbalance Settings for Category-based Classification

We examine the Prediction Penalty method on more imbalance settings for MNIST and CIFAR10 datasets to demonstrate the robustness of our Prediction Penalty method to various imbalance strengths. More settings are provided in Tab. 4.6. The strength of imbalance increases as the imbalance ID decreases.

Fig. 4.1 demonstrates the performance of the Prediction Penalty method on MNIST dataset when the model is optimized using ERM objective. Regardless of imbalance strength, our method significantly improves naive ERM and SR. It is notable that the difference in performance between ERM + P.P. and ERM + SR + P.P. for Group Imbalance setting is trivial. This implies that the effect of SR disappears when the model prediction is penalized. Detailed and additional performance on other datasets and GDROs can be found in Sec. 6.2.2, Sec. 6.2.3 and Sec. 6.2.4.

Datasets	Imbalance Name	Imbalance ID	Imbalance Population
MNIST	Tailed	1	[10, 20, 40, 80, 150, 250, 500, 1000, 2000, 5000]
	Minority	2	[20, 40, 80, 150, 300, 500, 1000, 2000, 5000, 5000]
		3	[50, 100, 200, 400, 800, 1500, 3000, 3000, 5000, 5000]
	Group	4	[25, 25, 25, 25, 25, 5000, 5000, 5000, 5000, 5000]
	Imbalance	5	[50, 50, 50, 50, 50, 5000, 5000, 5000, 5000, 5000]
		6	[100, 100, 100, 100, 100, 5000, 5000, 5000, 5000, 5000]
CIFAR10	Tailed	1	[225, 225, 450, 450, 900, 900, 1800, 2700, 3600, 4500]
	Minority	2	[450, 900, 1350, 1800, 2250, 2700, 3150, 3600, 4050, 4500]
		3	[900, 900, 1800, 1800, 2700, 2700, 3600, 3600, 4500, 4500]
	Group	4	[225, 225, 225, 225, 225, 4500, 4500, 4500, 4500, 4500]
	Imbalance	5	[450, 450, 450, 450, 450, 4500, 4500, 4500, 4500, 4500]
		6	[900, 900, 900, 900, 900, 4500, 4500, 4500, 4500, 4500]

Table 4.6. Additional imbalance populations for category-based classification for MNIST and CIFAR10 datasets. Results for imbalance ID 1 and 4 for MNIST and CIFAR10 datasets are shown in Sec. 4.1, and other results are provided in Sec. 6.2.

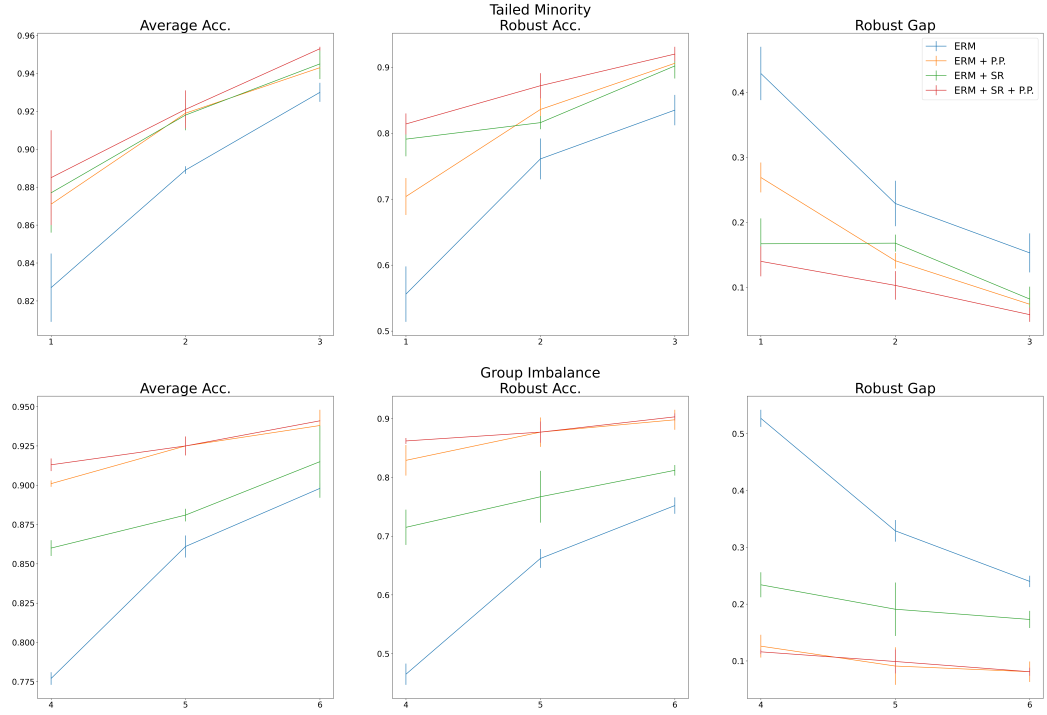


Figure 4.1. The Performance plot on MNIST dataset optimized using ERM objective. The detailed populations of the imbalanced training dataset are provided in Tab. 4.6.

4.4 On Distribution Match between Valid and Test

As our adaptive algorithm determines the best penalty function on the validation set, we confirm whether the superior performance on test data originated from carefully selecting a

valid dataset similar to the test one. We compare the train, valid, and test performance using the best penalty function selected on the validation dataset. Note that training samples are not penalized during training and these results are computed for this comparison.

Tab. 4.7 compares the train, valid, and test performance on WaterBird and CelebA datasets optimized using ERM objective and post-processed using our Prediction Penalty method. The penalty function is selected using the validation performance. It represents that the best-performing prediction function works well on the training and test datasets, which stands for that the choice is not over-fitted to the validation dataset. Additional results on other datasets can be found in Sec. 6.2.6.

			Average Acc.	Robust Acc.	Accuracy Gap.
WaterBird	ERM	train	0.929 ± 0.026	0.904 ± 0.036	0.096 ± 0.036
		valid	0.734 ± 0.064	0.661 ± 0.010	0.299 ± 0.009
		test	0.779 ± 0.016	0.693 ± 0.012	0.228 ± 0.020
	ERM + SR	train	0.906 ± 0.024	0.738 ± 0.057	0.262 ± 0.057
		valid	0.807 ± 0.017	0.793 ± 0.006	0.114 ± 0.005
		test	0.815 ± 0.009	0.788 ± 0.023	0.070 ± 0.020
CelebA	ERM	train	0.860 ± 0.013	0.822 ± 0.017	0.150 ± 0.016
		valid	0.826 ± 0.024	0.793 ± 0.003	0.164 ± 0.002
		test	0.843 ± 0.005	0.713 ± 0.013	0.244 ± 0.014
	ERM + SR	train	0.940 ± 0.007	0.859 ± 0.006	0.140 ± 0.006
		valid	0.874 ± 0.006	0.866 ± 0.002	0.040 ± 0.005
		test	0.808 ± 0.005	0.691 ± 0.033	0.204 ± 0.038

Table 4.7. Performance gap between the train, valid, and test datasets for selected penalty function in the attribute-based classification task. The best function is chosen according to the valid performance, and it works well in training and test data where the data distribution may differ.

Chapter 5

Conclusion

We propose the Prediction Penalty method, a novel post-processing method for bias elimination on a training dataset. We build on the tendency of over-parameterized neural networks to be over-confident in their predictions. It suggests an innovative perspective on the decision boundary of deep neural networks.

Our method penalizes the model predictions using the statistic vector computed from a training dataset. It effectively removes the model bias originating from data imbalance and balances out the generalization level of the predictions. Also, it allows an adaptive algorithm that evaluates multiple penalty functions simultaneously due to an isolated mechanism from the training procedures. Extensive experiments reveal the superiority of our method in various imbalanced settings. It consistently improves previous debiasing methods on the performance of both in average and robustness terms.

Chapter 6

Supplementary

We provide an example of penalty functions and additional results not contained in the main paper due to space limitations. Sec. 6.1 is about the penalty function pool we use for the adaptive algorithm. Sec. 6.2 demonstrates the additional results on imbalances settings and the ablation results.

6.1 An example pool of penalty functions

The only requirement for the penalty function is that it should output normalized values for effective penalizing. In our setting, the penalty gap among elements that exceed 1.0 means the infeasibility of the target. We provide a set of penalty functions for the adaptive prediction penalty algorithm as follows:

$$\mathcal{G} = [\frac{S}{\frac{C}{\sum_{i=1} S_i}}, \frac{S^2}{\frac{C}{\sum_{i=1} S_i^2}}, \frac{S^3}{\frac{C}{\sum_{i=1} S_i^3}}, \frac{S^4}{\frac{C}{\sum_{i=1} S_i^4}}], \tag{6.1}$$

$$\mathcal{G} = [\frac{S}{\max\limits_{i\in[1,C]} S_i}, \frac{S^2}{\max\limits_{i\in[1,C]} S_i^2}, \frac{S^3}{\max\limits_{i\in[1,C]} S_i^3}, \frac{S^4}{\max\limits_{i\in[1,C]} S_i^4}], \tag{6.2}$$

where S is the C-dimensional statistic vector. Note that the statistic vector can be any shape in general.

6.2 Additional Experiments

This section provides additional results that are not provided in the main paper due to space limitations.

6.2.1 Ablation results on group adjustment parameter of GDRO

The performance of GDRO varies by group adjustment parameter. We compare the robust accuracy for four parameter candidates 0.0, 1.0, 2.0, and 5.0. In the main paper, we provide the results from the best-performing parameter. This section demonstrates the detailed performance for varying group adjustment parameters not provided in the main text.

		Tailed Minority 1			Group Imbalance 4		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
GDRO	0	0.876 ± 0.030	0.798 ± 0.049	0.142 ± 0.019	0.872 ± 0.026	0.739 ± 0.073	0.232 ± 0.084
	1	0.909 ± 0.006	**0.847 ± 0.009**	0.105 ± 0.010	0.887 ± 0.001	0.730 ± 0.043	0.245 ± 0.058
	2	0.895 ± 0.008	0.829 ± 0.016	0.142 ± 0.015	0.885 ± 0.044	0.784 ± 0.025	0.181 ± 0.032
	5	0.861 ± 0.010	0.800 ± 0.011	0.159 ± 0.034	0.891 ± 0.009	**0.818 ± 0.028**	0.157 ± 0.022
GDRO + P.P.	0	0.896 ± 0.007	0.832 ± 0.024	0.116 ± 0.028	0.896 ± 0.009	**0.844 ± 0.020**	0.123 ± 0.045
	1	0.906 ± 0.011	**0.858 ± 0.013**	0.088 ± 0.007	0.901 ± 0.015	0.834 ± 0.031	0.136 ± 0.038
	2	0.900 ± 0.006	0.849 ± 0.024	0.113 ± 0.018	0.900 ± 0.005	0.833 ± 0.010	0.138 ± 0.026
	5	0.858 ± 0.015	0.764 ± 0.052	0.210 ± 0.040	0.892 ± 0.006	0.824 ± 0.002	0.149 ± 0.016
GDRO + SR	0	0.878 ± 0.010	0.780 ± 0.014	0.187 ± 0.035	0.824 ± 0.015	0.602 ± 0.056	0.366 ± 0.066
	1	0.900 ± 0.010	**0.811 ± 0.031**	0.151 ± 0.041	0.887 ± 0.007	0.707 ± 0.040	0.276 ± 0.041
	2	0.881 ± 0.016	0.803 ± 0.050	0.145 ± 0.069	0.883 ± 0.024	0.762 ± 0.034	0.195 ± 0.052
	5	0.846 ± 0.004	0.758 ± 0.005	0.191 ± 0.034	0.898 ± 0.006	**0.808 ± 0.020**	0.170 ± 0.030
GDRO + SR + P.P.	0	0.890 ± 0.004	0.826 ± 0.012	0.135 ± 0.040	0.889 ± 0.011	0.824 ± 0.025	0.135 ± 0.040
	1	0.909 ± 0.008	**0.863 ± 0.008**	0.088 ± 0.012	0.890 ± 0.022	0.808 ± 0.042	0.173 ± 0.041
	2	0.891 ± 0.007	0.833 ± 0.042	0.113 ± 0.057	0.892 ± 0.016	**0.836 ± 0.025**	0.125 ± 0.009
	5	0.848 ± 0.003	0.766 ± 0.040	0.195 ± 0.041	0.896 ± 0.005	0.819 ± 0.020	0.161 ± 0.025

Table 6.1. Results on ablation study on group adjustment parameter for MNIST dataset.

		Tailed Minority 1			Group Imbalance 4		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
GDRO	0	0.541 ± 0.035	0.321 ± 0.012	0.441 ± 0.075	0.483 ± 0.035	0.271 ± 0.045	0.426 ± 0.070
	1	0.565 ± 0.023	**0.380 ± 0.038**	0.445 ± 0.105	0.505 ± 0.022	0.318 ± 0.017	0.379 ± 0.114
	2	0.555 ± 0.027	0.355 ± 0.026	0.401 ± 0.033	0.509 ± 0.052	**0.339 ± 0.072**	0.341 ± 0.103
	5	0.509 ± 0.042	0.353 ± 0.014	0.338 ± 0.102	0.499 ± 0.020	0.298 ± 0.034	0.411 ± 0.026
GDRO + P.P.	0	0.559 ± 0.045	0.396 ± 0.014	0.427 ± 0.066	0.580 ± 0.011	**0.447 ± 0.001**	0.351 ± 0.060
	1	0.561 ± 0.012	**0.418 ± 0.025**	0.320 ± 0.062	0.559 ± 0.018	0.429 ± 0.023	0.398 ± 0.039
	2	0.553 ± 0.024	0.387 ± 0.006	0.379 ± 0.069	0.571 ± 0.032	0.428 ± 0.014	0.401 ± 0.030
	5	0.597 ± 0.033	0.384 ± 0.026	0.403 ± 0.044	0.553 ± 0.030	0.414 ± 0.030	0.345 ± 0.063
GDRO + SR	0	0.473 ± 0.040	**0.254 ± 0.047**	0.502 ± 0.073	0.390 ± 0.028	**0.171 ± 0.012**	0.501 ± 0.088
	1	0.452 ± 0.017	0.246 ± 0.065	0.483 ± 0.043	0.418 ± 0.041	0.144 ± 0.076	0.580 ± 0.268
	2	0.461 ± 0.008	0.243 ± 0.026	0.514 ± 0.091	0.429 ± 0.051	0.099 ± 0.048	0.644 ± 0.170
	5	0.475 ± 0.052	0.195 ± 0.043	0.584 ± 0.168	0.417 ± 0.032	0.165 ± 0.071	0.519 ± 0.142
GDRO + SR + P.P.	0	0.527 ± 0.031	0.319 ± 0.028	0.461 ± 0.057	0.536 ± 0.022	**0.388 ± 0.013**	0.378 ± 0.006
	1	0.486 ± 0.040	**0.323 ± 0.023**	0.436 ± 0.047	0.506 ± 0.012	0.372 ± 0.036	0.372 ± 0.035
	2	0.504 ± 0.052	0.282 ± 0.038	0.551 ± 0.012	0.506 ± 0.044	0.339 ± 0.006	0.419 ± 0.042
	5	0.551 ± 0.010	0.290 ± 0.013	0.546 ± 0.034	0.514 ± 0.014	0.354 ± 0.027	0.342 ± 0.058

Table 6.2. Results on ablation study on group adjustment parameter for CIFAR10 dataset.

		WaterBird			CelebA		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
GDRO	0	0.813 ± 0.028	0.722 ± 0.036	0.161 ± 0.059	0.884 ± 0.002	0.719 ± 0.011	0.183 ± 0.011
	1	0.848 ± 0.035	0.773 ± 0.009	0.116 ± 0.027	0.883 ± 0.010	0.682 ± 0.010	0.218 ± 0.026
	2	0.837 ± 0.009	0.752 ± 0.036	0.121 ± 0.032	0.882 ± 0.005	0.719 ± 0.034	0.175 ± 0.044
	5	0.849 ± 0.010	**0.820 ± 0.002**	0.051 ± 0.014	0.875 ± 0.010	**0.742 ± 0.084**	0.143 ± 0.103
GDRO + P.P.	0	0.844 ± 0.009	0.817 ± 0.011	0.065 ± 0.020	0.889 ± 0.000	0.710 ± 0.007	0.187 ± 0.006
	1	0.842 ± 0.010	0.821 ± 0.016	0.054 ± 0.023	0.888 ± 0.008	0.717 ± 0.032	0.179 ± 0.036
	2	0.845 ± 0.011	0.823 ± 0.022	0.053 ± 0.026	0.887 ± 0.002	0.725 ± 0.033	0.171 ± 0.039
	5	0.854 ± 0.005	**0.826 ± 0.010**	0.054 ± 0.022	0.888 ± 0.001	**0.730 ± 0.019**	0.171 ± 0.021
GDRO + SR	0	0.816 ± 0.031	0.692 ± 0.098	0.200 ± 0.121	0.868 ± 0.009	0.663 ± 0.050	0.239 ± 0.059
	1	0.796 ± 0.045	0.718 ± 0.013	0.151 ± 0.037	0.864 ± 0.008	0.691 ± 0.046	0.198 ± 0.059
	2	0.803 ± 0.030	0.729 ± 0.073	0.139 ± 0.074	0.864 ± 0.012	0.689 ± 0.026	0.200 ± 0.052
	5	0.819 ± 0.031	**0.786 ± 0.017**	0.083 ± 0.033	0.860 ± 0.013	**0.717 ± 0.059**	0.158 ± 0.080
GDRO + SR + P.P.	0	0.808 ± 0.027	0.773 ± 0.061	0.095 ± 0.072	0.881 ± 0.004	0.725 ± 0.061	0.164 ± 0.065
	1	0.819 ± 0.023	0.798 ± 0.030	0.075 ± 0.019	0.879 ± 0.006	**0.738 ± 0.026**	0.151 ± 0.027
	2	0.830 ± 0.011	0.817 ± 0.018	0.040 ± 0.021	0.881 ± 0.004	0.723 ± 0.004	0.169 ± 0.007
	5	0.829 ± 0.007	**0.820 ± 0.007**	0.027 ± 0.020	0.882 ± 0.001	0.693 ± 0.041	0.206 ± 0.044

Table 6.3. Results on ablation study on group adjustment parameter for WaterBird and CelebA. Pre-trained ResNet 50 architecture is used.

6.2.2 Additional MNIST experiments

		Tailed Minority 2			Group Imbalance 5		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
ERM		0.889 ± 0.002	0.761 ± 0.031	0.229 ± 0.035	0.861 ± 0.007	0.662 ± 0.016	0.329 ± 0.019
ERM + P.P.		0.919 ± 0.005	0.836 ± 0.019	0.141 ± 0.012	**0.925 ± 0.006**	**0.877 ± 0.025**	**0.091 ± 0.033**
ERM + SR		0.918 ± 0.008	0.816 ± 0.010	0.168 ± 0.013	0.881 ± 0.004	0.767 ± 0.044	0.191 ± 0.047
ERM + SR + P.P.		0.921 ± 0.010	0.872 ± 0.019	0.103 ± 0.022	**0.925 ± 0.006**	**0.877 ± 0.018**	**0.099 ± 0.021**
GDRO	0	0.907 ± 0.008	0.856 ± 0.014	0.101 ± 0.011	0.894 ± 0.009	0.788 ± 0.026	0.177 ± 0.039
	1	**0.925 ± 0.007**	0.883 ± 0.006	0.087 ± 0.006	0.914 ± 0.016	0.824 ± 0.014	0.159 ± 0.021
	2	**0.924 ± 0.001**	**0.894 ± 0.002**	**0.066 ± 0.007**	0.909 ± 0.013	0.812 ± 0.014	0.162 ± 0.012
	5	0.905 ± 0.005	0.856 ± 0.009	0.124 ± 0.007	0.917 ± 0.001	0.844 ± 0.024	0.134 ± 0.028
GDRO + P.P.	0	**0.926 ± 0.011**	0.871 ± 0.003	0.095 ± 0.013	0.918 ± 0.010	**0.872 ± 0.021**	**0.102 ± 0.027**
	1	**0.932 ± 0.008**	**0.896 ± 0.006**	0.074 ± 0.011	**0.927 ± 0.005**	0.867 ± 0.018	0.112 ± 0.014
	2	**0.926 ± 0.002**	0.889 ± 0.021	0.075 ± 0.025	0.914 ± 0.014	0.861 ± 0.013	0.119 ± 0.018
	5	0.904 ± 0.006	0.848 ± 0.017	0.119 ± 0.004	0.912 ± 0.008	0.862 ± 0.007	0.111 ± 0.008
GDRO + SR	0	0.902 ± 0.017	0.839 ± 0.021	0.124 ± 0.037	0.893 ± 0.005	0.773 ± 0.059	0.199 ± 0.080
	1	0.919 ± 0.006	0.866 ± 0.019	0.101 ± 0.031	0.912 ± 0.010	0.773 ± 0.054	0.213 ± 0.054
	2	**0.926 ± 0.003**	0.874 ± 0.013	0.096 ± 0.026	0.918 ± 0.006	0.782 ± 0.019	0.194 ± 0.020
	5	0.899 ± 0.011	0.838 ± 0.015	0.133 ± 0.014	0.912 ± 0.004	0.852 ± 0.010	**0.110 ± 0.020**
GDRO + SR + P.P.	0	0.919 ± 0.010	0.867 ± 0.019	0.098 ± 0.016	0.918 ± 0.006	**0.878 ± 0.006**	**0.095 ± 0.009**
	1	**0.927 ± 0.006**	0.885 ± 0.007	0.076 ± 0.014	**0.922 ± 0.007**	0.868 ± 0.010	0.107 ± 0.010
	2	0.922 ± 0.001	0.889 ± 0.011	0.081 ± 0.025	0.921 ± 0.006	0.862 ± 0.010	0.121 ± 0.011
	5	0.904 ± 0.005	0.859 ± 0.019	0.097 ± 0.024	0.912 ± 0.010	0.864 ± 0.024	0.106 ± 0.035
SGDRO		0.917 ± 0.005	0.861 ± 0.016	0.101 ± 0.029	0.904 ± 0.018	0.775 ± 0.021	0.207 ± 0.017
SGDRO + P.P.		0.921 ± 0.006	0.871 ± 0.004	0.096 ± 0.006	**0.926 ± 0.004**	**0.878 ± 0.014**	**0.100 ± 0.012**
SGDRO + SR		0.921 ± 0.008	0.844 ± 0.034	0.134 ± 0.035	0.890 ± 0.014	0.772 ± 0.020	0.203 ± 0.034
SGDRO + SR + P.P.		**0.928 ± 0.005**	0.883 ± 0.025	0.089 ± 0.034	**0.924 ± 0.005**	**0.874 ± 0.014**	0.106 ± 0.023

Table 6.4. Category-based classification on MNIST in imbalance settings 2 and 5.

		Tailed Minority 3			Group Imbalance 6		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
ERM		0.930 ± 0.005	0.835 ± 0.023	0.153 ± 0.030	0.898 ± 0.003	0.752 ± 0.014	0.240 ± 0.010
ERM + P.P.		0.943 ± 0.006	0.906 ± 0.014	0.074 ± 0.021	0.938 ± 0.010	0.898 ± 0.017	0.081 ± 0.018
ERM + SR		0.945 ± 0.008	0.902 ± 0.019	0.082 ± 0.019	0.915 ± 0.023	0.812 ± 0.009	0.173 ± 0.015
ERM + SR + P.P.		0.953 ± 0.001	0.920 ± 0.011	0.058 ± 0.011	**0.941 ± 0.000**	0.903 ± 0.006	0.081 ± 0.007
GDRO	0	0.945 ± 0.011	0.915 ± 0.010	0.057 ± 0.018	0.914 ± 0.019	0.847 ± 0.008	0.127 ± 0.020
	1	**0.954 ± 0.004**	0.924 ± 0.008	0.053 ± 0.016	0.929 ± 0.010	0.867 ± 0.015	0.100 ± 0.022
	2	**0.956 ± 0.002**	**0.929 ± 0.015**	**0.046 ± 0.019**	**0.944 ± 0.003**	0.877 ± 0.015	0.100 ± 0.016
	5	0.940 ± 0.001	0.898 ± 0.018	0.077 ± 0.012	**0.942 ± 0.004**	0.903 ± 0.011	**0.076 ± 0.008**
GDRO + P.P.	0	0.950 ± 0.003	**0.931 ± 0.005**	**0.043 ± 0.008**	0.930 ± 0.021	0.887 ± 0.024	0.096 ± 0.016
	1	0.953 ± 0.002	**0.930 ± 0.016**	**0.042 ± 0.017**	**0.941 ± 0.001**	**0.907 ± 0.005**	**0.070 ± 0.009**
	2	0.953 ± 0.005	**0.938 ± 0.011**	**0.034 ± 0.017**	0.939 ± 0.000	0.895 ± 0.010	0.088 ± 0.007
	5	0.938 ± 0.002	0.896 ± 0.015	0.076 ± 0.018	0.937 ± 0.008	0.897 ± 0.009	0.088 ± 0.010
GDRO + SR	0	0.941 ± 0.005	0.894 ± 0.006	0.084 ± 0.003	0.911 ± 0.015	0.824 ± 0.038	0.148 ± 0.039
	1	0.949 ± 0.004	0.913 ± 0.004	0.071 ± 0.008	0.925 ± 0.022	0.853 ± 0.002	0.125 ± 0.021
	2	0.949 ± 0.003	0.917 ± 0.013	0.065 ± 0.015	0.940 ± 0.002	0.869 ± 0.008	0.114 ± 0.010
	5	0.938 ± 0.009	0.907 ± 0.013	0.064 ± 0.022	0.940 ± 0.005	0.895 ± 0.016	0.082 ± 0.016
GDRO + SR + P.P.	0	0.905 ± 0.005	0.914 ± 0.013	0.057 ± 0.015	0.935 ± 0.014	0.886 ± 0.038	0.081 ± 0.035
	1	0.949 ± 0.004	0.921 ± 0.009	**0.050 ± 0.009**	0.940 ± 0.007	0.899 ± 0.003	0.080 ± 0.007
	2	0.948 ± 0.005	0.917 ± 0.006	0.055 ± 0.007	0.937 ± 0.000	0.902 ± 0.004	0.079 ± 0.009
	5	0.945 ± 0.003	0.905 ± 0.012	0.071 ± 0.019	0.939 ± 0.003	**0.909 ± 0.004**	**0.071 ± 0.008**
SGDRO		0.949 ± 0.004	0.897 ± 0.009	0.084 ± 0.014	0.934 ± 0.006	0.851 ± 0.021	0.124 ± 0.032
SGDRO + P.P.		0.949 ± 0.007	0.922 ± 0.013	**0.051 ± 0.021**	**0.942 ± 0.002**	0.904 ± 0.014	0.077 ± 0.017
SGDRO + SR		0.949 ± 0.006	0.896 ± 0.003	0.086 ± 0.009	0.935 ± 0.006	0.848 ± 0.019	0.134 ± 0.022
SGDRO + SR + P.P.		**0.955 ± 0.004**	0.921 ± 0.008	0.058 ± 0.010	0.939 ± 0.006	**0.910 ± 0.013**	**0.071 ± 0.005**

Table 6.5. Category-based classification on MNIST in imbalance settings 3 and 6.

6.2.3 FashionMNIST experiments

		Tailed Minority 1			Group Imbalance 4		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
ERM		0.747 ± 0.006	0.311 ± 0.084	0.678 ± 0.085	0.670 ± 0.018	0.155 ± 0.010	0.826 ± 0.011
ERM + P.P.		0.762 ± 0.018	0.389 ± 0.115	0.590 ± 0.117	0.721 ± 0.022	0.570 ± 0.040	**0.340 ± 0.046**
ERM + SR		0.783 ± 0.008	0.559 ± 0.035	0.386 ± 0.035	0.766 ± 0.026	0.499 ± 0.065	0.451 ± 0.073
ERM + SR + P.P.		0.777 ± 0.015	0.561 ± 0.075	0.368 ± 0.070	0.746 ± 0.029	0.523 ± 0.068	0.400 ± 0.104
GDRO	0	0.785 ± 0.007	0.576 ± 0.045	0.366 ± 0.042	0.759 ± 0.005	0.505 ± 0.065	0.447 ± 0.092
	1	0.770 ± 0.020	0.558 ± 0.027	0.365 ± 0.042	0.765 ± 0.022	0.482 ± 0.106	0.468 ± 0.108
	2	0.777 ± 0.019	**0.604 ± 0.005**	**0.326 ± 0.021**	0.757 ± 0.009	0.530 ± 0.024	0.419 ± 0.011
	5	0.697 ± 0.028	0.482 ± 0.021	0.450 ± 0.063	0.758 ± 0.009	0.536 ± 0.010	0.420 ± 0.022
GDRO + P.P.	0	0.776 ± 0.028	0.593 ± 0.032	**0.336 ± 0.007**	0.749 ± 0.024	**0.579 ± 0.012**	**0.343 ± 0.041**
	1	0.765 ± 0.026	0.565 ± 0.040	0.369 ± 0.036	0.728 ± 0.039	0.571 ± 0.036	**0.339 ± 0.054**
	2	0.773 ± 0.020	0.595 ± 0.015	**0.340 ± 0.013**	0.754 ± 0.014	**0.581 ± 0.038**	0.364 ± 0.033
	5	0.669 ± 0.006	0.449 ± 0.033	0.477 ± 0.068	0.766 ± 0.005	0.567 ± 0.046	0.380 ± 0.060
GDRO + SR	0	0.767 ± 0.016	0.548 ± 0.047	0.405 ± 0.040	0.754 ± 0.016	0.464 ± 0.036	0.476 ± 0.029
	1	0.765 ± 0.017	0.573 ± 0.014	0.358 ± 0.009	0.747 ± 0.023	0.494 ± 0.099	0.451 ± 0.082
	2	0.755 ± 0.032	0.564 ± 0.049	0.370 ± 0.065	0.743 ± 0.022	0.528 ± 0.063	0.399 ± 0.073
	5	0.713 ± 0.011	0.485 ± 0.069	0.445 ± 0.083	0.744 ± 0.016	0.540 ± 0.019	0.377 ± 0.022
GDRO + SR + P.P.	0	0.778 ± 0.003	**0.599 ± 0.010**	0.350 ± 0.017	0.748 ± 0.023	**0.582 ± 0.010**	**0.329 ± 0.034**
	1	0.764 ± 0.013	0.585 ± 0.028	0.361 ± 0.025	0.717 ± 0.027	**0.578 ± 0.059**	**0.308 ± 0.070**
	2	0.747 ± 0.008	0.538 ± 0.059	0.371 ± 0.054	0.753 ± 0.041	0.574 ± 0.029	**0.350 ± 0.018**
	5	0.673 ± 0.042	0.467 ± 0.115	0.501 ± 0.153	0.753 ± 0.003	0.557 ± 0.023	0.376 ± 0.018
SGDRO		**0.790 ± 0.014**	0.578 ± 0.033	0.375 ± 0.030	**0.780 ± 0.013**	0.519 ± 0.011	0.455 ± 0.017
SGDRO + P.P.		**0.795 ± 0.005**	0.598 ± 0.029	0.354 ± 0.032	0.748 ± 0.028	**0.595 ± 0.019**	**0.335 ± 0.028**
SGDRO + SR		0.763 ± 0.021	0.469 ± 0.091	0.485 ± 0.089	**0.768 ± 0.008**	0.485 ± 0.041	0.463 ± 0.045
SGDRO + SR + P.P.		0.771 ± 0.018	0.511 ± 0.065	0.414 ± 0.082	0.760 ± 0.016	0.564 ± 0.016	0.381 ± 0.029

Table 6.6. Category-based classification on FashionMNIST in imbalance settings 1 and 4.

		Tailed Minority 2			Group Imbalance 5		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
ERM		0.771 ± 0.018	0.451 ± 0.019	0.533 ± 0.018	0.717 ± 0.020	0.299 ± 0.023	0.686 ± 0.025
ERM + P.P.		0.789 ± 0.012	0.522 ± 0.035	0.451 ± 0.027	0.735 ± 0.003	0.601 ± 0.028	**0.311 ± 0.028**
ERM + SR		0.796 ± 0.008	0.612 ± 0.022	0.352 ± 0.048	0.804 ± 0.002	0.607 ± 0.029	0.350 ± 0.023
ERM + SR + P.P.		0.798 ± 0.007	**0.633 ± 0.002**	0.324 ± 0.025	0.752 ± 0.036	0.607 ± 0.014	0.333 ± 0.020
GDRO	0	0.789 ± 0.010	0.622 ± 0.039	0.325 ± 0.025	0.789 ± 0.007	0.610 ± 0.008	0.335 ± 0.022
	1	0.789 ± 0.024	0.625 ± 0.005	**0.303 ± 0.006**	0.796 ± 0.004	0.624 ± 0.021	**0.318 ± 0.018**
	2	0.792 ± 0.007	0.609 ± 0.006	0.342 ± 0.018	0.785 ± 0.011	0.614 ± 0.034	0.333 ± 0.036
	5	0.791 ± 0.009	0.602 ± 0.012	0.337 ± 0.041	0.780 ± 0.012	0.622 ± 0.026	0.322 ± 0.045
GDRO + P.P.	0	0.784 ± 0.016	**0.649 ± 0.021**	**0.278 ± 0.025**	0.770 ± 0.014	**0.638 ± 0.009**	**0.295 ± 0.024**
	1	0.794 ± 0.021	0.626 ± 0.023	0.316 ± 0.053	0.777 ± 0.027	**0.639 ± 0.044**	**0.300 ± 0.050**
	2	0.775 ± 0.003	0.610 ± 0.006	0.313 ± 0.038	0.777 ± 0.029	0.625 ± 0.029	**0.303 ± 0.042**
	5	0.769 ± 0.010	0.620 ± 0.028	0.334 ± 0.031	0.771 ± 0.039	**0.634 ± 0.011**	**0.312 ± 0.023**
GDRO + SR	0	0.774 ± 0.040	0.561 ± 0.006	0.383 ± 0.021	0.774 ± 0.013	0.581 ± 0.018	0.365 ± 0.034
	1	0.777 ± 0.006	0.587 ± 0.042	0.346 ± 0.025	0.784 ± 0.006	0.580 ± 0.035	0.350 ± 0.047
	2	0.783 ± 0.009	0.597 ± 0.060	0.335 ± 0.078	0.768 ± 0.025	0.593 ± 0.034	0.338 ± 0.050
	5	0.774 ± 0.014	0.610 ± 0.010	0.328 ± 0.049	0.777 ± 0.016	0.581 ± 0.049	0.345 ± 0.043
GDRO + SR + P.P.	0	0.761 ± 0.005	0.606 ± 0.021	0.320 ± 0.023	0.757 ± 0.030	0.617 ± 0.015	**0.294 ± 0.062**
	1	0.777 ± 0.006	0.602 ± 0.033	0.337 ± 0.040	0.779 ± 0.015	**0.637 ± 0.022**	**0.297 ± 0.027**
	2	0.784 ± 0.011	0.616 ± 0.023	0.316 ± 0.030	0.755 ± 0.042	0.617 ± 0.024	**0.306 ± 0.031**
	5	0.772 ± 0.021	0.619 ± 0.013	0.334 ± 0.042	0.759 ± 0.031	0.601 ± 0.014	**0.310 ± 0.030**
SGDRO		**0.807 ± 0.011**	0.587 ± 0.007	0.382 ± 0.010	**0.817 ± 0.008**	0.600 ± 0.043	0.374 ± 0.037
SGDRO + P.P.		**0.811 ± 0.008**	0.614 ± 0.036	0.328 ± 0.030	0.794 ± 0.030	0.628 ± 0.032	**0.313 ± 0.054**
SGDRO + SR		0.800 ± 0.008	0.569 ± 0.023	0.378 ± 0.021	0.795 ± 0.011	0.580 ± 0.014	0.376 ± 0.021
SGDRO + SR + P.P.		0.798 ± 0.007	0.595 ± 0.051	0.340 ± 0.060	0.767 ± 0.066	0.606 ± 0.023	0.337 ± 0.043

Table 6.7. Category-based classification on FashionMNIST in imbalance settings 2 and 5.

6.2.4 Additional CIFAR10 experiments

		Tailed Minority 3			Group Imbalance 6		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
ERM		0.809 ± 0.005	0.454 ± 0.055	0.525 ± 0.058	0.771 ± 0.013	0.446 ± 0.022	0.535 ± 0.021
ERM + P.P.		0.816 ± 0.007	0.558 ± 0.048	0.401 ± 0.063	0.771 ± 0.003	0.647 ± 0.017	0.294 ± 0.016
ERM + SR		0.824 ± 0.007	0.634 ± 0.046	0.331 ± 0.037	0.820 ± 0.002	0.643 ± 0.004	0.322 ± 0.007
ERM + SR + P.P.		0.807 ± 0.005	0.610 ± 0.042	0.339 ± 0.053	0.802 ± 0.039	0.650 ± 0.009	0.306 ± 0.025
GDRO	0	0.817 ± 0.003	0.657 ± 0.019	0.294 ± 0.027	0.807 ± 0.007	0.647 ± 0.009	0.299 ± 0.014
	1	0.807 ± 0.010	0.647 ± 0.020	0.286 ± 0.026	0.802 ± 0.024	0.640 ± 0.011	0.309 ± 0.044
	2	0.822 ± 0.017	0.666 ± 0.015	0.282 ± 0.021	0.810 ± 0.019	0.659 ± 0.003	0.295 ± 0.014
	5	0.813 ± 0.011	**0.681 ± 0.014**	**0.248 ± 0.013**	0.810 ± 0.002	0.669 ± 0.018	0.277 ± 0.017
GDRO + P.P.	0	0.815 ± 0.002	0.669 ± 0.007	0.280 ± 0.018	0.797 ± 0.021	0.671 ± 0.015	**0.264 ± 0.015**
	1	0.810 ± 0.013	0.652 ± 0.025	0.284 ± 0.033	0.798 ± 0.006	0.652 ± 0.015	0.298 ± 0.008
	2	0.809 ± 0.011	**0.683 ± 0.013**	**0.250 ± 0.020**	0.797 ± 0.006	**0.685 ± 0.019**	**0.239 ± 0.034**
	5	0.796 ± 0.017	**0.678 ± 0.020**	0.270 ± 0.036	0.784 ± 0.015	0.665 ± 0.009	0.270 ± 0.011
GDRO + SR	0	0.807 ± 0.016	0.629 ± 0.062	0.307 ± 0.069	0.795 ± 0.013	0.604 ± 0.064	0.347 ± 0.064
	1	0.809 ± 0.015	0.647 ± 0.029	0.300 ± 0.051	0.795 ± 0.002	0.649 ± 0.013	0.298 ± 0.028
	2	0.808 ± 0.016	0.642 ± 0.056	0.288 ± 0.067	0.780 ± 0.019	0.613 ± 0.044	0.315 ± 0.060
	5	0.814 ± 0.010	0.654 ± 0.031	0.275 ± 0.011	0.796 ± 0.014	0.664 ± 0.010	0.283 ± 0.018
GDRO + SR + P.P.	0	0.808 ± 0.024	0.667 ± 0.029	0.275 ± 0.018	0.801 ± 0.005	0.662 ± 0.015	0.289 ± 0.028
	1	0.814 ± 0.022	0.668 ± 0.014	0.274 ± 0.027	0.798 ± 0.022	0.652 ± 0.020	0.283 ± 0.023
	2	0.802 ± 0.016	0.662 ± 0.014	0.263 ± 0.024	0.796 ± 0.026	0.658 ± 0.033	0.272 ± 0.039
	5	0.804 ± 0.024	0.657 ± 0.017	0.281 ± 0.023	0.793 ± 0.022	0.669 ± 0.022	0.273 ± 0.034
SGDRO		0.835 ± 0.004	0.649 ± 0.028	0.312 ± 0.011	**0.825 ± 0.004**	0.648 ± 0.030	0.317 ± 0.040
SGDRO + P.P.		**0.840 ± 0.001**	0.662 ± 0.026	0.310 ± 0.024	0.820 ± 0.014	**0.681 ± 0.009**	**0.266 ± 0.003**
SGDRO + SR		0.820 ± 0.012	0.602 ± 0.021	0.364 ± 0.027	0.820 ± 0.001	0.625 ± 0.039	0.328 ± 0.039
SGDRO + SR + P.P.		0.819 ± 0.006	0.631 ± 0.024	0.316 ± 0.040	0.789 ± 0.041	0.647 ± 0.023	0.294 ± 0.019

Table 6.8. Category-based classification on FashionMNIST in imbalance settings 3 and 6.

		Tailed Minority 2			Group Imbalance 5		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
ERM		0.597 ± 0.012	0.346 ± 0.021	0.526 ± 0.024	0.535 ± 0.010	0.171 ± 0.013	0.723 ± 0.010
ERM + P.P.		0.590 ± 0.026	0.469 ± 0.018	0.331 ± 0.022	0.553 ± 0.005	0.433 ± 0.020	**0.275 ± 0.044**
ERM + SR		0.535 ± 0.021	0.354 ± 0.015	0.409 ± 0.042	0.466 ± 0.025	0.256 ± 0.038	0.519 ± 0.017
ERM + SR + P.P.		0.543 ± 0.019	0.416 ± 0.037	**0.318 ± 0.179**	0.480 ± 0.014	0.329 ± 0.028	0.361 ± 0.015
GDRO	0	0.644 ± 0.030	0.490 ± 0.031	0.361 ± 0.062	0.577 ± 0.040	0.370 ± 0.029	0.474 ± 0.033
	1	0.667 ± 0.035	0.521 ± 0.018	**0.316 ± 0.101**	0.580 ± 0.004	0.387 ± 0.011	0.376 ± 0.038
	2	0.676 ± 0.010	0.506 ± 0.015	0.366 ± 0.037	0.580 ± 0.010	0.363 ± 0.032	0.444 ± 0.107
	5	0.684 ± 0.022	0.515 ± 0.019	0.353 ± 0.032	0.553 ± 0.034	0.385 ± 0.034	0.341 ± 0.126
GDRO + P.P.	0	0.681 ± 0.029	**0.547 ± 0.023**	**0.314 ± 0.032**	**0.636 ± 0.029**	**0.484 ± 0.037**	0.380 ± 0.044
	1	**0.709 ± 0.016**	**0.546 ± 0.011**	0.342 ± 0.010	**0.636 ± 0.015**	**0.486 ± 0.025**	0.338 ± 0.066
	2	0.674 ± 0.025	**0.538 ± 0.031**	**0.313 ± 0.011**	**0.640 ± 0.043**	**0.494 ± 0.022**	**0.308 ± 0.063**
	5	**0.693 ± 0.019**	0.526 ± 0.014	**0.324 ± 0.028**	**0.636 ± 0.012**	**0.494 ± 0.020**	0.385 ± 0.066
GDRO + SR	0	0.612 ± 0.035	0.436 ± 0.021	0.421 ± 0.093	0.466 ± 0.028	0.255 ± 0.035	0.459 ± 0.022
	1	0.625 ± 0.029	0.432 ± 0.040	0.429 ± 0.108	0.459 ± 0.036	0.218 ± 0.034	0.532 ± 0.087
	2	0.611 ± 0.047	0.443 ± 0.032	0.348 ± 0.113	0.499 ± 0.034	0.225 ± 0.073	0.563 ± 0.138
	5	0.611 ± 0.019	0.407 ± 0.029	0.439 ± 0.083	0.491 ± 0.069	0.216 ± 0.041	0.491 ± 0.194
GDRO + SR + P.P.	0	0.656 ± 0.009	0.508 ± 0.021	0.351 ± 0.045	0.603 ± 0.023	0.421 ± 0.032	0.408 ± 0.030
	1	0.636 ± 0.045	0.479 ± 0.029	**0.321 ± 0.019**	0.603 ± 0.019	0.425 ± 0.017	0.380 ± 0.023
	2	0.642 ± 0.057	0.498 ± 0.035	0.331 ± 0.033	0.596 ± 0.018	0.450 ± 0.033	0.398 ± 0.062
	5	0.677 ± 0.026	0.495 ± 0.041	0.371 ± 0.046	0.605 ± 0.008	0.435 ± 0.009	0.401 ± 0.012
SGDRO		0.569 ± 0.034	0.367 ± 0.024	0.432 ± 0.023	0.511 ± 0.013	0.257 ± 0.057	0.500 ± 0.106
SGDRO + P.P.		0.583 ± 0.002	0.396 ± 0.015	0.403 ± 0.070	0.533 ± 0.011	0.382 ± 0.038	0.342 ± 0.071
SGDRO + SR		0.503 ± 0.008	0.318 ± 0.039	0.409 ± 0.042	0.465 ± 0.004	0.222 ± 0.057	0.521 ± 0.070
SGDRO + SR + P.P.		0.532 ± 0.011	0.406 ± 0.011	0.337 ± 0.128	0.489 ± 0.006	0.361 ± 0.022	0.325 ± 0.029

Table 6.9. Category-based classification on CIFAR10 in imbalance settings 2 and 5.

6.2.5 Attribute-based classification on ResNet 18 architecture

		Tailed Minority 3			Group Imbalance 6		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
ERM		0.610 ± 0.015	0.390 ± 0.045	0.488 ± 0.040	0.587 ± 0.014	0.287 ± 0.018	0.593 ± 0.043
ERM + P.P.		0.597 ± 0.006	0.469 ± 0.010	0.358 ± 0.059	0.593 ± 0.004	0.446 ± 0.009	**0.291 ± 0.020**
ERM + SR		0.536 ± 0.049	0.337 ± 0.055	0.408 ± 0.133	0.519 ± 0.029	0.278 ± 0.041	0.495 ± 0.115
ERM + SR + P.P.		0.548 ± 0.006	0.402 ± 0.022	0.334 ± 0.028	0.541 ± 0.013	0.403 ± 0.019	0.341 ± 0.040
GDRO	0	0.684 ± 0.053	0.512 ± 0.075	0.330 ± 0.048	0.636 ± 0.043	0.426 ± 0.009	0.426 ± 0.010
	1	**0.704 ± 0.022**	0.531 ± 0.030	0.368 ± 0.021	0.644 ± 0.016	0.436 ± 0.026	0.386 ± 0.025
	2	0.677 ± 0.033	0.515 ± 0.027	0.366 ± 0.050	0.635 ± 0.015	0.456 ± 0.070	0.386 ± 0.160
	5	0.688 ± 0.035	0.510 ± 0.044	0.358 ± 0.137	0.641 ± 0.020	0.456 ± 0.025	0.369 ± 0.090
GDRO + P.P.	0	**0.707 ± 0.014**	0.557 ± 0.014	0.353 ± 0.015	**0.685 ± 0.016**	0.529 ± 0.022	0.348 ± 0.024
	1	**0.704 ± 0.017**	**0.579 ± 0.005**	0.309 ± 0.030	**0.689 ± 0.012**	**0.557 ± 0.018**	0.334 ± 0.040
	2	0.685 ± 0.030	0.557 ± 0.027	0.313 ± 0.048	**0.683 ± 0.044**	**0.540 ± 0.020**	0.342 ± 0.019
	5	**0.700 ± 0.017**	0.558 ± 0.034	0.339 ± 0.085	**0.689 ± 0.026**	0.538 ± 0.019	0.361 ± 0.008
GDRO + SR	0	0.688 ± 0.023	0.490 ± 0.045	0.416 ± 0.053	0.548 ± 0.078	0.279 ± 0.053	0.462 ± 0.068
	1	0.676 ± 0.026	0.509 ± 0.050	0.376 ± 0.065	0.536 ± 0.013	0.349 ± 0.085	0.464 ± 0.112
	2	0.680 ± 0.019	0.484 ± 0.021	0.397 ± 0.081	0.552 ± 0.018	0.339 ± 0.028	0.429 ± 0.148
	5	0.663 ± 0.016	0.498 ± 0.027	0.386 ± 0.054	0.583 ± 0.020	0.330 ± 0.054	0.573 ± 0.072
GDRO + SR + P.P.	0	0.678 ± 0.010	0.537 ± 0.019	0.345 ± 0.020	0.667 ± 0.009	0.482 ± 0.033	0.387 ± 0.021
	1	0.688 ± 0.028	0.529 ± 0.072	0.346 ± 0.097	0.633 ± 0.029	0.496 ± 0.035	0.379 ± 0.020
	2	0.687 ± 0.014	0.529 ± 0.032	0.371 ± 0.035	0.667 ± 0.006	0.498 ± 0.027	0.389 ± 0.027
	5	0.668 ± 0.015	0.545 ± 0.032	**0.276 ± 0.007**	0.609 ± 0.039	0.476 ± 0.029	0.358 ± 0.071
SGDRO		0.585 ± 0.015	0.402 ± 0.046	0.371 ± 0.125	0.537 ± 0.007	0.332 ± 0.047	0.393 ± 0.024
SGDRO + P.P.		0.575 ± 0.019	0.411 ± 0.044	0.367 ± 0.105	0.576 ± 0.008	0.422 ± 0.028	0.360 ± 0.023
SGDRO + SR		0.553 ± 0.007	0.345 ± 0.021	0.455 ± 0.061	0.493 ± 0.051	0.284 ± 0.062	0.457 ± 0.123
SGDRO + SR + P.P.		0.554 ± 0.011	0.375 ± 0.003	0.378 ± 0.064	0.552 ± 0.020	0.381 ± 0.027	0.375 ± 0.022

Table 6.10. Category-based classification on CIFAR10 in imbalance settings 3 and 6.

		WaterBird			CelebA		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
ERM		0.712 ± 0.009	0.218 ± 0.036	0.773 ± 0.036	0.867 ± 0.004	0.245 ± 0.031	0.740 ± 0.031
ERM + P.P.		0.727 ± 0.015	0.680 ± 0.008	0.228 ± 0.011	0.864 ± 0.007	0.736 ± 0.028	0.225 ± 0.028
ERM + SR		0.782 ± 0.029	0.587 ± 0.099	0.279 ± 0.180	0.892 ± 0.007	0.685 ± 0.048	0.240 ± 0.057
ERM + SR + P.P.		0.802 ± 0.009	0.774 ± 0.010	0.090 ± 0.022	0.895 ± 0.002	**0.762 ± 0.016**	0.147 ± 0.019
GDRO	0	0.804 ± 0.044	0.687 ± 0.035	0.177 ± 0.020	**0.902 ± 0.000**	0.706 ± 0.062	0.219 ± 0.062
	1	0.790 ± 0.010	0.667 ± 0.125	0.188 ± 0.123	0.889 ± 0.013	**0.783 ± 0.052**	**0.122 ± 0.057**
	2	0.828 ± 0.012	0.710 ± 0.044	0.192 ± 0.084	0.892 ± 0.006	**0.768 ± 0.026**	**0.135 ± 0.023**
	5	0.812 ± 0.034	0.753 ± 0.024	0.115 ± 0.042	0.891 ± 0.006	0.740 ± 0.026	0.166 ± 0.024
GDRO + P.P.	0	0.826 ± 0.015	0.793 ± 0.005	**0.053 ± 0.011**	0.900 ± 0.003	0.749 ± 0.026	0.159 ± 0.030
	1	0.822 ± 0.030	0.783 ± 0.029	0.090 ± 0.037	0.897 ± 0.004	**0.779 ± 0.030**	**0.130 ± 0.024**
	2	**0.851 ± 0.004**	**0.819 ± 0.008**	0.056 ± 0.011	0.901 ± 0.003	0.743 ± 0.018	0.169 ± 0.019
	5	0.834 ± 0.008	0.776 ± 0.042	0.100 ± 0.047	0.897 ± 0.007	0.742 ± 0.051	0.167 ± 0.052
GDRO + SR	0	0.775 ± 0.028	0.570 ± 0.085	0.343 ± 0.103	0.895 ± 0.011	0.667 ± 0.022	0.252 ± 0.033
	1	0.789 ± 0.020	0.551 ± 0.022	0.374 ± 0.082	0.895 ± 0.002	0.676 ± 0.043	0.243 ± 0.044
	2	0.796 ± 0.024	0.672 ± 0.039	0.234 ± 0.079	0.887 ± 0.012	0.732 ± 0.086	0.164 ± 0.105
	5	0.815 ± 0.017	0.763 ± 0.015	0.094 ± 0.020	0.894 ± 0.009	0.712 ± 0.060	0.200 ± 0.066
GDRO + SR + P.P.	0	0.800 ± 0.012	0.760 ± 0.010	0.104 ± 0.018	0.895 ± 0.002	0.753 ± 0.019	0.156 ± 0.021
	1	0.805 ± 0.019	0.784 ± 0.027	0.074 ± 0.041	0.893 ± 0.002	**0.764 ± 0.006**	**0.142 ± 0.004**
	2	0.819 ± 0.012	0.785 ± 0.010	0.065 ± 0.009	0.894 ± 0.000	0.749 ± 0.037	0.160 ± 0.035
	5	0.827 ± 0.000	0.782 ± 0.014	0.070 ± 0.012	0.897 ± 0.003	0.751 ± 0.033	0.163 ± 0.036
SGDRO		0.798 ± 0.047	0.706 ± 0.035	0.176 ± 0.034	0.897 ± 0.001	0.699 ± 0.054	0.223 ± 0.058
SGDRO + P.P.		0.801 ± 0.007	0.794 ± 0.005	**0.040 ± 0.015**	0.900 ± 0.005	0.747 ± 0.017	0.162 ± 0.017
SGDRO + SR		0.779 ± 0.017	0.589 ± 0.070	0.288 ± 0.090	0.889 ± 0.001	0.691 ± 0.028	0.218 ± 0.035
SGDRO + SR + P.P.		0.808 ± 0.007	0.790 ± 0.015	**0.053 ± 0.007**	0.894 ± 0.002	0.738 ± 0.026	0.168 ± 0.030

Table 6.11. Attribute-based classification results on WaterBird and CelebA datasets for Res 18 architecture.

6.2.6 Comparison of train and valid performance on selected penalty function

The distributional similarity between valid and test datasets can cause the superior performance of our adaptive prediction penalty algorithm. To validate that this is not the case,

we compare our method's train and valid performances on the best penalty function and the worst-case best model. Tables demonstrate a typical generalization gap between these performances, and the best case chosen from the validation dataset does not fail on the training dataset. These results suggest that our method does not overfit the validation set and provides a proper penalty to the model trained on the training dataset.

			Tailed Minority			Group Imbalance		
			Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
	ERM	train	0.966 ± 0.024	0.943 ± 0.039	0.053 ± 0.035	0.945 ± 0.010	0.932 ± 0.015	0.068 ± 0.015
		valid	0.863 ± 0.008	0.686 ± 0.044	0.293 ± 0.042	0.899 ± 0.003	0.841 ± 0.013	0.111 ± 0.008
	ERM + SR	train	0.957 ± 0.050	0.848 ± 0.143	0.152 ± 0.143	0.964 ± 0.004	0.913 ± 0.012	0.087 ± 0.012
		valid	0.882 ± 0.025	0.814 ± 0.015	0.134 ± 0.032	0.909 ± 0.003	0.866 ± 0.004	0.099 ± 0.011
	GDRO 0	train	0.923 ± 0.016	0.898 ± 0.018	0.102 ± 0.018	0.907 ± 0.012	0.890 ± 0.016	0.110 ± 0.016
		valid	0.892 ± 0.007	0.829 ± 0.036	0.122 ± 0.042	0.896 ± 0.008	0.851 ± 0.020	0.113 ± 0.042
	GDRO 0 + SR	train	0.967 ± 0.018	0.863 ± 0.091	0.137 ± 0.091	0.944 ± 0.019	0.856 ± 0.057	0.144 ± 0.057
		valid	0.886 ± 0.005	0.818 ± 0.015	0.144 ± 0.034	0.882 ± 0.010	0.821 ± 0.014	0.127 ± 0.018
	GDRO 1	train	0.892 ± 0.039	0.846 ± 0.069	0.154 ± 0.069	0.918 ± 0.030	0.902 ± 0.036	0.098 ± 0.036
		valid	0.902 ± 0.013	0.864 ± 0.021	0.098 ± 0.034	0.900 ± 0.018	0.852 ± 0.013	0.112 ± 0.015
	GDRO 1 + SR	train	0.976 ± 0.003	0.898 ± 0.014	0.102 ± 0.014	0.950 ± 0.017	0.875 ± 0.042	0.125 ± 0.042
		valid	0.913 ± 0.005	0.879 ± 0.006	0.081 ± 0.003	0.886 ± 0.022	0.821 ± 0.023	0.158 ± 0.015
MNIST	GDRO 2	train	0.851 ± 0.056	0.792 ± 0.089	0.208 ± 0.089	0.907 ± 0.018	0.870 ± 0.026	0.130 ± 0.026
		valid	0.890 ± 0.004	0.834 ± 0.008	0.119 ± 0.011	0.898 ± 0.006	0.847 ± 0.013	0.112 ± 0.008
	GDRO 2 + SR	train	0.955 ± 0.016	0.830 ± 0.089	0.170 ± 0.089	0.941 ± 0.012	0.844 ± 0.023	0.156 ± 0.023
		valid	0.885 ± 0.012	0.840 ± 0.015	0.099 ± 0.024	0.883 ± 0.020	0.834 ± 0.022	0.121 ± 0.008
	GDRO 5	train	0.769 ± 0.053	0.688 ± 0.041	0.312 ± 0.041	0.889 ± 0.013	0.851 ± 0.016	0.149 ± 0.016
		valid	0.851 ± 0.015	0.775 ± 0.024	0.187 ± 0.016	0.886 ± 0.004	0.841 ± 0.004	0.111 ± 0.015
	GDRO 5 + SR	train	0.885 ± 0.012	0.580 ± 0.201	0.420 ± 0.201	0.954 ± 0.004	0.868 ± 0.011	0.132 ± 0.011
		valid	0.841 ± 0.010	0.765 ± 0.009	0.199 ± 0.020	0.896 ± 0.010	0.838 ± 0.006	0.136 ± 0.008
	SGDRO	train	0.933 ± 0.047	0.910 ± 0.063	0.090 ± 0.063	0.923 ± 0.010	0.906 ± 0.014	0.094 ± 0.014
		valid	0.880 ± 0.016	0.763 ± 0.062	0.197 ± 0.061	0.898 ± 0.005	0.844 ± 0.011	0.122 ± 0.011
	SGDRO + SR	train	0.958 ± 0.054	0.859 ± 0.138	0.141 ± 0.138	0.968 ± 0.004	0.923 ± 0.009	0.077 ± 0.009
		valid	0.892 ± 0.019	0.813 ± 0.026	0.164 ± 0.018	0.905 ± 0.012	0.853 ± 0.020	0.116 ± 0.018

Table 6.12. Performance gap between the train and valid for selected penalty function on MNIST dataset. The best function is chosen according to the valid performance and works well in training data where the data distribution can differ. Validation performance in this table is the approximation of the test performance represented in other tables.

			Tailed Minority			Group Imbalance		
			Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
	ERM	train	0.988 ± 0.012	0.817 ± 0.071	0.183 ± 0.071	0.781 ± 0.058	0.660 ± 0.076	0.340 ± 0.076
		valid	0.783 ± 0.015	0.425 ± 0.092	0.562 ± 0.090	0.732 ± 0.016	0.594 ± 0.023	0.318 ± 0.030
	ERM + SR	train	0.892 ± 0.017	0.560 ± 0.064	0.440 ± 0.064	0.907 ± 0.022	0.623 ± 0.155	0.375 ± 0.153
		valid	0.783 ± 0.014	0.599 ± 0.064	0.335 ± 0.058	0.758 ± 0.028	0.584 ± 0.024	0.344 ± 0.056
	GDRO 0	train	0.800 ± 0.141	0.708 ± 0.190	0.292 ± 0.190	0.817 ± 0.061	0.674 ± 0.075	0.326 ± 0.075
		valid	0.789 ± 0.024	0.606 ± 0.015	0.325 ± 0.014	0.763 ± 0.015	0.623 ± 0.010	0.317 ± 0.036
	GDRO 0 + SR	train	0.923 ± 0.022	0.804 ± 0.019	0.180 ± 0.040	0.914 ± 0.030	0.756 ± 0.083	0.244 ± 0.083
		valid	0.796 ± 0.010	0.648 ± 0.022	0.300 ± 0.026	0.752 ± 0.016	0.586 ± 0.019	0.341 ± 0.016
	GDRO 1	train	0.790 ± 0.123	0.610 ± 0.132	0.364 ± 0.095	0.736 ± 0.062	0.597 ± 0.071	0.403 ± 0.071
		valid	0.776 ± 0.023	0.601 ± 0.020	0.330 ± 0.017	0.735 ± 0.038	0.604 ± 0.043	0.324 ± 0.051
	GDRO 1 + SR	train	0.905 ± 0.057	0.753 ± 0.079	0.233 ± 0.064	0.872 ± 0.026	0.636 ± 0.057	0.364 ± 0.057
		valid	0.777 ± 0.013	0.629 ± 0.021	0.306 ± 0.023	0.734 ± 0.023	0.621 ± 0.042	0.273 ± 0.054
FMNIST	GDRO 2	train	0.828 ± 0.043	0.717 ± 0.018	0.283 ± 0.018	0.810 ± 0.043	0.647 ± 0.049	0.340 ± 0.064
		valid	0.782 ± 0.016	0.630 ± 0.013	0.304 ± 0.028	0.769 ± 0.018	0.616 ± 0.018	0.331 ± 0.036
	GDRO 2 + SR	train	0.798 ± 0.028	0.557 ± 0.052	0.426 ± 0.061	0.918 ± 0.029	0.745 ± 0.059	0.255 ± 0.059
		valid	0.762 ± 0.011	0.583 ± 0.049	0.327 ± 0.067	0.766 ± 0.035	0.618 ± 0.015	0.326 ± 0.036
	GDRO 5	train	0.692 ± 0.071	0.412 ± 0.069	0.588 ± 0.069	0.832 ± 0.040	0.734 ± 0.068	0.266 ± 0.068
		valid	0.676 ± 0.003	0.474 ± 0.044	0.437 ± 0.099	0.774 ± 0.006	0.608 ± 0.016	0.338 ± 0.010
	GDRO 5 + SR	train	0.776 ± 0.037	0.465 ± 0.134	0.535 ± 0.134	0.901 ± 0.031	0.683 ± 0.125	0.317 ± 0.125
		valid	0.680 ± 0.035	0.475 ± 0.099	0.473 ± 0.153	0.767 ± 0.009	0.594 ± 0.016	0.332 ± 0.022
	SGDRO	train	0.926 ± 0.023	0.699 ± 0.042	0.301 ± 0.042	0.804 ± 0.041	0.651 ± 0.056	0.349 ± 0.056
		valid	0.808 ± 0.012	0.630 ± 0.005	0.323 ± 0.016	0.755 ± 0.035	0.630 ± 0.026	0.301 ± 0.027
	SGDRO + SR	train	0.841 ± 0.048	0.413 ± 0.147	0.587 ± 0.147	0.928 ± 0.009	0.739 ± 0.054	0.261 ± 0.054
		valid	0.787 ± 0.007	0.564 ± 0.037	0.356 ± 0.046	0.776 ± 0.017	0.598 ± 0.022	0.359 ± 0.040

Table 6.13. Performance gap between the train and valid for selected penalty function on FMNIST dataset. The best function is chosen according to the valid performance and works well in training data where the data distribution can differ. Validation performance in this table is the approximation of the test performance represented in other tables.

			Tailed Minority			Group Imbalance		
			Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
	ERM	train	0.996 ± 0.003	0.972 ± 0.019	0.028 ± 0.019	0.995 ± 0.004	0.990 ± 0.006	0.010 ± 0.006
		valid	0.535 ± 0.010	0.322 ± 0.023	0.469 ± 0.020	0.530 ± 0.014	0.417 ± 0.012	0.275 ± 0.075
	ERM + SR	train	0.677 ± 0.035	0.430 ± 0.054	0.564 ± 0.051	0.968 ± 0.022	0.924 ± 0.047	0.076 ± 0.047
		valid	0.444 ± 0.023	0.291 ± 0.028	0.467 ± 0.054	0.467 ± 0.007	0.351 ± 0.023	0.305 ± 0.022
	GDRO 0	train	0.692 ± 0.174	0.470 ± 0.245	0.530 ± 0.245	0.731 ± 0.063	0.665 ± 0.099	0.335 ± 0.099
		valid	0.561 ± 0.042	0.423 ± 0.019	0.404 ± 0.071	0.589 ± 0.010	0.464 ± 0.021	0.369 ± 0.052
	GDRO 0 + SR	train	0.869 ± 0.117	0.647 ± 0.159	0.329 ± 0.130	0.901 ± 0.007	0.781 ± 0.014	0.219 ± 0.014
		valid	0.538 ± 0.035	0.323 ± 0.024	0.474 ± 0.053	0.538 ± 0.022	0.405 ± 0.026	0.362 ± 0.052
	GDRO 1	train	0.680 ± 0.102	0.581 ± 0.136	0.370 ± 0.158	0.787 ± 0.059	0.749 ± 0.075	0.251 ± 0.075
		valid	0.565 ± 0.016	0.457 ± 0.028	0.298 ± 0.025	0.598 ± 0.014	0.465 ± 0.011	0.348 ± 0.020
	GDRO 1 + SR	train	0.787 ± 0.180	0.663 ± 0.225	0.260 ± 0.137	0.884 ± 0.066	0.722 ± 0.184	0.278 ± 0.184
		valid	0.485 ± 0.056	0.323 ± 0.013	0.443 ± 0.141	0.537 ± 0.010	0.413 ± 0.017	0.341 ± 0.039
CIFAR10	GDRO 2	train	0.662 ± 0.149	0.564 ± 0.165	0.404 ± 0.140	0.726 ± 0.060	0.667 ± 0.075	0.333 ± 0.075
		valid	0.564 ± 0.034	0.426 ± 0.015	0.376 ± 0.023	0.571 ± 0.009	0.436 ± 0.006	0.341 ± 0.030
	GDRO 2 + SR	train	0.703 ± 0.085	0.540 ± 0.093	0.328 ± 0.149	0.914 ± 0.020	0.808 ± 0.032	0.192 ± 0.032
		valid	0.477 ± 0.020	0.336 ± 0.041	0.429 ± 0.057	0.546 ± 0.015	0.399 ± 0.007	0.375 ± 0.029
	GDRO 5	train	0.644 ± 0.088	0.550 ± 0.097	0.371 ± 0.077	0.699 ± 0.058	0.623 ± 0.066	0.377 ± 0.066
		valid	0.532 ± 0.011	0.420 ± 0.002	0.289 ± 0.016	0.578 ± 0.026	0.439 ± 0.010	0.367 ± 0.084
	GDRO 5 + SR	train	0.776 ± 0.166	0.451 ± 0.236	0.506 ± 0.228	0.962 ± 0.030	0.912 ± 0.068	0.088 ± 0.068
		valid	0.486 ± 0.058	0.318 ± 0.007	0.380 ± 0.137	0.571 ± 0.008	0.428 ± 0.015	0.272 ± 0.040
	SGDRO	train	0.497 ± 0.149	0.379 ± 0.117	0.497 ± 0.027	0.901 ± 0.078	0.884 ± 0.087	0.116 ± 0.087
		valid	0.450 ± 0.036	0.291 ± 0.032	0.338 ± 0.055	0.512 ± 0.014	0.376 ± 0.021	0.316 ± 0.020
	SGDRO + SR	train	0.645 ± 0.134	0.347 ± 0.122	0.631 ± 0.089	0.954 ± 0.054	0.884 ± 0.133	0.116 ± 0.133
		valid	0.440 ± 0.029	0.327 ± 0.009	0.260 ± 0.030	0.456 ± 0.010	0.355 ± 0.013	0.262 ± 0.049

Table 6.14. Performance gap between the train and valid for selected penalty function on CIFAR10 dataset. The best function is chosen according to the valid performance and works well in training data where the data distribution can differ. Validation performance in this table is the approximation of the test performance represented in other tables.

		Waterbird			CelebA		
		Average Acc.	Robust Acc.	Accuracy Gap.	Average Acc.	Robust Acc.	Accuracy Gap.
ERM	train	0.929 ± 0.026	0.904 ± 0.036	0.096 ± 0.036	0.860 ± 0.013	0.822 ± 0.017	0.150 ± 0.016
	valid	0.734 ± 0.064	0.661 ± 0.010	0.299 ± 0.009	0.826 ± 0.024	0.793 ± 0.003	0.164 ± 0.002
ERM + SR	train	0.906 ± 0.024	0.738 ± 0.057	0.262 ± 0.057	0.940 ± 0.007	0.859 ± 0.006	0.140 ± 0.006
	valid	0.807 ± 0.017	0.793 ± 0.006	0.114 ± 0.005	0.874 ± 0.006	0.866 ± 0.002	0.040 ± 0.005
GDRO 0	train	0.841 ± 0.025	0.809 ± 0.024	0.191 ± 0.024	0.897 ± 0.004	0.889 ± 0.004	0.109 ± 0.005
	valid	0.832 ± 0.013	0.819 ± 0.015	0.063 ± 0.025	0.886 ± 0.005	0.880 ± 0.004	0.023 ± 0.005
GDRO 0 + SR	train	0.912 ± 0.027	0.752 ± 0.076	0.248 ± 0.076	0.938 ± 0.005	0.872 ± 0.008	0.126 ± 0.009
	valid	0.806 ± 0.005	0.775 ± 0.029	0.145 ± 0.026	0.877 ± 0.005	0.868 ± 0.006	0.035 ± 0.021
GDRO 1	train	0.878 ± 0.010	0.849 ± 0.021	0.151 ± 0.021	0.899 ± 0.016	0.886 ± 0.018	0.113 ± 0.017
	valid	0.839 ± 0.010	0.825 ± 0.008	0.073 ± 0.011	0.884 ± 0.007	0.879 ± 0.003	0.029 ± 0.003
GDRO 1 + SR	train	0.931 ± 0.040	0.808 ± 0.064	0.188 ± 0.061	0.929 ± 0.011	0.857 ± 0.019	0.139 ± 0.016
	valid	0.816 ± 0.008	0.801 ± 0.017	0.112 ± 0.010	0.874 ± 0.003	0.872 ± 0.002	0.024 ± 0.008
GDRO 2	train	0.876 ± 0.016	0.846 ± 0.018	0.154 ± 0.018	0.893 ± 0.012	0.877 ± 0.013	0.123 ± 0.013
	valid	0.839 ± 0.014	0.832 ± 0.009	0.030 ± 0.017	0.883 ± 0.004	0.880 ± 0.002	0.045 ± 0.003
GDRO 2 + SR	train	0.946 ± 0.001	0.836 ± 0.008	0.164 ± 0.008	0.934 ± 0.011	0.861 ± 0.017	0.137 ± 0.015
	valid	0.826 ± 0.006	0.817 ± 0.010	0.060 ± 0.022	0.876 ± 0.005	0.872 ± 0.002	0.028 ± 0.009
GDRO 5	train	0.865 ± 0.011	0.848 ± 0.016	0.152 ± 0.016	0.898 ± 0.006	0.880 ± 0.007	0.120 ± 0.007
	valid	0.851 ± 0.006	0.836 ± 0.009	0.034 ± 0.024	0.885 ± 0.005	0.880 ± 0.004	0.026 ± 0.012
GDRO 5 + SR	train	0.926 ± 0.001	0.830 ± 0.008	0.170 ± 0.008	0.946 ± 0.001	0.868 ± 0.009	0.132 ± 0.009
	valid	0.830 ± 0.009	0.823 ± 0.003	0.044 ± 0.003	0.875 ± 0.007	0.869 ± 0.003	0.042 ± 0.004
SGDRO	train	0.805 ± 0.014	0.770 ± 0.013	0.230 ± 0.013	0.884 ± 0.008	0.865 ± 0.010	0.133 ± 0.011
	valid	0.819 ± 0.008	0.805 ± 0.005	0.070 ± 0.002	0.877 ± 0.005	0.871 ± 0.004	0.054 ± 0.006
SGDRO + SR	train	0.910 ± 0.005	0.757 ± 0.007	0.243 ± 0.007	0.928 ± 0.011	0.853 ± 0.020	0.145 ± 0.019
	valid	0.805 ± 0.025	0.796 ± 0.017	0.091 ± 0.015	0.873 ± 0.007	0.868 ± 0.005	0.029 ± 0.012

Table 6.15. Performance gap between the train and valid for selected penalty function in the attribute-based classification task. The best function is chosen according to the valid performance and works well in training data where the data distribution can differ. Validation performance in this table is the approximation of the test performance represented in other tables.

Bibliography

[1] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, 2017.

[2] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[4] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[7] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

[8] Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: closing the subgroup performance gap with data augmentation. In *ICLR*, 2021.

[9] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[10] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, pages 3315–3323, 2016.

[11] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] 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 (CVPR)*, June 2016.

[14] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: a systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

[15] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Felix Last, Georgios Douzas, and Fernando Bacao. Oversampling for imbalanced learning based on k-means and smote. *arXiv preprint arXiv:1711.00837*, 2017.

[18] Charles X. Ling and Chenghui Li. Data mining for direct marketing: problems and solutions. In *Knowledge Discovery and Data Mining*, pages 73–79, 1998.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[20] Pranay Lohia. Priority-based post-processing bias mitigation for individual and group fairness. *arXiv preprint arXiv:2102.00417*, 2021.

[21] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE, 2019.

[22] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

[23] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393. PMLR, 10–15 Jul 2018.

[24] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[26] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS*, 2021.

[27] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. FR-train: A mutual information-based approach to fair and robust training. 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 8147–8157. PMLR, 13–18 Jul 2020.

[28] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 815–827. Curran Associates, Inc., 2021.

[29] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*, 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. In *ICLR*, 2020.

[31] Sahil Verma and Julia Rubin. Fairness definitions explained. In *International Workshop on Software Fairness (Fairware)*, 2018.

[32] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

[33] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[34] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio Calmon. Optimized score transformation for fair classification. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1673–1683. PMLR, 26–28 Aug 2020.

[35] Show-Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3, Part 1):5718–5727, 2009.

[36] Jy yong Sohn, Liang Shang, Hongxu Chen, Jaekyun Moon, Dimitris S. Papailiopoulos, and Kangwook Lee. Genlabel: Mixup relabeling using generative models. *CoRR*, abs/2201.02354, 2022.

[37] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, page 335–340, 2018.

[38] Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. 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 12857–12867. PMLR, 18–24 Jul 2021.

[39] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006.

[40] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthaus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *CVPR*, 2022.

Abstract in Korean

특정 범주에서는 충분한 표본을 얻는 것이 굉장히 힘들기 때문에 실제 문제에서 범주 별로 균형 잡힌 큰 규모의 데이터셋을 구축하는 것은 현실적으로 어려움이 있다. 이러한 불균형은 주로 소수 범주의 자연적인 특징이나 잠재적인 특성의 계층적인 구조에 의한다. 이는 집단 간의 성능 차이나 불공정성을 야기한다. 소수 집단을 강조하는 방식으로 그룹들을 공정하게 대하려는 다양한 방법들이 존재한다. 또한 데이터 증대나 생성 모델들 또한 이러한 문제를 해결하여 일반화 성능을 증대하기 위해 사용되었다. 하지만 이러한 접근법들은 소수 집단의 다양성 부족에 의한 과접합 문제에 의한 부정적인 영향을 완전히 제거하는 데에는 실패했다. 본 논문에서는 분류기들이 각자의 예측을 과신하는 경향을 실증한다. 또한 과반수 범주에 대한 예측에 불이익을 가해 소수 범주의 성능을 증대시키는 새로운 후처리 방법을 제안한다. 이 방법은 기존의 방법들과 양립하며, 최적의 불이익 함수를 얻기 위한 적응 알고리즘 또한 제안한다. 본 방법은 데이터의 불균형이나 편향에 강건한 예측 경계를 구축하는 새로운 관점을 제시한다. 다양한 데이터셋과 불균형 환경에 대한 다양한 실험 결과들을 통해 평균 성능과 최소 성능 두 측면에서의 상당한 향상을 보이고 새로운 예측 경계의 장점을 제안한다.

주요어 : 불균형 학습, 후처리 방법, 이미지 분류
학 번 : 2021-27158