EXACT GROUP FAIRNESS REGULARIZATION VIA CLASSWISE ROBUST OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing group fairness-aware training methods typically employ some heuristics, such as re-weighting underrepresented groups based on some rules or using approximated surrogates for the exact fairness metrics as regularization terms, which result in models with sub-optimal accuracy-fairness trade-offs. The reason for using such heuristics is that the fairness metrics are usually non-differentiable or non-convex, and exactly incorporating those metrics in a tractable learning objective is challenging. To that end, we propose a principled method that indeed can incorporate an *exact* form of a well-justified group fairness metric, Difference of Conditional Accuracy (DCA), as a regularizer using a *classwise* distributionally robust optimization (DRO) framework. Namely, we first show that the DCA is equivalent (up to a constant) to the average (over the classes) of the roots of the *variances* of group losses, then employ the Group DRO formulation for each class *separately* to convert the non-differentiable DCA (or variance) regularized group-balanced empirical risk minimization to a more tractable minimax optimization. We further develop an efficient iterative optimization algorithm and show that our resulting method, dubbed as FairDRO, makes an interesting connection between the re-weighting based and regularization-based fairness-aware learning. Our experiments show that FairDRO is scalable, easily adaptable to diverse applications, and consistently improves the group fairness on several benchmark datasets in terms of the accuracy-fairness trade-off, compared to recent state-of-the-art baselines.

1 INTRODUCTION

Machine learning algorithms are increasingly used in learning various decision-making applications that have societal impact; *e.g.*, crime assessment (Julia Angwin & Kirchner, 2016), credit estimation (Khandani et al., 2010), facial recognition (Buolamwini & Gebru, 2018; Wang et al., 2019), automated filtering in social media (Fan et al., 2021), AI-assisted hiring (Nguyen & Gatica-Perez, 2016) and law enforcement (Garvie, 2016). A critical issue in such application is the potential discrepancy of model performance, *e.g.*, accuracy, across different sensitive groups (*e.g.*, race or gender) (Buolamwini & Gebru, 2018), which is easily observed in the models trained with a vanilla empirical risk minimization (ERM) (Valiant, 1984) when the training data has unwanted bias. To address such issue, the *fairness-aware* learning has recently drawn attention in the AI research community.

One of the objectives of fairness-aware learning is to achieve *group fairness*, which focuses on the statistical parity of the model prediction across sensitive groups. The so-called *in-processing* methods typically employ additional machinery for achieving group fairness while training, *e.g.*, re-weighting sample weights or adding fairness-promoting regularizations. However, since exactly enforcing the given group fairness criterion into the training objective typically results in a non-convex optimization problem, most methods boil down to either using heuristic rules for re-weighting underrepresented samples (*e.g.*, RW (Kamiran & Calders, 2012) or LBC (Jiang & Nachum, 2020)) or using approximated surrogate loss for group fairness as a regularization term (*e.g.*, covariance (Zafar et al., 2017b;a), HSIC (Quadrianto et al., 2019), Rényi (Baharlouei et al., 2020), or MMD (Jung et al., 2021b)). To that end, while the in-processing methods have been easily adopted to train *fair* complex models like DNN on large-scale datasets, such as vision (Jung et al., 2021b; Quadrianto et al., 2019; Roh et al., 2020) datasets, most of them actually lack theoretically clear justifications for achieving group fairness explicitly.

In this paper, we devise a new in-processing method in a more principled way, dubbed as Fairness-aware DRO (FairDRO). The goal of our method is to incorporate an exact form, not a surrogate, of a well-justified group fairness metric, *Difference of Conditional Accuracy* (DCA) (Berk et al., 2021), as a regularizer in the training objective. To this end, we first present that DCA is equivalent (up to a constant) to the average (over the classes) of the roots of the *variances* of groupwise 0-1 losses. Then, we employ the Group DRO formulation, which uses the χ^2-divergence ball including *quasi-probabilities* as the uncertainty set, for each class *separately* to convert the non-differentiable DCA (or variance) regularized group-balanced empirical risk minimization (ERM) to a more tractable minimax optimization. We further develop an efficient iterative optimization algorithm, which automatically produces the correct (sometimes even negative) re-weights for the samples in each group during the optimization process, making an interesting unified connection between the re-weighting and regularization-based fairness-aware learning schemes.

In our experiments, we empirically show that our FairDRO is scalable and easily adaptable to diverse application scenarios, including tabular (Julia Angwin & Kirchner, 2016; Dua et al., 2017), vision (Zhang et al., 2017), and natural language text (Koh et al., 2021) datasets. We compare with several representative in-processing baselines that apply either re-weighting schemes or surrogate regularizations for group fairness, and show that FairDRO consistently achieves the state-of-the-art performance on all datasets in terms of accuracy-fairness trade-off.

2 Related works

2.1 Fairness-aware training

A comprehensive study on group fairness-aware learning has been conducted under the following three categories (Caton & Haas, 2020): *pre-processing*, *in-processing*, and *post-processing* methods. In this paper, we focus on the in-processing methods, because they can flexibly incorporate the fairness notion during training without losing information from pre-processing the data or being limited by a pre-trained model as in the post-processing methods.

In-processing methods for group fairness Recently proposed in-processing methods can be categorized into *re-weighting based* methods (Kamiran & Calders, 2012; Jiang & Nachum, 2020; Roh et al., 2020; Agarwal et al., 2018) and *regularization based* methods (Zafar et al., 2017b; Quadrianto et al., 2019; Baharlouei et al., 2020; Jung et al., 2021b; Chuang & Mroueh, 2021). The former methods typically determine to give higher weights to the samples from the underrepresented groups. For example, Kamiran & Calders (2012) proposed a re-weighting scheme (RW) based on the ratio of the number of data points per each group. Label Bias Correction (LBC) (Jiang & Nachum, 2020) and FairBatch (Roh et al., 2020) adaptively adjust weights and mini-batch sizes based on the average loss per group, respectively. Although they can be easily adapted to various domains due to their simplicity, they lack sound theoretical justifications for enforcing group fairness and result in poor performance on some benchmark datasets. On the other hand, the regularization-based methods apply surrogate constraints to their training objectives; *e.g.*, Cov (Zafar et al., 2017b), Rényi (Baharlouei et al., 2020), and FairHSIC (Quadrianto et al., 2019) employ as fairness constraints a covariance approximation, Rényi correlation, and Hilbert Schmidt Independence Criterion (Gretton et al., 2005) w.r.t. a group label and a model prediction, respectively. Jung et al. (2021b) devised MFD that uses Maximum Mean Discrepancy (MMD) (Gretton et al., 2005) based regularization for distilling knowledge and promoting group fairness at the same time. Despite their success, we note that their performance could be sub-optimal due to using the surrogates of the fairness metrics during learning.

2.2 Distributionally robust optimization (DRO)

Distributionally robust optimization (DRO) (Ben-Tal et al., 2009) is a general robust optimization framework that was originally used for learning a model that is robust to potential test distribution shift from the training set, by optimizing for the worst-case distribution in an uncertainty set. The DRO framework is now widely adopted in other applications, including algorithmic fairness.

DRO in algorithmic fairness Numerous works in algorithmic fairness literature have recently utilized the DRO frameworks in various settings. For example, a line of works (Jiang et al., 2020; Taskesen et al., 2020; Wang et al., 2021; Mandal et al., 2020) embeds the DRO frameworks into the fairness-constrained optimization problem to make a fair model robust to distribution shifts of a test

dataset. Another related line of works (Wang et al., 2020; Hashimoto et al., 2018) proposed DRO formulations for achieving fairness in more challenging settings such as learning with noisy group labels or sequential learning. In spite of their effectiveness, both of above lines of work do not use DRO as a direct tool for achieving group fairness. Perhaps the closest related work to ours would be (Rezaei et al., 2020; Yurochkin et al., 2019), in which DRO frameworks are adopted for the purpose of achieving fairness. However, the key difference is that they either require the group labels at test time or target *individual* fairness. On the other hand, our FairDRO directly aims to achieve group fairness while not requiring group labels at test time. A more detailed discussion is in Appendix B.

DRO in other applications Applications of DRO are actively studied in other contexts, *e.g.*, debiasing (Sagawa et al., 2020; Lahoti et al., 2020), multilingual machine translation (Oren et al., 2019; Zhou et al., 2021), long-tailed classification (Samuel & Chechik, 2021) or out-of-distribution generalization (Krueger et al., 2021; Xie et al., 2020). For example, Sagawa et al. (2020) proposed Group DRO which constructs the uncertainty set of joint distributions over the class-group label pairs to capture the spurious correlations between the class and group labels. Although these works are effective in their applications, however, the notion of statistical parity across groups (group fairness) has not been explicitly pursued as in our FairDRO.

3 Notations and fairness criterion

Notations We consider a classification problem where each sample consists of an input $\boldsymbol{x} \in \mathcal{X}$, a class label $y \in \mathcal{Y}$ and a group label $a \in \mathcal{A}$. The group label is defined by sensitive attributes such as *gender* or *race*. Given a training dataset $\mathcal{D}$ with N samples (*i.e.*, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^N$), we aim to find a fair classifier $\boldsymbol{\theta}^{\text{FAIR}} : \mathcal{X} \to \mathcal{Y}$ satisfying the given fairness criterion while maintaining the classification accuracy. Moreover, given a loss function $\ell(\boldsymbol{\theta}, (\boldsymbol{x}, y)) : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, we denote $\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}) \triangleq \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \ell(\boldsymbol{\theta}, (\boldsymbol{x}_i, y_i))$. Additionally, $\mathcal{D}_a$ and $\mathcal{D}_a^y$ denote the subsets of $\mathcal{D}$ that are confined to the samples with group label a or class-group label pair (y, a), respectively.

Fairness criterion The notion of fairness may depend on the different points of view on how discrimination is defined (Hardt et al. (2016); Dwork et al. (2012); Chouldechova (2017); Berk et al. (2021)). For example, one may argue that a classifier is fair if its prediction is not dependent on a (*Demographic Parity*), while the other can insist that the classifier should be conditionally independent on a (*Equalized Odds*). In our work, we focus on the *Equalized Conditional Accuracy* (ECA) (Berk et al., 2021), which can naturally cover the multi-class, multi-group label setting.

A classifier $\boldsymbol{\theta}$ satisfies ECA if all of the accuracies among groups are the same for each given class, *i.e.*, $\forall a, a' \in \mathcal{A}, y \in \mathcal{Y}, P(\widehat{Y} = y | A = a, Y = y) = P(\widehat{Y} = y | A = a', Y = y)$, where $\widehat{Y}$ is the prediction of $\boldsymbol{\theta}$. We measure the degree of unfairness with *Difference of Conditional Accuracy* (DCA), denoted as Δ_{DCA}, by taking the average of the maximum accuracy gaps among groups over y:

$$\Delta_{\text{DCA}} \triangleq \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \Delta_y, \ \Delta_y \triangleq \max_{a,a'} |P(\widehat{Y} = y | A = a, Y = y) - P(\widehat{Y} = y | A = a', Y = y)|. \quad (1)$$

Remark: We note that ECA has a close connection with the two popular group fairness criteria, *Equalized Odds* (EO) and *Equal Opportunity* (EOpp) (Hardt et al. (2016)). EO requires the conditional independence between $\widehat{Y}$ and A given Y, hence, we can easily observe that ECA is *equivalent* to EO for the binary classification setting. Furthermore, we argue that ECA is a natural extension of EOpp to the multi-class setting; namely, ECA requires equality of TPR among groups for each class label, which resembles the argument for justifying EOpp.

4 Fairness-aware DRO (FairDRO)

Our goal is to incorporate the *exact* DCA (1), not its approximated surrogate, in the training objective. However, since DCA is non-differentiable, it is not straightforward to use DCA directly as a regularization term. To that end, we first show in Sec. 4.1 that DCA is equivalent to the average (over the classes) of the roots of the *variances* of groupwise 0-1 losses. Then, we utilize the previous result (Xie et al., 2020) in Sec. 4.2 to make the connection between the Group DRO with the uncertainty set of

χ^2-divergence ball including the quasi-probability and the variance regularized group-balanced empirical loss minimization. Subsequently, in Sec. 4.3, we present our training objective that applies Group DRO *separately* for each class y, which essentially converts the non-differentiable DCA (or variance) regularized group-class balanced ERM to a more tractable minimax optimization problem. Finally, Sec. 4.4 proposes an efficient iterative algorithm that does alternating minimization-maximization of our training objective.

4.1 Equivalence between DCA and variance of groupwise losses

Proposition 1 *Assume $\ell(\boldsymbol{\theta}) = \mathbb{1}\{\boldsymbol{\theta}(\boldsymbol{x}) \neq y\}$. Then, the following inequality holds for all $y \in \mathcal{Y}$,*

$$\sqrt{2\,\mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\}_{a\in\mathcal{A}}\right)} \leq \widehat{\Delta}_y \leq \sqrt{2|\mathcal{A}|^2\,\mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\}_{a\in\mathcal{A}}\right)}, \tag{2}$$

in which $\widehat{\Delta}_y$ is the empirical version of Δ_y and $\mathrm{Var}(\{x_i\}_{i=1}^d) \triangleq \frac{1}{d}\sum_i (x_i - \frac{1}{d}\sum_i x_i)^2$ is the variance of d numbers $\{x_i\}_{i=1}^d$. Thus, $\widehat{\Delta}_y$ of a classifier is 0 if and only if $\mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\}_{a\in\mathcal{A}}\right) = 0$.

The proof is in Appendix A. By taking the average over y in (2), Proposition 1 implies that DCA is equivalent to the average (over y) of the square roots of the variances of the groupwise 0-1 losses, up to a constant. Building this equivalence, we are now able to utilize the Group DRO formulation given in the next subsection to make a connection between DCA regularization and robust optimization.

4.2 Prior work on variance regularization via Group DRO

While ERM minimizes the empirical loss, DRO optimizes for the worst-case distributions in an uncertainty set, as mentioned in Sec. 2.2. Since general DRO (Duchi et al., 2019) may lead to a pessimistic model that optimizes for implausible worst-case distributions, Sagawa et al. (2020) recently proposed the *Group* DRO framework, which performs at the group level and can incorporate prior knowledge about the groups[1]; it aims to obtain

$$\boldsymbol{\theta}^{\mathrm{GDRO}} \triangleq \arg\min_{\boldsymbol{\theta}} \max_{\boldsymbol{q}\in\Delta^{|\mathcal{A}|}} \sum_{a\in\mathcal{A}} q_a \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a), \tag{3}$$

in which $\Delta^{|\mathcal{A}|}$ is the $|\mathcal{A}|$-simplex.

The applications of Group DRO are actively studied in various contexts, *e.g.*, debiasing (Sagawa et al., 2020) or machine translation (Zhou et al., 2021). In particular, Xie et al. (2020) proposed Risk Variance Penalization (RVP), an OOD generalization method, using a variance regularization with respect to risks of each domain (a formulation first introduced by V-REx (Krueger et al., 2021)). The authors showed the connection between the variance regularization and Group DRO, when the uncertainty set is the following χ^2-divergence ball with radius $\rho > 0$

$$\mathcal{Q}_\rho \triangleq \Big\{\boldsymbol{q} \in \mathbb{R}^{|\mathcal{A}|} : \sum_{a\in\mathcal{A}} q_a = 1,\ D_\phi\Big(\boldsymbol{q}\,\Big|\Big|\,\frac{1}{|\mathcal{A}|}\mathbb{1}\Big) \leq \rho\Big\}, \tag{4}$$

in which $D_\phi(p\|q) \triangleq \sum_i q_i \phi(p_i/q_i)$ with $\phi(t) = (t-1)^2$, and $\frac{1}{|\mathcal{A}|}\mathbb{1} \in \mathbb{R}^{|\mathcal{A}|}$ denotes the uniform distribution. Note (4) does not have the nonnegativity constraint of $\boldsymbol{q}$, hence, depending on ρ, $\mathcal{Q}_\rho$ can also include *quasi-probabilities*, which has negative components as well. Namely, Xie et al. (2020) shows the following lemma that the inner maximization of Group DRO with $\mathcal{Q}_\rho$ uncertainty set is equivalent to the group-balanced empirical loss plus the group-wise variance term as a regularizer.

Lemma 1 *(Xie et al., 2020, Proposition 1) For any finite loss $\ell(\boldsymbol{\theta}, (\boldsymbol{x}, y))$ and with $\mathcal{Q}_\rho$ in (4), we have*

$$\max_{\boldsymbol{q}\in\mathcal{Q}_\rho} \sum_{a\in\mathcal{A}} q_a \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a) = \frac{1}{|\mathcal{A}|}\sum_a \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a) + \sqrt{\rho\,\mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a)\}_{a\in\mathcal{A}}\right)}. \tag{5}$$

To be self-contained, the proof of the lemma is given in Appendix A.

[1]The meaning of the word "group" is used slightly differently only in this subsection. The group here is used more inclusively. Namely, the group can be defined by information of any hierarchical structure in the data distribution, *e.g.*, domain, environment, and sensitive attribute.

4.3 TRAINING OBJECTIVE OF FAIRDRO

To make the connection between Proposition 1 and Lemma 1 and realize the exact DCA regularization for fairness-aware training, we propose to apply the Group DRO *separately* for each class y. Namely, we define our FairDRO as minimizing the average of the worst-case losses defined for each class y:

$$\boldsymbol{\theta}^{\text{FairDRO}} \triangleq \arg\min_{\boldsymbol{\theta}} \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \max_{\boldsymbol{q}^y \in \mathcal{Q}_\rho} \sum_{a \in \mathcal{A}} q_a^y \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y), \tag{6}$$

in which $\mathcal{Q}_\rho$ is as defined in (4), $\boldsymbol{q}^y = \{q_a^y\}_{a \in \mathcal{A}}$ is a quasi-probability vector, and ρ is a tunable hyperparameter. We stress that while applying Group DRO by treating each pair (y, a) as a "group" has been considered in Sagawa et al. (2020), applying Group DRO separately for each class as in FairDRO has not been considered before. Now, we present our key result which follows by combining Proposition 1 and Lemma 1.

Corollary 1 *Assume $\ell(\boldsymbol{\theta}) = \mathbb{1}\{\boldsymbol{\theta}(\boldsymbol{x}) \neq y\}$. Then, $\boldsymbol{\theta}^{\text{FairDRO}}$ also achieves*

$$\min_{\boldsymbol{\theta}} \Big\{ \frac{1}{|\mathcal{Y}||\mathcal{A}|} \sum_{(y,a)} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) + C\sqrt{\rho}\widehat{\Delta}_{DCA} \Big\}, \tag{7}$$

for some positive constant C in the range of $[\sqrt{2}, \sqrt{2}|\mathcal{A}|]$.

The corollary implies that when the 0-1 loss is used, FairDRO is also a *regularization-based method*, as described in Sec. 2.1, with the *exact* DCA serving as a regularizer. Note we do not solve the minimization in (7) directly; however, since $\widehat{\Delta}_{\text{DCA}}$ can be biased when estimated from a small mini-batch, particularly when the number of groups is large, and is hard to minimize due to non-differentiability. Instead, we solve a much more tractable minimax optimization in (6) where the maximization is a simple linear optimization and the minimization is simply for the *re-weighted* group losses, and *no* estimation of $\widehat{\Delta}_{\text{DCA}}$ is involved.

At this point, it is now clear that during solving (6), FairDRO produces the re-weights $\{q_a^y\}$'s for each (y, a), which makes it also categorized as a *re-weighting based method*. The following proposition shows how the re-weights are characterized and controlled by the hyperparameter ρ.

Proposition 2 *For $\mathcal{Q}_\rho$ in (4) with $\rho > 0$, each element of the optimum $\boldsymbol{q}^{*y}$ for the inner-maximization of (6) can be obtained by a closed-form solution:*

$$q_a^{*y} = \frac{1}{|\mathcal{A}|} + \sqrt{\frac{\rho}{|\mathcal{A}|}} \times \frac{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a \in \mathcal{A}} \frac{1}{|\mathcal{A}|} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)}{\|\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a \in \mathcal{A}} \frac{1}{|\mathcal{A}|} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\|_2}, \tag{8}$$

and is in the range of $\Big[\frac{1}{|\mathcal{A}|} - \sqrt{\rho}\frac{\sqrt{(|\mathcal{A}|-1)}}{|\mathcal{A}|}, \frac{1}{|\mathcal{A}|} + \sqrt{\rho}\frac{\sqrt{(|\mathcal{A}|-1)}}{|\mathcal{A}|}\Big]$.

The proof is in Appendix A. Proposition 2 shows that some elements of $\boldsymbol{q}^{*y}$ can be negative depending on ρ and the groupwise loss $\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)$. Moreover, it shows that the range of each element of the optimum $\boldsymbol{q}^{*y}$ is automatically determined by a radius ρ and the number of groups $|\mathcal{A}|$, which means FairDRO controls the degree of the penalization by varying ρ. We will show in Sec. 5 that we can find a proper degree of penalization by tuning ρ so that the performances on various datasets in terms of accuracy-fairness trade-off are significantly improved.

A crucial difference from typical re-weighting based methods is that our re-weights are obtained as the solution of an optimization problem as opposed to those in (Kamiran & Calders, 2012; Jiang & Nachum, 2020) that are heuristically obtained based on the ratio of the number of data points or losses for each group. Also, we consider negative values for q_a^y in $\mathcal{Q}_\rho$. Such negative values for q_a^y enable penalizing a high accuracy group more aggressively than typical re-weighting methods that only control the *positive* weights for each group a and class y, to achieve group fairness.

Remark: For Corollary 1 to hold, we use 0-1 loss in (6), which makes our training objective still non-differentiable with respect to $\boldsymbol{\theta}$. Thus, when solving (6), we use the cross-entropy loss for the outer minimization, whereas the 0-1 loss is used for the inner maximization. Since the cross-entropy is

the convex surrogate for 0-1 loss, and performing ERM with a convex surrogate has a well-established theory (Mohri et al., 2018) in practice, the near-optimal solution of (6) can be achieved by our strategy. We again note that the non-differentiable regularization term in (7) is substituted into a more feasible problem through our DRO formulation.

4.4 An efficient iterative optimization for FairDRO

A canonical way of solving the minimax optimization in DRO is to use a primal-dual method (Nemirovski et al., 2009) in which one alternates between a regular gradient descent on $\boldsymbol{\theta}$ and an exponentiated gradient (EG) ascent (Kivinen & Warmuth, 1997) on $\boldsymbol{q}$. However, the EG algorithm is widely used when the variables lie in the probability space because it is a special case of mirror descent when the convex function for the Bregman divergence is the Shannon Entropy. Thus, EG cannot be applied to solving (6) since our uncertainty set also allows the quasi-probability for $\boldsymbol{q}^y$.

Hence, we employ the *smoothed* Iterated Best Response (IBR) which is a variant of IBR recently used for solving a DRO problem in (Zhou et al., 2021). Using the standard IBR, we update $\boldsymbol{q}$ as the closed-form solution of the maximization (*i.e.*, (8)), instead of the gradient ascent of $\boldsymbol{q}$. It is known that such IBR procedure eventually converges to equilibrium under some mild regularity assumptions (Zhou et al., 2021). However, we empirically observed the instability during training; *i.e.*, the $\boldsymbol{q}^{*y,t}$ oscillates when the loss for each group fluctuates, hence, we applied the learning rate scheduling of $\eta_{\boldsymbol{q}}^t = 1 - \frac{t}{T}$ with T being the number of total epochs to update $\boldsymbol{q}^{y,t}$ *smoothly*. Namely, our update rules are

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta_{\boldsymbol{\theta}}^t \nabla_{\boldsymbol{\theta}} \Big(\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \sum_{a \in \mathcal{A}} q_a^{y,t} \mathcal{L}(\boldsymbol{\theta}^t, \mathcal{D}_a^y) \Big), \quad \text{(update } \boldsymbol{\theta}\text{)} \tag{9}$$

$$\boldsymbol{q}^{y,t+1} \leftarrow (1 - \eta_{\boldsymbol{q}}^t) \boldsymbol{q}^{y,t} + \eta_{\boldsymbol{q}}^t \boldsymbol{q}^{*y,t}, \ \ \forall y \in \mathcal{Y}, \quad \text{(update } \boldsymbol{q}^y\text{)} \tag{10}$$

$$\text{in which } \ \boldsymbol{q}^{*y,t} \triangleq \ \underset{\boldsymbol{q}^y \in \mathcal{Q}_\rho}{\arg\max} \sum_{a \in \mathcal{A}} q_a^y \mathcal{L}(\boldsymbol{\theta}^{t+1}, \mathcal{D}_a^y) \tag{11}$$

Algorithm 1: FairDRO Iterative Optimization

Input : Dataset $\{\mathcal{D}_a^y\}_{y \in \mathcal{Y}, a \in \mathcal{A}}$, radius ρ, epoch T, iteration I, mini-batch size B

Randomly initialize $\boldsymbol{\theta}^0$
Set $\boldsymbol{q}^{y,0} \leftarrow \frac{1}{|\mathcal{A}|}\mathbf{1}$ for $y \in \mathcal{Y}$, in which $\mathbf{1} \in \mathbb{R}^{|\mathcal{A}|}$
for $t = 0$ **to** $T - 1$ **do**
  // $\boldsymbol{\theta}$-update
  for $i = 0$ **to** $I - 1$ **do**
    Sample a mini-batch $\{(\boldsymbol{x}_j, y_j, a_j)\}_{j \in [B]}$, equally from $\mathcal{D}_a^y$ for each $(y, a) \in \mathcal{Y} \times \mathcal{A}$.
    Update $\boldsymbol{\theta}^t$ with (9) using the mini-batch
  end
  // $\boldsymbol{q}$-update
  Update $\boldsymbol{q}^{y,t}$ for each $y \in \mathcal{Y}$ with (10)
end

Output : $\boldsymbol{\theta}^T$, $\{\boldsymbol{q}^{y,T}\}_{y \in \mathcal{Y}}$

and $\eta_{\boldsymbol{\theta}}^t$ and $\eta_{\boldsymbol{q}}^t$ denote the learning rates for $\boldsymbol{\theta}$ and $\boldsymbol{q}^y$, respectively. By applying smoothing for update $\boldsymbol{q}^y$ as in (10) , we empirically observe improvements in the group fairness and more stable convergence of $\boldsymbol{q}^y$ (result given in Appendix D.5). Algorithm 1 illustrates the summary of FairDRO optimization scheme.

5 Experimental results

In this section, we demonstrate the versatility of FairDRO by comparing it with various baseline methods on three modalities of datasets: (1) two tabular datasets, UCI Adult (Dua et al., 2017) and ProPublica COMPAS (Julia Angwin & Kirchner, 2016), (2) two vision datasets, UTKFace (Zhang et al., 2017), CelebA (Liu et al., 2015) (in Appendix D.2) and (3) a language dataset, CivilComments-WILDS (Koh et al., 2021). We use logistic regression, ResNet18 (He et al., 2016), and a pre-trained BERT (Devlin et al., 2019; Wolf et al., 2019) as the classifier for each modality of datasets, respectively. For the evaluation, we report the trade-offs between the accuracy and Δ_{DCA} by varying hyperparameters of each method. Since the most datasets are severely skewed toward a specific group or class, we report the balanced accuracy over all combinations of groups and classes (the issue discussed more in Appendix D.6). We further give a thorough ablation study on our design choice in (6), namely, the classwise DRO formulation and the uncertainty set (4) in Sec. 5.4 and Appendix D.3, respectively. We further visualize of re-weights produced by FairDRO in Sec. 5.4. Further implementation details are given in Appendix C.

	Adult		COMPAS	
	Acc. (↑)	Δ_{DCA} (↓)	Acc. (↑)	Δ_{DCA} (↓)
Scratch	76.10	4.79	**63.48**	15.10
RW (Kamiran & Calders, 2012)	77.04	8.42	63.40	13.45
LBC (Jiang & Nachum, 2020)	76.22	4.13	61.12	6.02
FairBatch (Roh et al., 2020)	76.29	4.24	61.18	5.95
Cov (Zafar et al., 2017a)	76.09	4.74	60.68	8.93
Rényi (Baharlouei et al., 2020)	72.79	2.67	60.35	5.50
FairDRO	**79.23**	**1.99**	61.97	**5.20**

Table 1: **The best performances on Adult and COMPAS datasets**. The target accuracy (higher is better) and DCA (1) (lower is better) are shown. We follow the proposed model selection criterion. The standard deviation of each result is shown in Appendix D.1.

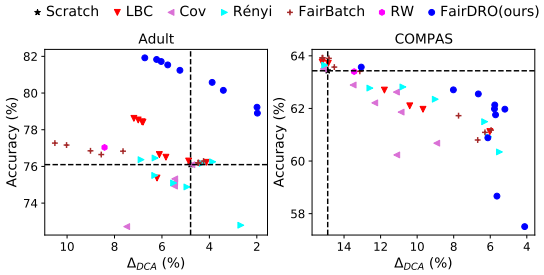

Figure 1: **The trade-offs between accuracy and DCA on Adult (left) and COMPAS (right)**. The "Scatch" performances are highlighted by the dotted lines.

Comparison methods. We compare FairDRO with seven in-processing fair-training methods; Three re-weighting based methods (RW (Kamiran & Calders, 2012), LBC (Jiang & Nachum, 2020), FairBatch (Roh et al., 2020)) and four regularization-based methods (Cov (Zafar et al., 2017b), FairHSIC (Quadrianto et al., 2019), Rényi (Baharlouei et al., 2020), MFD (Jung et al., 2021b)). We note that Cov and Rényi are task-dependent, and FairHSIC and MFD are model-dependent. Namely, Cov is designed only for binary classification tasks. Also, applying Rényi to non-binary group labels is not straightforward with mini-batch training. Meanwhile, FairHSIC and MFD are limited to be only applied for DNN architectures with a sound extractor. Scratch denotes logistic regression model (for Adult and COMPAS), ResNet18 (for UTKFace), and BERT (for CivilComments) trained without any fairness constraints. We note that all baselines except RW were implemented for targeting EO. Considering that ECA is equivalent to EO only in binary classification, we also reported the difference of EO (DEO) for multi-class classification tasks in Appendix D.1. We provide more details of description and implementation of the baselines in Appendix C.

Model selection. The fairness-accuracy trade-off may exist in many real-world applications (Dutta et al., 2020); better accuracy leads to worse fairness, and vice versa. For example, Fig. 1 shows the trade-off for the Adult and COMPAS datasets. Therefore, the model selection criterion is critical for fair evaluation. To that end, we explore varying hyperparameters (HPs) of each method for controlling the strength between accuracy and fairness. We then select the best model showing the best fairness criterion Δ_{DCA} while achieving at least 95% of the accuracy of the vanilla-trained model. The HPs for each method were extensively searched, but fairly in terms of computation budgets. We note that we only report a single score for RW because it has no controllable HP.

5.1 Tabular Datasets

Two tabular datasets, UCI Adult (Dua et al., 2017) (Adult) and ProPublica COMPAS (Julia Angwin & Kirchner, 2016) (COMPAS), are used for the benchmark. Both tasks are binary classification tasks (predicting whether the income of an individual exceeds $50K per a year and whether a defendant re-offends, respectively) with binary sensitive groups ("Female" and "Male", and "Caucasian" and "Non-Caucasin", respectively). Following pre-processing of Bellamy et al. (2018), each dataset includes 45K and 5K rows, respectively.

Tab. 1 and Fig. 1 show the best performances and the fairness-accuracy trade-offs on Adult and COMPAS. In Tab. 1, FairDRO outperforms other methods in the fairness metric (DCA); while the best baselines achieve Δ_{DCA} of 2.67 and 5.50 on Adult and COMPAS, respectively, our FairDRO achieves 1.99 and 5.20 for the same datasets, respectively. We also emphasize that although the accuracy of FairDRO is slightly worse than RW and Scratch on COMPAS, ours show significantly better DCA (5.20) compared to RW (13.45) and Scratch (15.10). In Fig. 1, FairDRO is placed to the most top-right area among the methods, showing the best trade-off. In addition, other re-weighting based methods perform worse than FairDRO on Adult in terms of DCA, which implies that FairDRO successfully finds the better weights (*i.e.*, quasi-probability $\boldsymbol{q}^y$) over groups for achieving group fairness. Finally, we note that the balanced accuracy of other baselines and FairDRO are higher than Scratch on Adult due to the severe imbalance of class and group labels although Scratch does not use any fairness constraints.

	UTKFace		CivilComments	
	Acc. (↑)	Δ_{DCA} (↓)	Acc. (↑)	Δ_{DCA} (↓)
Scratch	80.40	22.58	74.82	10.60
RW (Kamiran & Calders, 2012)	80.27	20.17	77.17	7.78
LBC (Jiang & Nachum, 2020)	78.63	19.50	74.83	7.69
FairBatch (Roh et al., 2020)	79.63	21.00	75.93	11.30
Cov (Zafar et al., 2017a)	N/A		75.48	**6.40**
FairHSIC (Quadrianto et al., 2019)	81.60	18.58	N/A	
MFD (Jung et al., 2021b)	**81.83**	16.08	76.70	11.36
FairDRO	79.48	**14.25**	**77.65**	7.07

Table 2: **The best performances on UTKFace and CivilComments**. Details are the same as Tab. 1. "N/A" denotes the method is not trainable.

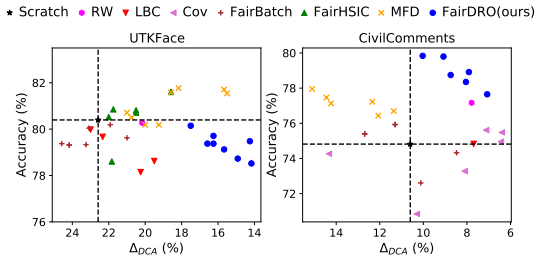

Figure 2: **The trade-offs between accuracy and DCA on UTKFace (left) and CivilComments (right)**. The "Scatch" performances are highlighted by the dotted lines

5.2 Vision Datasets

We also evaluate FairDRO on UTKFace (Zhang et al., 2017), a face dataset with multi-class and multi-group labels. UTKFace includes more than 20K images with age, gender and ethnicity labels. We divide the age labels into three classes ("0 to 19", "20 to 40", and "more than 40"), and set these as the class labels. We also use four ethnic classes (*"White", "Black", "Asian",* and *"Indian"*) as the group labels. Results on another vision dataset, CelebA (Liu et al., 2015), are given in Appendix D.2.

Tab. 2 and Fig. 2 show the best performances and the accuracy-fairness trade-offs on UTKFace for each method. We confirm that FairDRO significantly outperforms the baselines in terms of DCA. Although MFD also achieves competitive trade-offs compared to FairDRO, we note that MFD requires additional computations for training a teacher model for knowledge distillation. We also observe that the re-weighting based baselines are ineffective in achieving fairness on UTKFace, while FairDRO systematically optimizes the weights for reducing DCA. In addition, we report the difference of EO (DEO) in Appendix D.1 and show FairDRO again outperforms them in DEO.

5.3 Language Dataset

In order to show the versatility of FairDRO, we additionally consider a language dataset, CivilComments-WILDS (Koh et al., 2021). CivilComments is a large collection of comments on online article taken from the Civil Comments platform. It comprises 450K comments which are annotated with identity attributes by 10 crowdworkers and majority votes. The CivilComments task is to predict whether a given comment is toxic or not. We set ethnicity as the sensitive attribute and extract the subset of comments that contain identities w.r.t. ethnicity (*"Black", "White", "Asian", "Latino"* and *"Ohters"*). Note that the dataset has continuous values for the identities (between 0 and 1), hence, we binarized the continuous group values by setting the maximum as 1 and the others as 0, because the existing in-processing methods are only available to handle discrete group labels.

The results are shown in Tab. 2 and Fig. 2. Note that since a pre-trained BERT is fine-tuned only for small epochs, the update for $\boldsymbol{q}^y$ of FairDRO and weights of re-weighting baselines is executed per 100 iterations, instead of every epoch. FairDRO consistently achieves superior performances in terms of both metrics and a better trade-off on CivilComments. In Tab. 2, we marked results of FairHSIC as N/A because it fails to converge due to the gradient exploding. Despite of the best DCA of Cov, it does not achieve good performance on tabular datasets and cannot bet applied for multi-class classification tasks. Although MFD is the best baseline on UTKFace, it does not show effectiveness in achieving fairness on CivilComments.

5.4 Analysis

Ablation study. We analyze the effect of our DRO formulation and the classwise treatment employed by FairDRO. Tab. 3 shows the accuracy and Δ_{DCA} of each method on each dataset. RVP (w/ classwise) directly employs the variance of groupwise cross-entropy losses as a regularization term for each class separately, *i.e.*, the class-separated version of the RHS of (5) with cross-entropy loss. FairDRO (w/o classwise) defines only one uncertainty set over the (y, a) pair in (6) for the ablation study of the classwise treatment. In Tab. 3, the result of RVP (w/ classwise) on CivilComments is

Table 3: **Ablation study of FairDRO**. RVP (w/ classwise) and FairDRO (w/o classwise) are for the ablation study of our key two components, the DRO formulation, and classwise treatment, respectively. We follow the same model selection criterion used in Tab. 1.

	Adult		COMPAS		UTKFace		CivilComments	
	Acc. (↑)	Δ_{DCA} (↓)	Acc (↑)	Δ_{DCA} (↓)	Acc. (↑)	Δ_{DCA} (↓)	Acc. (↑)	Δ_{DCA} (↓)
Scratch	76.10	4.79	**63.48**	15.10	**80.40**	22.58	74.82	10.60
RVP (w/ classwise)	**80.10**	4.62	61.99	9.74	77.25	**12.33**	N/A	
FairDRO (w/o classwise)	78.45	**1.62**	60.72	6.28	78.25	20.75	77.42	8.85
FairDRO	79.23	1.99	61.97	**5.20**	79.48	14.25	**77.65**	**7.07**

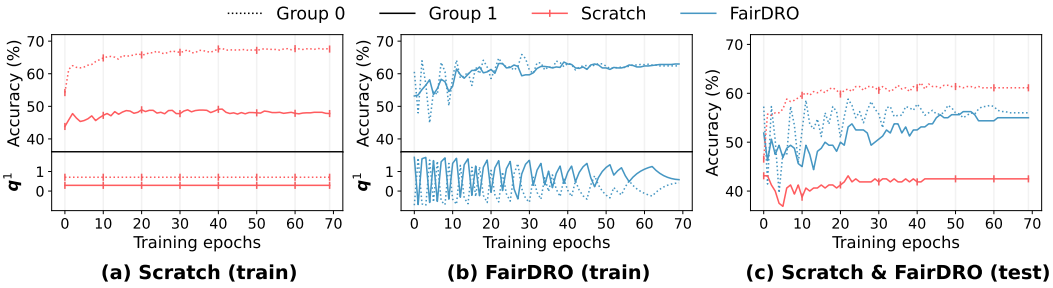

Figure 3: **Accuracies and loss weights for each group (q) on COMPAS**. The groupwise training accuracies and loss weights during training for Scratch (a) and FairDRO (b) are shown. Groupwise test accuracies are shown in (c). The reported accuracies and weights are only for label 1, while label 0 results are in Appendix D.4.

marked as N/A because of the failure of convergence by the gradient exploding. Followings are our observations. First, we see that although RVP (w/ classwise) improves DCA on COMPAS and UTKFace, it sometimes fails to produce a fair model because its variance regularization using the cross entropy loss is a crude surrogate for DCA. Second, we clearly observe that FairDRO (w/o classwise) achieves sub-optimal accuracy-DCA trade-offs (except for Adult), which implies that our classwise DRO formulation is critical for the optimal performances since it makes the connection to the DCA regularization more clearly.

Visualization of $\boldsymbol{q}^y$. We visualize the learning dynamics of the weights ($\boldsymbol{q}^y$), including training accuracies (Fig. 3 (a) and (b)) and test accuracies (Fig. 3 (c)) for Scratch and FairDRO on COMPAS. For simplicity, we report the result for $y = 1$ (for $y = 0$ is in Appendix D.4). In Fig. 3 (a), $\boldsymbol{q}^1$ simply stands for the ratio of samples for each group, confirming the data imbalance between the two groups. At the top of the plot, Scratch shows a large discrepancy in accuracy between the two groups throughout training. On the other hand, Fig. 3 (b) shows FairDRO adjusts $\boldsymbol{q}^1$ by assigning low values to the high accuracy group (Group 0) and high values to the low accuracy group (Group 1). Note that $\boldsymbol{q}^1$ initially fluctuates and even assigns *negative* values for the high accuracy group but eventually converges via learning rate scheduling in (10). As a result, we clearly observe both the training and test accuracy gaps between the groups smoothly become negligible as the training continues. We believe this example clearly shows the effectiveness of introducing the quasi-probability in $\mathcal{Q}_\rho$ and our FairDRO framework in achieving small test accuracy gaps between the groups, *i.e.*, Δ_{DCA}.

6 Concluding Remark

We proposed a theoretically well-justified in-processing method for achieving the group fairness. In order to make the non-differentiable DCA regularized problem more tractable, our FairDRO employs the classwise Group DRO framework with χ^2-divergence ball including quasi-probability as an uncertainty set. As a result, the FairDRO training objective directly incorporates DCA as a regularizer and can be effectively solved by our proposed optimization scheme. Furthermore, our DRO formulation enables FairDRO to have the strengths of both re-weighing based and regularization based methods. Indeed, we showed through our experiments that our FairDRO are applicable to various settings and consistently achieve superior accuracy-fairness trade-off on benchmark datasets.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Int. Conf. Mach. Learn.*, pp. 60–69. PMLR, 2018.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *Int. Conf. Learn. Represent.*, 2020.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. Robust optimization. In *Robust optimization*. Princeton university press, 2009.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conf. Fairness, Accountability and Transparency*, pp. 77–91. PMLR, 2018.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *Int. Conf. Learn. Represent.*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

Dheeru Dua, Casey Graff, et al. Uci machine learning repository. `http://archive.ics.uci.edu/ml`, 2017.

John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2, 2019.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, pp. 2803–2813. PMLR, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Conf. Innov. Theor. Compu. Scien.*, 2012.

Hong Fan, Wu Du, Abdelghani Dahou, Ahmed A Ewees, Dalia Yousri, Mohamed Abd Elaziz, Ammar H Elsheikh, Laith Abualigah, and Mohammed AA Al-qaness. Social media toxicity classification using deep learning: Real-world application uk brexit. *Electronics*, 10(11):1332, 2021.

Clare Garvie. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Int. Conf. Algo. Learn. Theory*, pp. 63–77. Springer, 2005.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Adv. Neural Inform. Process. Syst.*, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Int. Conf. Learn. Represent.*, pp. 1929–1938. PMLR, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 770–778, 2016.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *Int. Conf. Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pp. 862–872. PMLR, 2020.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *Int. Conf. Mach. Learn.*, pp. 4880–4889. PMLR, 2020.

Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. There's software used across the country to predict future criminals. and its biased against blacks. *ProPublica*, 2016. URL `https://github.com/propublica/compas-analysis`.

Sangwon Jung, Sanghyuk Chun, and Taesup Moon. Learning fair classifiers with partially annotated group labels. *arXiv preprint arXiv:2111.14581*, 2021a.

Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 12115–12124, 2021b.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Int. Conf. Mach. Learn.*, pp. 5637–5664. PMLR, 2021.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Int. Conf. Mach. Learn.*, pp. 5815–5826. PMLR, 2021.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Adv. Neural Inform. Process. Syst.*, 33:728–740, 2020.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Int. Conf. Comput. Vis.*, pp. 3730–3738, 2015.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Int. Conf. Learn. Represent.*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, 2019.

Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. Ensuring fairness beyond the training data. *Advances in neural information processing systems*, 33:18445–18456, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Laurent Son Nguyen and Daniel Gatica-Perez. Hirability in the wild: Analysis of online conversational video resumes. *IEEE Trans. Multimedia*, 18(7):1422–1437, 2016.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, volume 32, 2019.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8227–8236, 2019.

Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. Fairness for robust log loss classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5511–5518, 2020.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *Int. Conf. Learn. Represent.*, 2020.

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. *Int. Conf. Learn. Represent.*, 2020.

Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *Int. Conf. Comput. Vis.*, pp. 9495–9504, 2021.

Xinyue Shen, Steven Diamond, Yuantao Gu, and Stephen Boyd. Disciplined convex-concave programming. In *IEEE 55th Conference on Decision and Control*, pp. 1009–1014. IEEE, 2016.

Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Int. Conf. Comput. Vis.*, pp. 692–702, 2019.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. *Advances in Neural Information Processing Systems*, 33:5190–5203, 2020.

Yijie Wang, Viet Anh Nguyen, and Grani A Hanasusanto. Wasserstein robust classification with fairness constraints. *arXiv preprint arXiv:2103.06828*, 2021.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Chuanlong Xie, Fei Chen, Yue Liu, and Zhenguo Li. Risk variance penalization: From distributional robustness to causality. *ArXiv*, abs/2006.07544, 2020.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Int. Conf. Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017b.

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*, pp. 335–340, 2018.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5810–5818, 2017.

Chunting Zhou, Daniel Levy, Xian Li, Marjan Ghazvininejad, and Graham Neubig. Distributionally robust multilingual machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Punta Cana, Dominican Republic, November 2021.

APPENDIX

We provide additional materials in this document. We include our proofs of proposed Lemma and Propositions in Appendix A and implementation details such as optimization, implementations of baselines, and hyperparameter search in Appendix C. Finally, we report additional results in Appendix D

A PROOFS

Proof of Proposition 1 We define $\mathbf{L}^y \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ for brevity:

$$\mathrm{L}^y_{i,j} \triangleq \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_i) - \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_j), 1 \leq \forall i, j \leq |\mathcal{A}|.$$

If we denote $\boldsymbol{\Lambda}^y \in \mathbb{R}^{|\mathcal{A}|^2}$ as

$$\boldsymbol{\Lambda}^y \triangleq (\mathrm{L}^y_{1,1}, ..., \mathrm{L}^y_{|\mathcal{A}|,1}, \mathrm{L}^y_{1,2}, ..., \mathrm{L}^y_{|\mathcal{A}|,2}, ..., \mathrm{L}^y_{|\mathcal{A}|,|\mathcal{A}|}),$$

then, we have the following chain of equalities:

$$\begin{aligned}
\mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a)\}_{a\in\mathcal{A}}\right) =& \frac{1}{|\mathcal{A}|}\sum_a \left(\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a)\right)^2 - \left(\frac{1}{|\mathcal{A}|}\sum_a \left(\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a)\right)\right)^2 \\
=& \frac{1}{2|\mathcal{A}|^2}\sum_a\sum_{a\prime}\left(\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a) - \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_{a\prime})\right)^2 \\
=& \frac{1}{2|\mathcal{A}|^2}\|\boldsymbol{\Lambda}^y\|_2^2.
\end{aligned} \tag{A.1}$$

Note that as we consider ℓ to be zero-one loss function, we have the followings:

$$\begin{aligned}
\widehat{\Delta}_y \triangleq& \max_{a,a\prime}\left|\frac{1}{|\mathcal{D}^y_a|}\sum_{i=1}^{|\mathcal{D}^y_a|}\mathbb{1}\{\boldsymbol{\theta}(\boldsymbol{x}_i) \neq y\} - \frac{1}{|\mathcal{D}^y_{a\prime}|}\sum_{i=1}^{|\mathcal{D}^y_{a\prime}|}\mathbb{1}\{\boldsymbol{\theta}(\boldsymbol{x}_i) \neq y\}\right| \\
=& \max_{a,a\prime}|\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a) - \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_{a\prime})| \\
=& \|\boldsymbol{\Lambda}^y\|_\infty.
\end{aligned}$$

Since $\|\boldsymbol{\Lambda}^y\|_\infty = \max_i |\Lambda^y_i| \leq \sqrt{\sum_i (\Lambda^y_i)^2} = \|\boldsymbol{\Lambda}^y\|_2$, we obtain the second inequality of (2). Note also that $\|\boldsymbol{\Lambda}^y\|_2 = \sqrt{\sum_i (\boldsymbol{\Lambda}^y_i)^2} \leq \sqrt{\sum_i \|\boldsymbol{\Lambda}^y\|^2_\infty} = \sqrt{|\mathcal{A}|^2\|\boldsymbol{\Lambda}^y\|^2_\infty} = |\mathcal{A}|\|\boldsymbol{\Lambda}^y\|_\infty$. We thereby obtain the first inequality of (2).

From (2), we obtain $\widehat{\Delta}_y = 0$ if $\mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a)\}_{a\in\mathcal{A}}\right) = 0$ for all $y \in \mathcal{Y}$. Considering variance is always non-negative, we obtain the necessary condition that $\widehat{\Delta}_y = 0$ implies $\mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a)\}_{a\in\mathcal{A}}\right) = 0$ for all $y \in \mathcal{Y}$. ■

Proof of Lemma 1 Suppose that a model parameter $\boldsymbol{\theta}$ is given. As $\sum_{a\in\mathcal{A}}\left(q^y_a - \frac{1}{|\mathcal{A}|}\right) = 0$ and $\sum_{a\in\mathcal{A}}\frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a)$ is a constant, we can decompose $\sum_{a\in\mathcal{A}} q^y_a\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a)$ into

$$\begin{aligned}
\sum_{a\in\mathcal{A}} q^y_a\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a) =& \sum_{a\in\mathcal{A}}\frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a) + \sum_{a\in\mathcal{A}}(q^y_a - \frac{1}{|\mathcal{A}|})\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a) \\
=& \sum_{a\in\mathcal{A}}\frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a) + \sum_{a\in\mathcal{A}}(q^y_a - \frac{1}{|\mathcal{A}|})\Big(\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a) - \sum_{a\in\mathcal{A}}\frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^y_a)\Big).
\end{aligned}$$

Let $\boldsymbol{v}^y \triangleq \boldsymbol{q}^y - \frac{1}{|\mathcal{A}|}\mathbb{1}$, where $\mathbb{1} \in \mathbb{R}^{|\mathcal{A}|}$. From $\sum_{a\in\mathcal{A}} q^y_a = 1$, we have $\sum_{a\in\mathcal{A}} v^y_a = 0$.
Considering $D_\phi\Big(\boldsymbol{q}\,||\,\frac{1}{|\mathcal{A}|}\mathbb{1}\Big) = \sum_{a\in\mathcal{A}}\frac{1}{|\mathcal{A}|}(|\mathcal{A}|q^y_a - 1)^2 = \sum_{a\in\mathcal{A}}|\mathcal{A}|(q^y_a - \frac{1}{|\mathcal{A}|})^2 = |\mathcal{A}|\sum_{a\in\mathcal{A}}(v^y_a)^2$,

the following problems are equivalent:

$$\max_{\boldsymbol{q}^y \in \mathcal{Q}_\rho} \sum_{a\in\mathcal{A}} q_a^y \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) = \max_{\boldsymbol{v}^y} \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) + \sum_{a\in\mathcal{A}} v_a^y \Big(\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\Big)$$
$$\text{subject to } \sum_{a\in\mathcal{A}} v_a^y = 0, \|\boldsymbol{v}^y\|_2^2 \leq \frac{\rho}{|\mathcal{A}|}.$$

Also, we have the following inequalities:

$$\sum_{a\in\mathcal{A}} v_a^y \Big(\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\Big) \leq \sqrt{\sum_{a\in\mathcal{A}} (v_a^y)^2} \sqrt{\sum_{a\in\mathcal{A}} \Big(\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\Big)^2} \tag{A.2}$$

$$\leq \sqrt{\frac{\rho}{|\mathcal{A}|}} \sqrt{|\mathcal{A}| \times \mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\}_{a\in[\mathcal{A}]}\right)}, \tag{A.3}$$

in which (A.2) holds by the Cauchy-Schwartz inequality. The second inequality follows from

$$\mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\}_{a\in\mathcal{A}}\right) = \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \Big(\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\Big)^2,$$
$$\|\boldsymbol{v}^y\|_2^2 \leq \frac{\rho}{|\mathcal{A}|}.$$

The equality in (A.3) is attained if and only if the vector $\boldsymbol{v}^y$ satisfies: (i) $\boldsymbol{v}^y$ and the $|\mathcal{A}|$-dimensional vector whose a-th element is $\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)$ are in the same direction; (ii) $||\boldsymbol{v}^y||_2^2 = \frac{\rho}{|\mathcal{A}|}$. This implies that the a-th element of $\boldsymbol{v}^y$ should be

$$v_a^y = \sqrt{\frac{\rho}{|\mathcal{A}|}} \times \frac{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)}{\sqrt{\sum_{a\in\mathcal{A}} \Big(\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\Big)^2}}$$
$$= \sqrt{\frac{\rho}{|\mathcal{A}|}} \times \frac{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)}{\sqrt{|\mathcal{A}| \times \mathrm{Var}\left(\{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\}_{a\in\mathcal{A}}\right)}}.$$

Thus, Lemma 1 holds if and only if

$$q_a^y = \frac{1}{|\mathcal{A}|} + \sqrt{\frac{\rho}{|\mathcal{A}|}} \times \frac{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)}{\|\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\|_2}, \tag{A.4}$$

for all $a \in \mathcal{A}$. ■

Proof of Proposition 2 From the proof of Lemma 1, the equality in (5) holds if and only if

$$q_a^y = \frac{1}{|\mathcal{A}|} + \sqrt{\frac{\rho}{|\mathcal{A}|}} \times \frac{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)}{\|\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\|_2}, \tag{A.5}$$

for $\forall a \in \mathcal{A}$. We now consider each element of $\boldsymbol{q}^{*y}$. For given $y \in \mathcal{Y}$, we define $\boldsymbol{\gamma}^y \in \mathbb{R}^{|\mathcal{A}|}$ as follows:

$$\gamma_a^y \triangleq \frac{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_a \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)}{\|\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y) - \sum_a \frac{1}{|\mathcal{A}|}\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_a^y)\|_2},$$

for any $a \in \mathcal{A}$. Fix $i \in \mathcal{A}$ by symmetry. Then, we have:

$$\sum_{a\in\mathcal{A}} \gamma_a^y = 0, \text{and thus, } \gamma_i^y = -\sum_{a\in\mathcal{A}\setminus i} \gamma_a^y, \tag{A.6}$$

$$\sum_{a\in\mathcal{A}} (\gamma_a^y)^2 = 1, \text{and thus, } (\gamma_i^y)^2 = 1 - \sum_{a\in\mathcal{A}\setminus i} (\gamma_a^y)^2. \tag{A.7}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}|\gamma_i^y| &= |(1,...,1)\cdot(\gamma_1^y,...,\gamma_{i-1}^y,\gamma_{i+1}^y,...,\gamma_{|\mathcal{A}|}^y)| \\ &\leq \sqrt{|\mathcal{A}|-1}\cdot\sqrt{(\gamma_1^y)^2+\cdots+(\gamma_{i-1}^y)^2+(\gamma_{i+1}^y)^2+\cdots+(\gamma_{|\mathcal{A}|}^y)^2} \\ &= \sqrt{|\mathcal{A}|-1}\cdot\sqrt{1-(\gamma_i^y)^2},\end{aligned}$$

in which the first and last equality follow from (A.6) and (A.7), respectively.

From the above inequality, we have the followings:

$$\begin{aligned}(\gamma_i^y)^2 &\leq (|\mathcal{A}|-1)\cdot(1-(\gamma_i^y)^2), \\ |\mathcal{A}|(\gamma_i^y)^2 &\leq |\mathcal{A}|-1, \\ |(\gamma_i^y)| &\leq \sqrt{\frac{|\mathcal{A}|-1}{|\mathcal{A}|}}.\end{aligned} \tag{A.8}$$

Considering (A.6), the equality in (A.8) is attained if and only if

$$\begin{aligned}\gamma_i^y &= \pm\sqrt{\frac{|\mathcal{A}|-1}{|\mathcal{A}|}}, \\ \gamma_a^y &= \mp\sqrt{\frac{1}{|\mathcal{A}|(|\mathcal{A}|-1)}}, \forall a\in\mathcal{A}\backslash i.\end{aligned}$$

Thus, we can derive the range described in Proposition 2. ■

B ADDITIONAL RELATED WORKS

Much more works related to algorithmic fairness have recently utilized the DRO frameworks for their own goals. A line of works embeds the DRO frameworks into the fairness-constraint optimization problem to make a model fair, and robust to distribution shifts of a test dataset (Jiang et al., 2020; Taskesen et al., 2020; Wang et al., 2021; Mandal et al., 2020). For example, Mandal et al. (2020) aimed to train fair classifiers to be with respect to weighted perturbations in the training distribution by considering all possible weight set over training samples. Wang et al. (2021) and Taskesen et al. (2020) proposed DRO objectives with Wasserstein uncertainty sets which are combined by group fairness constraints. Although all aforementioned methods showed improvements in group fairness and robustness, the key differences from FairDRO is that we leverage the DRO framework for directly promoting group fairness, not robustness, and consider the uncertainty set over groups, not samples.

Some works proposed DRO formulations for achieving fairness in more challenging settings like learning with imperfect group labels or sequential learning. Wang citewang2020robust proposed how to achieve the group fairness successfully given noisy group labels only, by considering the worst-case distribution of group labels. Hashimoto *et al.* Hashimoto et al. (2018) devised repeated loss minimization using the DRO-based training objective with the χ^2-divergence uncertainty set for encouraging long-term fairness in a sequential setting where the group populations changes over time. However, they do not also use DRO as a direct tool for achieving group fairness.

Several works adopted min-max optimization techniques for group fairness (although they are not based on DRO frameworks). Zhang et al. (2018) proposed an adversarial debiasing technique that eliminates group information of latent features of a neural network (NN) model through the mini-max game of an adversary and the NN model. Baharlouei et al. (2020), and Jiang *et al.* Jiang et al. (2020) proposed novel regularization terms for group fairness using Rényi correlation and Wasserstein distance respectively, which lead to min-max formulations. We emphasize that these methods rely on additional terms or adversaries for promoting group fairness, while our method directly uses the DRO-based min-max formulation as a fairness regularization term.

There are some recent works perhaps closest to our method in terms of using DRO for achieving the fairness. Rezaei et al. (2020) incorporated a group fairness constraint into statistic-matching based robust log-loss minimization problem for training fair logistic regression models. Although

they showed nice theoretical benefits in terms of its convexity and convergence, their method has a limitation that group information is needed at the test time. Yurochkin et al. (2019) aims to achieve the fairness of a classifier using DRO frameworks. However, they only focus on a novel *individual* fairness notion called as distributionally robust fairness (DRF), not group fairness.

C MORE IMPLEMENTATION DETAILS

We used PyTorch (Paszke et al., 2019); Experiments are performed on a server with AMD Ryzen Threadripper PRO 3975WX CPUs and NVIDIA RTX A5000 GPUs. All models are evaluated on separate test sets, and all experiments are repeated with 4 different random seeds. All the reported results are the averaged results.

C.1 OPTIMIZATION

For tabular and vision datasets, we train all models with the AdamW optimizer (Loshchilov & Hutter, 2019) for 70 epochs. We set the mini-batch size and the weight decay as 128 and 0.001, respectively. The initial learning rate is set as 0.001 and decayed by cosine annealing technique (Loshchilov & Hutter, 2017).

For the language dataset, we fine-tune pre-trained BERT with the AdamW optimizer for 3 epochs. We set the mini-batch size and the weight decay as 24 and 0.001, respectively. The initial learning rate is set as 0.00002 and adjusted with a learning rate schedule using a warm-up phase followed by a linear decay. All results are reported for the model at the last epoch.

C.2 IMPLEMENTATION DETAILS OF IN-PROCESSING BASELINES

Here, we describe the implementation details of each in-processing baseline used for the experiments.

The original LBC (Jiang & Nachum, 2020) requires multiple full-training iterations by alternatively re-weighting each group based on the given fairness metric and re-training the full dataset. Since this optimization procedure needs a very high computation budget, we modify full-training iterations to 5 epochs and 14 iterations for vision and language datasets.

FairBatch (Roh et al., 2020) and RW (Kamiran & Calders, 2012) are implemented in the way they were originally proposed.

Cov (Zafar et al., 2017b) utilized a fairness constraint based on Covariance between the group label and the signed distance of the feature vectors from the decision boundary of a classifier. Although they used the Disciplined convex-concave program (DCCP) (Shen et al., 2016) solver for the constrained problem, we apply our optimization procedure (*i.e.*, AdamW) by setting the covariance-based constraint as a regularization term.

MFD (Jung et al., 2021b) and FairHSIC (Quadrianto et al., 2019) use additional fairness-promoting regularization terms based on MMD and HSIC. For FairHSIC, we only implement the second term of their decomposition loss (*i.e.*, the HSIC loss between the feature representations and the group labels). For implementing their regularization terms, we use the Gaussian RBF kernel of which the variance parameter is set as the mean of squared distance between all data points, following Jung et al. (2021a).

Baharlouei et al. (2020) attempted to remove the non-linear dependency between the model prediction and the sensitive attribute by using Rényi correlation as a regularization term. The resulting training objective is a mini-max optimization problem where the maximization is for computing the Rényi correlation, and the minimization is for learning a model. When the group label is binary, we use modified Algorithm 2 in Baharlouei et al. (2020) where the minimization step (*i.e.*, line 4 in Algorithm 2) is processed in a mini-batch manner. Although Algorithm 1 is designed for the non-binary group label, applying the algorithm to our experiments with the non-binary group label is not straightforward because the second term in the RHS of equation 6 is a biased estimator so the mini-batch optimization can be significantly inaccurate.

Table C.1: **Hyperparameter search spaces.** We perform the grid search to find the best hyperparameters for each method.

Method	Hyperparameter	Search range
Cov (Zafar et al., 2017b)	Covariance strength λ	$[10^{-2}, 10^{2}]$
Rényi (Baharlouei et al., 2020)	Rényi correlation strength λ	$[10^{-2}, 10^{3}]$
MFD (Jung et al., 2021b)	MMD strength λ	$[10^{-2}, 10^{4}]$
FairHSIC (Quadrianto et al., 2019)	HSIC strength λ	$[10^{-2}, 10^{4}]$
LBC (Jiang & Nachum, 2020)	LR for re-weights η	$[10^{-3}, 10^{2}]$
FairBatch (Roh et al., 2020)	LR for mini-batch size of each group α	$[10^{-3}, 10^{2}]$
FairDRO	radius of χ^2-divergence ball ρ	$[10^{-2}, 10^{2}]$

Table D.1: **The best performances on Adult and COMPAS datasets**. The number in the parentheses with $\pm$ stands for the standard deviation of each metric obtained from several independent runs with 4 different seeds. Other details are the same as Tab. 1.

	Adult		COMPAS	
	Acc. (↑)	Δ_{DCA} (↓)	Acc. (↑)	Δ_{DCA} (↓)
Scratch	76.10 ($\pm$0.00)	4.79 ($\pm$0.03)	**63.48** ($\pm$0.00)	15.10 ($\pm$0.00)
RW (Kamiran & Calders, 2012)	77.04 ($\pm$0.01)	8.42 ($\pm$0.02)	63.40 ($\pm$0.16)	13.45 ($\pm$0.79)
LBC (Jiang & Nachum, 2020)	76.22 ($\pm$0.02)	4.13 ($\pm$0.33)	61.12 ($\pm$0.17)	6.02 ($\pm$1.08)
FairBatch (Roh et al., 2020)	76.29 ($\pm$0.17)	4.24 ($\pm$0.08)	61.18 ($\pm$0.23)	5.95 ($\pm$1.13)
Cov (Zafar et al., 2017a)	76.09 ($\pm$0.00)	4.74 ($\pm$0.02)	60.68 ($\pm$1.07)	8.93 ($\pm$1.51)
Rényi (Baharlouei et al., 2020)	72.79 ($\pm$1.53)	2.67 ($\pm$0.50)	60.35 ($\pm$0.23)	5.50 ($\pm$1.12)
FairDRO	**79.23** ($\pm$0.11)	**1.99** ($\pm$0.38)	61.97 ($\pm$0.24)	**5.20** ($\pm$0.92)

C.3 Hyperparameters for main results

We perform the grid search on the hyperparameter candidates for every method. The full hyperparameter search space is illustrated in Tab. C.1.

D Additional results

D.1 Result tables

We additionally report the standard deviation of each result in Tab. 1 and Tab. 2 in Tab. D.1 and Tab. D.2. In Tab. D.2, we also provide *difference of Equalized Odds* (DEO) on UTKFace, denoted by Δ_{DEO}, which is defined as follows:

$$\Delta_{\text{DEO}} \triangleq \frac{1}{|\mathcal{Y}|^2} \sum_{y,y' \in \mathcal{Y}} \max_{a,a'} |P(\widehat{Y} = y' | A = a, Y = y) - P(\widehat{Y} = y' | A = a', Y = y)|. \tag{D.1}$$

Note that Δ_{DCA} and Δ_{DEO} are the same for binary labels.

From Tab. D.1 and Tab. D.2, we observe FairDRO achieve the best Δ_{DCA} with moderate variances. Moreover, although LBC, FairBatch, and FairHSIC were devised for achieving EO, we again show that FairDRO outperforms baselines on UTKFace with respect to DEO.

D.2 CelebA results

To show consistent improvements, we conducted an additional experiment on the CelebA dataset, which includes about 200K celebrities' face images with 40 annotated binary attributes. Among the attributes, we select "blond hair" as the class label and "gender" as the group label, which is the set of attributes widely used in many previous works Sagawa et al. (2020); Jung et al. (2021a); Chuang & Mroueh (2021); Liu et al. (2021).

Table D.2: **The best performances on UTKFace and CivilComments**. The number in the parentheses with $\pm$ stands for the standard deviation of each metric obtained from several independent runs with 4 different seeds. "N/A" denotes the method is not trainable. All other details are the same as Tab. 2.

	UTKFace			CivilComments	
	Acc. (↑)	Δ_{DCA} (↓)	Δ_{DEO} (↓)	Acc. (↑)	Δ_{DCA} (↓)
Scratch	80.40 ($\pm$0.32)	22.58 ($\pm$0.28)	15.56 ($\pm$0.14)	74.82 ($\pm$1.42)	10.60 ($\pm$4.12)
RW (Kamiran & Calders, 2012)	80.27 ($\pm$0.40)	20.17 ($\pm$2.36)	18.00 ($\pm$0.00)	77.17 ($\pm$0.92)	7.78 ($\pm$1.87)
LBC (Jiang & Nachum, 2020)	78.63 ($\pm$0.49)	19.50 ($\pm$3.41)	13.61 ($\pm$2.43)	74.83 ($\pm$2.39)	7.69 ($\pm$1.51)
FairBatch (Roh et al., 2020)	79.63 ($\pm$0.64)	21.00 ($\pm$1.62)	14.44 ($\pm$1.05)	75.93 ($\pm$0.86)	11.30 ($\pm$1.96)
Cov (Zafar et al., 2017a)	N/A			75.48 ($\pm$1.85)	**6.40** ($\pm$1.36)
FairHSIC (Quadrianto et al., 2019)	81.60 ($\pm$0.33)	18.58 ($\pm$1.30)	12.94 ($\pm$0.78)	N/A	
MFD (Jung et al., 2021b)	**81.83** ($\pm$0.95)	16.08 ($\pm$0.76)	11.08 ($\pm$0.55)	76.70 ($\pm$1.09)	11.36 ($\pm$2.78)
FairDRO	79.48 ($\pm$0.59)	**14.25** ($\pm$1.26)	**11.06** ($\pm$1.02)	**77.65** ($\pm$1.75)	**7.07** ($\pm$1.44)

	CelebA	
	Acc. (↑)	Δ_{DCA} (↓)
Scratch	78.75	20.56
RW (Kamiran & Calders, 2012)	77.60	15.90
LBC (Jiang & Nachum, 2020)	80.69	13.61
FairBatch (Roh et al., 2020)	79.20	14.24
Cov (Zafar et al., 2017a)	85.97	13.06
FairHSIC (Quadrianto et al., 2019)	86.63	9.10
MFD (Jung et al., 2021b)	**90.97**	4.44
FairDRO	89.55	**3.68**

Table D.3: **The best performances on CelebA**. Details are the same as Tab. 1

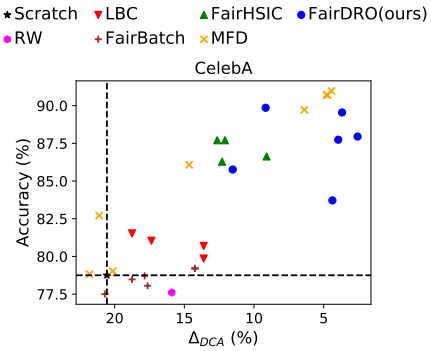

Figure D.1: **Trade-offs between accuracy and DCA on CelebA**. The "Scratch" performances are highlighted by the dotted lines

Tab. D.3 and Fig. D.1 show a similar result trend to one of UTKFace in Sec. 5.2. FairDRO again outperforms the baselines in terms of DCA and achieves a competitive trade-off with MFD. We also observe re-weighting based baselines show worse DCA than FairDRO, meaning that FairDRO finds the better weights over groups for lower DCA.

D.3 Ablation Study for the Choice of the Uncertainty Set

We analyze the effect of each constraint employed by FairDRO. Tab. D.4 shows the overall accuracy, Δ_{DCA} and the worst accuracy over groups for each ablation case. "Classwise", "χ^2-divergence" and "Quasi-prob" ((4)) indicate each component added when defining the uncertainty set of Group DRO (*i.e.*, simplex over groups) as described in Sec. 4. For a fair comparison, we apply the same optimization scheme (described in Sec. 4.4) for all DRO variants.

As the supremum in 3 is attained at the vertex of the simplex $\Delta^{|\mathcal{Y}|\times|\mathcal{A}|}$, Group DRO minimizes the worst-case loss over the (y, a) pair, thereby it can prevent the lowest accuracy across the group and class from becoming too low. Thus, Group DRO substantially improves and achieves the best worst-case accuracy over groups compared to ERM (*i.e.*, 1st $\rightarrow$ 2nd row). However, it shows the worst DCA among the DRO variants, confirming the mismatch between the criterion Group DRO uses and the group fairness metric. We observe that changing the simplex uncertainty set of Group DRO to the proposed χ^2-divergence and quasi-prob uncertainty sets (4) (*i.e.*, 2nd $\rightarrow$ 3rd row) has an impact on the group fairness mostly. In addition, reducing the size of uncertainty sets via classwise treatment (*i.e.*, 3rd $\rightarrow$ 6th row) further improves the group fairness. Finally, including negative values in the uncertainty set also helps to improve the group fairness (*i.e.*, 5th $\rightarrow$ 6th row) on most datasets, by more severely penalizing the major groups with high accuracy with negative weights.

Table D.4: **Ablation study of FairDRO**. We show the effects of each component of the proposed FairDRO: GDRO and the constraints on the uncertainty set, including classwise uncertainty set, χ^2-divergence ball, and quasi-probability. All numbers are average results over Adult, COMPAS, UTKFace, and CivilComments datasets. GDRO is an abbreviation for Group DRO.

GDRO	Classwise	χ^2-div.	Q-prob.	Adult			COMPAS			UTKFace			CivilComments		
				Acc.	Δ_{DCA}	Worst acc.	Acc.	Δ_{DCA}	Worst acc.	Acc.	Δ_{DCA}	Worst acc.	Acc.	Δ_{DCA}	Worst acc.
✗	✗	✗	✗	76.10	4.79	57.52	**63.48**	15.10	41.88	**80.40**	22.58	51.00	74.82	10.60	51.25
✓	✗	✗	✗	**81.89**	6.18	**78.62**	62.06	6.68	**57.57**	79.54	18.17	61.00	**80.32**	12.63	69.72
✓	✗	✓	✓	78.45	**1.62**	77.46	60.72	6.28	55.94	78.25	20.75	45.75	77.42	8.85	63.70
✓	✓	✗	✗	81.68	5.89	77.67	62.10	6.41	53.59	78.69	14.50	63.75	79.99	9.16	**71.52**
✓	✓	✓	✗	81.53	5.76	77.38	62.32	6.42	53.91	78.37	**13.75**	**66.50**	78.56	7.89	67.24
✓	✓	✓	✓	79.23	1.99	75.60	61.97	**5.20**	53.91	79.48	14.25	65.75	77.65	**7.07**	62.85

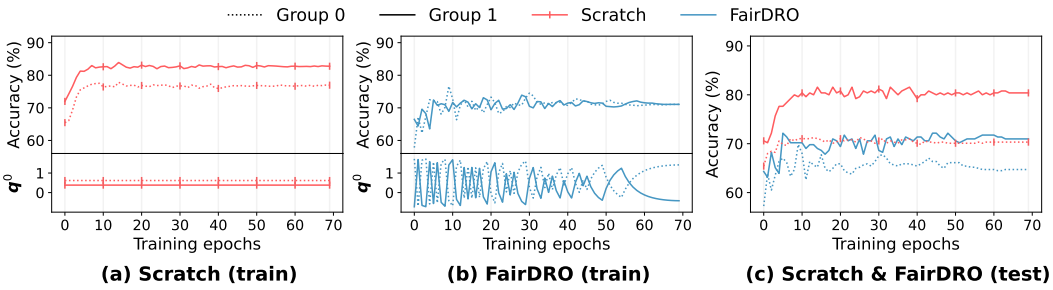

Figure D.2: **Accuracies and loss weights for each group (q) on COMPAS** ($y = 0$). Details are the same as Fig. 3.

D.4 Visualization of q^0

We continue from Sec. 5.4 and provide visualization results including the weights (q^y), training accuracy, and test accuracy for $y = 0$. In Fig. D.2 (a), Scratch shows that the significant accuracy gap between two groups is kept during all training time. On the other hand, we observe again in Fig. D.2 (b) that q^0 initially fluctuates and eventually converges by our optimization procedure, and as a result, the large discrepancy of training accuracy is reduced at the end of training. Finally, we show from Fig. D.2 (c) that the test accuracy gap of FairDRO is than that of Scratch, *i.e.*, FairDRO achieves lower DCA.

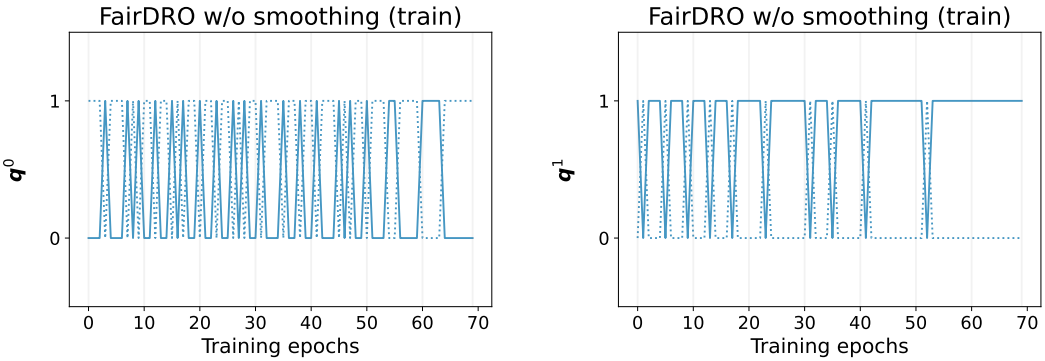

Figure D.3: **q on COMPAS**. The groupwise loss weights during training for FairDRO without smoothed IBR are shown. The reported weights are for label 0 (left) and 1 (right).

Table D.5: **The best performances on Adult and COMPAS datasets**. We follow the same model selection criterion used in [Table 1, Manuscript].

	Adult		COMPAS	
	Acc. (↑)	Δ_{DCA} (↓)	Acc. (↑)	Δ_{DCA} (↓)
Scratch	76.10	4.79	**63.48**	15.10
FairDRO (w/o smoothing)	**81.55**	5.76	61.86	5.50
FairDRO	79.23	**1.99**	61.97	**5.20**

Table D.6: **The number of samples in Adult, COMPAS, and CelebA datasets.**

	Adult		COMPAS		CelebA	
	$a=0$	$a=1$	$a=0$	$a=1$	$a=0$	$a=1$
$y=0$	13026	20988	2080	1278	17809	67054
$y=1$	1669	9539	1987	822	23060	1567

Table D.7: **The number of samples in UTKFace and CivilComments datasets.**

	UTKFace				CivilComments				
	$a=0$	$a=1$	$a=2$	$a=3$	$a=0$	$a=1$	$a=2$	$a=3$	$a=4$
$y=0$	2049	336	1025	638	11485	17632	8154	4048	9823
$y=1$	3735	3116	1856	2221	4564	5892	828	545	1286
$y=2$	4294	1074	553	1116			-		

D.5 Ablation study of smoothed IBR

We experimentally observe the effectiveness of smoothed IBR updates introduced in Sec. 4.4 on FairDRO. Although the IBR procedure is known to converge (Jin et al., 2020), in our experiment, we observe that $\boldsymbol{q}$ values oscillate when IBR without smoothing was utilized in FairDRO. Fig. D.3 visualizes $\boldsymbol{q}$ of FairDRO without smoothed IBR. The figure shows the large oscillation of FairDRO without smoothing compared to FairDRO. Tab. D.5 is the result of FairDRO without smoothed IBR in the same setting as Tab. 1. For the reported result, the same model selection criterion introduced in Sec. 5 is applied. Tab. D.5 shows that the use of smoothed learning rate schedule for FairDRO achieves improvements in terms of DCA. As we observe that the oscillation leads to performance degrades in terms of DCA, we identify the effectiveness of smoothed IBR on FairDRO.

D.6 Issue of standard accuracy

In the experiments, we reported the balanced accuracy for measuring the performance of the models instead of the standard accuracy. While the balanced accuracy offers several advantages in measuring performance on imbalanced datasets, but also has drawbacks. In general, it may underestimate the performance of certain groups with relatively large numbers. As in most cases, the groups with large numbers often become majority groups. The balanced accuracy metric may not capture accuracy reduction in majority group. In real-world applications, there is also concern about a decrease in majority-group accuracy.

However, we argue that the vanilla standard accuracy would be problematic when the dataset is severely class-imbalanced. For example, Tab. D.6 shows the number of samples in each dataset with respect to the class and group labels. For the case of Adult dataset, 75.2% of the samples are labeled as $y=0$, hence, a naive predictor that always predict $\hat{Y}=0$ will achieve the standard accuracy of 75.2% and DCA$=0$, which are clearly overestimating the (accuracy, fairness) performance of the classifier. Therefore, to avoid such issue, we use the group-class balanced accuracy, rather than the standard accuracy.