Alpha and Prejudice: Improving α-sized Worst-case Fairness via Intrinsic Reweighting

Jing Li, Yinghua Yao, Yuangang Pan, Xuanqian Wang, Ivor W. Tsang*, *Fellow, IEEE*, and Xiuju Fu

***Abstract*—Achieving worst-case group fairness typically relies on maximizing the utility of the worst-off demographic group. However, in practice, demographic information is often unavailable, making direct max-min formulations infeasible. To address this, recent work introduces a relaxed setting, using a lower bound α on the minimal group size—referred to as "α-sized worst-case fairness" in this paper. We first motivate the importance of this setting by highlighting its relevance to data privacy, a critical yet underexplored perspective. Rather than simply retraining on worst-off samples, we propose a reweighting approach that assigns sample weights based on their intrinsic contributions to fairness. To handle the global nature of worst-case objectives efficiently, we develop a stochastic learning algorithm that simplifies training without sacrificing performance. We also address the impact of outliers by introducing a robust variant of our method. Through theoretical analysis and extensive experiments on standard fairness benchmarks, we show that our methods not only connect naturally to existing fairness-through-reweighting approaches but also outperform strong baselines.**

***Index Terms*—Fairness learning, worst-case formulation, reweighted samples, stochastic optimization, outliers.**

I. Introduction

NUMEROUS sectors within society have been identified as revealing instances of prejudice across demographic groups, such as credit offerings [1], recidivism prediction [2], allocation of health-care resources [3], and more. As AI systems become extensively integrated into contemporary decision-making processes [4], concerns regarding discrimination hold significance in the realm of machine learning model development [5]. With prepared demographic information, group fairness [6]–[9], a.k.a. statistical fairness [10], has been established, which typically restricts the trained model to behave consistently across different groups. Although stated in an in-processing manner, group fairness achieved through pre-processing [11]–[16] and post-processing [17]–[19] can be also interpreted in a similar fashion; the former requires the learned representation of each group to be statistically aligned with each other and the latter treats the model after standard training as a black box and focuses on how to barely adjust model outputs to remove group variations.

Jing Li, Yinghua Yao, Yuangang Pan, and Ivor W. Tsang are with the Center for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore and also with the Institute of High-Performance Computing, Agency for Science, Technology and Research, 1 Fusionopolis Way, 138632, Singapore. E-mail: {kyle.jingli, eva.yh.yao, yuangang.pan, ivor.tsang}.gmail.com

Xuanqian Wang is with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China. Email: wwxxqq@buaa.edu.cn

Xiuju Fu is with the Institute of High-Performance Computing, Agency for Science, Technology and Research, 1 Fusionopolis Way, 138632, Singapore. Email: fuxj@ihpc.a-star.edu.sg

*Corresponding author.

In the case of a multi-valued sensitive attribute, such as race or religion, a naive approach to achieving fairness can be characterized by minimizing the summed discrimination across all pairs of groups. Given a predefined model utility, John Rawls' theory [20] promotes maximizing the utility (e.g., classification accuracy) of the worst-off group, thereby ensuring the consent of minorities. Because if a model's capacity is limited, encouraging the utility of the worst-off group at each iteration will eventually reduce group disparity. Following the convention of [21], [22], we use *worst-case fairness* to refer to this type of fairness formulations. Note that worst-case fairness is distinct from notions like disparate treatment or disparate impact [23], which focuses on decisions. Instead, it aligns more closely with utility-based fairness research, such as fair PCA [24], [25] or fair regression [26], where the group disparity on reconstruction errors or regression errors is considered.

Recent developments in worst-case fairness focus on the scenarios where demographic information is not available during the training process due to laws or privacy in practice [27], [28]. This restriction introduces a new challenge for training fair models, as we are no longer have access to group information, which was previously accessible through data annotations. Some works propose using proxy attributes, either by selecting observed ones [29], [30] or predicting from data [31], [32]. We follow the assertion of [21] that an interested attribute could arbitrarily partition the training data, implying that proxy attributes may not be sufficient in this case. To still enable a worst-case formulation without demographics, they propose that group size should be constrained by a parameter α, as has been argued in distributional robust optimization [33]. Although starting from the necessity of modeling, they pave the way for learning worst-case fairness with group prior used during model training only, followed by numerous subsequent works [34]–[36].

We believe it will be interesting to explain why it is possible to obtain prior information about group size even when demographics are too sensitive to acquire beforehand. In this paper, We adopt the perspective of data privacy [37], opening up opportunities for multidisciplinary research collaborations. On the other hand, by reviewing existing methods we realize that the solutions of achieving worst-case fairness can be further improved, because the state-of-the-art methods transform the population-based formulation into greedily upweighting the higher-loss samples instance-wise[1], which however overlooks

[1]This is also understood as an over-fitting issue [21], which they mitigate by adopting cross-validation.

the true attribution of each example and thus fails to obtain satisfactory model performance. Our assertion stems from influence function [38], which reveals that sample importance can be characterized from individual gradients rather than their loss values only. Interestingly, we realized that the individual gradients used to compute sample importance provide a valuable representation that could be further applied to address the outliers problem [35], [39], which has been identified as one of the key risks in worst-case formulations.

We summarize the contributions of this work as follows.

- We revisit the *α-sized worst-case fairness* problem where demographic labels are unavailable, but side information—a lower bound on group proportions—can be obtained or estimated. We discuss how this information can guide fairness interventions and highlight its practical usefulness, including situations where it is derived through privacy-aware techniques.
- We propose the Intrinsic ReWeighting (IRW) method to address α-sized worst-case fairness, reweighting each training sample by computing its individual contribution to improving the utility of the worst-off group. IRW is a gradient-based approach that provides a more reliable reweighting compared to existing methods that which heuristically assign larger weights to samples with higher losses.
- We enhance the efficiency of IRW by incorporating stochastic updates which overcome the limitations imposed by the global nature of worst-case formulations. Additionally, we extend IRW to handle scenarios with outliers by leveraging individual gradients as representations, allowing the method to effectively identify and remove potential outliers.
- We demonstrate the superiority of IRW through experimental comparisons with existing methods across various datasets. Our experiments also reveal the connections of all methods from the per-sample reweight strategy and exploit the potential impact of different choices of evaluation metrics.

II. Related Work

Despite the vast body of literature on fairness learning, we revisit three branches of research that are closely related to our work.

Utility-centric Fairness. Following Rawlsian fairness [20], many fairness learning works [40]–[44] have concentrated on model's utility across different groups. In classification tasks, the accuracy of each group data is expected to be same with each other, also termed as subgroup robustness [21], [45]. Regarding regression tasks, the utility is typically characterized by regression error [46], while in PCA the utility can be described into reconstruction error [24], [25]. With model utility in consideration, fairness can be achieved by minimizing utility discrepancies across groups. Recent work [47] treats learning on every group as a multiple coupled objective, and then seeks a set of optimal classifier parameters along the Pareto front that best balances the utility among all groups. Our work falls in the accuracy-based fair classification tasks but also consider alternative evaluations metrics, such as F1 score for imbalanced datasets.

Fairness without Demographics. A direct workaround for this problem is using proxy attributes. One can perform clustering on observed features and treat cluster indicators as groups [48], or select observed attributes under the assumption that these attributes are usually correlated [29]. Apart from these empirically approaches, fairness researchers proposed to approximate the sensitive attributes using a defined causal graph [31] or employ proxy models for prediction [32]. Given a group size, the Rawlsian fairness can be approximated without need of demographics, such as in [33], [40]. The key insight behind is that true worst-off group utility can be lower bounded when group size prior is known. We emphasize that such an approximation is dependent on the utility-centric fairness.

Sample Reweighting. Reweighting [49] is one of the earliest strategies to mitigate bias. However, standard reweighting techniques are designed for cohort-based approaches [12], [50], making them inapplicable to demographics-unaware scenarios. A pioneering work that connects worst-case fairness and sample reweighting is [40], which observes that the samples with higher loss tend to receive more focus during training. This finding inspires the subsequent research [21], [42] to optimize learnable sample weights. Our work builds on this insight by evaluating each sample's importance through its influence on the fairness objective. However, unlike [36], our method is not dependent on any additional validation set with demographic annotation, thereby being free of additional requirement on training dataset.

III. Problem Formulation

A. Worst-case Fairness

Let predictive accuracy serve as the utility of a classification model f_θ parameterized by θ. If a sensitive attribute splits training data into K groups, i.e., $G = \{G_1, G_2, ..., G_K\}$, Rawlsian max-min fairness, a.k.a. the worst-case fairness [21], optimizes the following problem,

$$\theta^* = \arg\max_{\theta} \min_{k\in[K]} \text{ACC}_k(\theta), \tag{1}$$

where $\text{ACC}_k(\theta)$ represents the accuracy of classifier f_θ on group G_k, i.e., $\text{ACC}_k(\theta) = \frac{1}{|G_k|}\sum_{i=1}^{|G_k|} \mathbb{1}(f_\theta(x) = y)$. Let worst-off group accuracy $\text{ACC}_{wg}(\theta) = \min_{k\in[K]} \text{ACC}_k(\theta)$. As $\text{ACC}_k(\theta) \leq 1 (\forall k \in [K])$, we have

$$|\text{ACC}_i(\theta) - \text{ACC}_j(\theta)| \leq 1 - \text{ACC}_{wg}(\theta), \ \ \forall i, j \in [K] \tag{2}$$

which indicates the utility difference between any two groups are upper bounded by $1 - \text{ACC}_{wg}(\theta)$. With this insight, Eq. (1) optimizes θ to maximize $\text{ACC}_{wg}(\theta)$ iteratively, thereby reducing the utility gap among groups.

Remark. We notice that worst-case study in fairness is also interpreted as fair treatment to tails samples [51], [52]. The line of works are built on extreme value theory [53] often works with provided sensitive attributes.

To avert the non-differentiable indicator function $\mathbb{1}(\cdot)$ during optimization, Eq. (1) is practically converted to minimizing the expected risk over the worst-off group,

$$\theta^* = \arg\min_{\theta} \max_{k\in[K]} \mathcal{J}_k(\theta), \quad \mathcal{J}_k(\theta) = \mathbb{E}_{z\sim P_k}[\ell(\theta; z)] \tag{3}$$

where the variable $z = (x, y)$, P_k denotes the distribution from which the members of G_k are drawn, and $\ell(;)$ is any feasible classification loss function, e.g., cross entropy. Similar to $\text{ACC}_{wg}(\theta)$, we denote $\mathcal{J}_{wg}(\theta) = \max_{k\in[K]} \mathcal{J}_k(\theta)$ for simple statement later. In particular, if f_θ belongs to the hypothesis class which achieves Pareto optimal classifiers, Eq. (3) is able to achieve the well-known harmless fairness if a binary sensitive attribute is applied. While for multi-value sensitive attributes, it has been found difficult to find an equal-group-loss plane [47].

When demographic information is unknown, group partition G then cannot not be accessible, making the inner optimization of Eq. (3) intractable. Section II has revisited how existing studies remedy this issue. Although not all particularly devised for the worst-case fairness, they are potentially used as workarounds if corresponding conditions are provided.

B. Setting: α-sized Worst-case Fairness

1) Minimal Group Proportion for Bounding: Distributional robustness around the empirical distribution with radius shrinking [54] offers a bridge between group and global empirical risks. Inspired by this finding, the pioneer work [40] addresses worst-case fairness without demographics by introducing the side information about group partitions. Specifically, they presume a lower bound on the group proportions $\alpha \leq \min_{k\in[K]} \frac{|G_k|}{|G|}$ by which a surrogate objective can be formulated. Formally, it minimizes the expected risk over a perturbed distribution, i.e., Q

$$\mathcal{J}_{dr}(\theta; \alpha) := \sup_{Q\in\mathcal{B}(P,r)} \mathbb{E}_{z\sim Q}[\ell(\theta; z)], \tag{4}$$

where $\mathcal{B}(P, r)$ is chi-squared ball centered at the data distribution P with radius $r = (1/\alpha - 1)^2$. $\mathcal{J}_{dr}(\theta; \alpha)$ is provably to bound the risk of each group $\mathcal{J}_k(\theta)$, and hence the worst-off group risk $\mathcal{J}_{wg}(\theta)$. In fact, given the minimal proportion size α and the total number of training examples N, it is verified that the empirical risk (denoted by adding a $\hat{}$ for differentiation, and same to others) of the worst-off group $\hat{\mathcal{J}}_{wg}(\theta)$ is upper bounded by the average risk of examples which achieve top $N\alpha$ largest losses

$$\begin{aligned}\hat{\mathcal{J}}_{wg}(\theta) &= \max_{k\in[K]} \frac{1}{|G_k|} \sum_{i=1}^{|G_k|} \ell(\theta; z_i) \\ &\leq \frac{1}{N\alpha} \sum_{i=1}^{N\alpha} \ell(\theta; z^{(i)}) := \hat{\mathcal{J}}_{N\alpha}(\theta; \alpha)\end{aligned} \tag{5}$$

where $\ell(\theta; z^{(i)})$ denotes i-th largest loss yielded by $z^{(i)}$, and $\hat{\mathcal{J}}_{N\alpha}(\theta; \alpha)$ is defined as the empirical risk on the group whose members have largest $N\alpha$ losses. Although the inequality of Eq. (5) is easy to reach, due to the nature of $\mathcal{J}_{dr}(\theta; \alpha)$, we instantly have the following proposition.

Proposition 1: Given a bound α for the minimal group proportion, we have $\mathcal{J}_{wg}(\theta) \leq \mathcal{J}_{N\alpha}(\theta; \alpha) \leq \mathcal{J}_{dr}(\theta; \alpha)$ for all $\theta \in \Theta$.

The proof of Proposition 1 is left to Appendix. Although optimizing $\mathcal{J}_{dr}(\theta; \alpha)$ leads to a convex formulation [54], we focus on minimizing $\hat{\mathcal{J}}_{N\alpha}(\theta; \alpha)$ of Eq. (5) that is more efficient to improve the worst-off group performance. Now, we formally frame the definition of the problem setting.

Definition 1: **(α-sized Worst-case Fairness)** A classifier f_θ is said to satisfy α-sized worst-case fairness principle on a training set of N examples if it maximizes the utility of the subgroup S, denoted by U_S, which is composed of $N\alpha$ examples and is with the lowest utility.

$$\theta^* = \arg\max_{\theta\in\Theta} \min_{|S|=N\alpha} U_S(\theta). \tag{6}$$

Similar to Eq. (3), by taking predictive accuracy as utility and selecting $N\alpha$ examples with largest losses, the inner objective of Eq. (6) can be implemented by the negative $\hat{\mathcal{J}}_{N\alpha}(\theta; \alpha)$. Hence, the full objective is written as

$$\min_{\theta} \frac{1}{N\alpha} \sum_{i=1}^{N\alpha} \ell(\theta; z^{(i)}). \tag{7}$$

Note that $z^{(i)}(i = 1, 2, ..., N\alpha)$ is dependent on θ.

2) Deriving α from Privacy Perspective: Note that both $\mathcal{J}_{dr}(\theta; \alpha)$ and $\mathcal{J}_{N\alpha}(\theta; \alpha)$ introduce the group proportion bound α in a mathematical manner. In practice, demographics are not available often because of privacy concerns, but we show in the following example that α could be properly estimated by a privacy technique called randomized response [55] even without true sensitive attribute labeling.

Example. Suppose gender is considered a sensitive attribute, with "female" being the protected group. When collecting data, we instruct each participant to perform a coin flip. If tails, they should respond truthfully. If heads, they should flip a second coin and respond "female" if heads and "male" if tails. We denote the true female proportion and the frequency of "female" answers by α and β, respectively. We know that the expected number of "female" answers is $1/4$ times the number of participants who are male plus $3/4$ times the actual number of female participants, that is, $\beta = (1/4)(1 - \alpha) + (3/4)\alpha$. Therefore, we can estimate the female proportion by $\alpha = 2\beta - 1/2$.

This example shows that the collected answers about gender can be very noisy (Please refer to Appendix for a robust analysis for this additional concern), as each individual has a plausible deniability to their outcomes, from which privacy comes. From the theory of differential privacy, we have the following theorem.

Theorem 1: The randomized response described in the above example is $(\ln 3, 0)$-differentially private.

The proof can be referred to Chapter 3 of [37]. The random response strategy has been adopted in the U.S. Census [56] to address related concerns. Note that the probability of deniability can be changeable if one is apt to adopt some other strategies. Generally, the estimated α will become more accurate if the number of participants is larger. And we use this α as an approximation of the minimal group proportion.

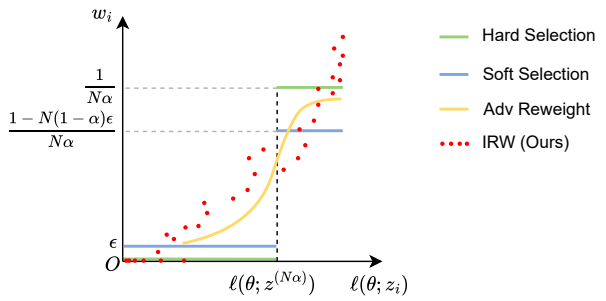


Fig. 1. Comparison of four weights learning strategies. The weight of our IRW method is not solely determined by loss; therefore we use discrete scatter plots to represent the variance introduced by other factors, i.e., gradients.

For multi-valued attributes such as race, if the protected group is known, one can reduce the setting to a binary case (e.g., 'Black' vs. 'non-Black') following this example. While this approach relies on a strong prior, which may still hold in certain real-world applications, it also highlights an opportunity for future research in the social sciences to explore more nuanced treatments of group protection.

IV. Reweighting Solution

A. Per-sample Weight Learning

Optimizing Eq. (7) is also referred as minimizing top-$N\alpha$ loss in the recent study [36], which essentially encourages the model to fit more difficult examples only, that is,

$$\min_{\theta} \frac{1}{N\alpha} \sum_{i=1}^{N} [\ell(\theta; z_i) - \ell(\theta; z^{(N\alpha+1)})]_+ + \ell(\theta; z^{(N\alpha+1)}), \quad (8)$$

where $[a]_+ = \max(a, 0)$. Note that this objective minimizes the α-fraction of training points that have the highest loss [45]. If α is small, there will be very few examples that participate the training per iteration, incurring an inferior model utility (See the results of CVaR [33] in Table I). Essentially, this sample selection process can be understood as assigning a set of weights to the training examples, where Eq. (8) effectively performs a hard selection. To better utilize the training samples, a soft selection approach can be adopted by maintaining a minimum weight for samples whose losses are below the threshold. Since the weights lie on a $(N-1)$-dimensional simplex, denoted as Δ^{N-1}, we can write the per-sample weight as

$$w_i = \begin{cases} \epsilon, & \text{if } \ell(\theta; z_i) < \ell(\theta; z^{(N\alpha)}) \\ \frac{1-N(1-\alpha)\epsilon}{N\alpha}, & \text{otherwise.} \end{cases} \quad (9)$$

Obviously, Eq. (9) reduces to the hard selection if $\epsilon = 0$. Instead of manually assigning binary weights, adversarial reweighting fairness works [21], [42] propose that these weights are learnable through the following form:

$$\max_{w \in \Delta^{N-1}} \sum_{i=1}^{N} w_i \ell(\theta; z_i) \;\; w \in C, \quad (10)$$

where C denotes any predefined constraint which could be a specific bound for weights [21] or a complexity control applied to the function that generates weights [42], preventing from trivial solutions to Eq. (10).

Theorem 2: If the feasible set defined by C is convex, the optimal solution of Eq. (10) implies that a higher loss corresponds strictly to a higher weight.

The proof of Theorem 2 is left to Appendix. Theorem 2 suggests that the fairness methods which optimize Eq. (10) are essentially searching for a set of weights that are monotonically non-decreasing with respect to the loss, guided by a prior belief encoded in C. We summarize these weight distributions into Fig. 1, where both hard and soft selection appear as staircase-like lines, while adversarial reweighting corresponds to a smooth curve that non-decreases with the loss. In particular, following the formulation of [21], the minimum weight ϵ is maintained for the samples whose losses are relative smaller in soft selection and adversarial reweighting, which serves as a hyper-parameter in corresponding algorithms implementation.

B. Intrinsic Reweighting (IRW)

The weight of each training sample should reflect its actual contribution to the fairness objective, i.e., average loss of top-$N\alpha$ samples. We question the weight strategy that samples with larger losses will be strictly assigned to larger weights according to Theorem 2. This approach overlooks the fact that it is the gradient that drives model updates during training; thus a sample with a large loss but a zero gradient should not be considered contributive. We present two numerical cases to justify this claim.

Numerical Case 1. We consider a sigmoid model $\hat{y} = \sigma(w^T x)$ with a binary cross-entropy loss $\ell(y, \hat{y}) = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$. The gradient with respect to w is $\nabla_w \ell = (\hat{y} - y) \cdot x$. Given two samples with scalar inputs $x_1 = 2, y_1 = 0$, $x_2 \approx -0.7461, y_2 = 0$, and $w = 1$, we observe that their losses differ $\ell_1 \approx 0.1269$ and $\ell_2 \approx 0.3881$, while their gradient are nearly same: $\nabla_w \ell_1 \approx \nabla_w \ell_2 \approx -0.24$. Note that the gradient can be exactly equal by solving for a more precise x_2.

Numerical Case 2. We consider the same model and loss function with case 1, but use two new samples: $x_1 = 4, y_1 = 1$ and $x_2 = 0.5, y_2 = 0$. Again by setting $w = 1$, we find that two samples have nearly same $\ell_1 \approx \ell_2 \approx 0.0182$. However, their gradients differ significantly: $\nabla_w \ell_1 \approx -0.0720$ and $\nabla_w \ell_2 \approx 0.3113$. Notably, the gradients point in opposite directions.

The sample weights are used to influence the learning dynamics so that the trained model improves worst-case fairness. Traditional loss-based weighting strategies assume that samples with larger losses should receive larger weights. However, since model training is fundamentally driven by gradients, not loss values alone, this approach can be misleading—samples with high loss but negligible gradients do not meaningfully influence the model update. Please see the illustrative comparison in Fig. 1, where the "true" sample contribution can be inconsistent with adversarial reweighting. Influence function [38] offers a principled solution to this

challenge; it estimates how a small change in the weight of a training sample would affect the model's learned parameters. By capturing the first-order effect of a sample on the model, influence functions allow us to identify and prioritize samples that truly impact the model's worst-off performance.

With this understanding in mind, we look back to Eq. (7) and consider reassigning the weights to the training examples, which can directly reduce the largest $N\alpha$ losses. That means, we intend to learn a weight assignment such that by minimizing reweighted losses we can also achieve the α-sized worst-case fairness:

$$\theta^*(w) = \arg\min_{\theta} \sum_{i=1}^{N} w_i \ell(\theta; z_i), \tag{11a}$$

$$w^* = \arg\min_{w} \underbrace{\frac{1}{N\alpha} \sum_{i=1}^{N} [\ell(\theta^*(w); z_i) - \ell(\theta^*(w); z^{(N\alpha+1)})]_+}_{\mathcal{F}(\theta^*(w))}. \tag{11b}$$

Finding the optimal w^* in Eq. (11b) is difficult because we need to first optimize Eq. (11a). Inspired by the idea of examples reweighing [57] via influence function, for each example z_i we approximate its weight at t-th iteration by first perturbing its weight by $\epsilon_{t,i}$ and taking the single step gradient evaluated at zero

$$v_{t,i} = \left. \frac{\partial \mathcal{F}(\theta_t)}{\partial \epsilon_{t,i}} \right|_{\epsilon_{t,i}=0} \approx (\nabla_\theta \mathcal{F}(\theta_t))^\top \nabla_\theta \ell(\theta_t; z_i). \tag{12}$$

Without reiterating the details of how the approximation is derived, we can observe that the contribution of each sample depends on its gradient similarity with $\nabla_\theta \mathcal{F}(\theta_t)$, the gradient of worst-off group. To avoid the negative weights, we further rectify each $v_{t,i}$ and normalize the weights,

$$w_{t,i} = \frac{[v_{t,i}]_+}{\sum_j [v_{t,j}]_+ + \delta\left(\sum_j [v_{t,j}]_+\right)}, \tag{13}$$

where $\delta(a) = 1$ if $a = 0$ and otherwise $\delta(a) = 0$, which rules out a zero denominator.

We can see from Eq. (12), the value of $v_{t,i}$ is determined by the gradient similarity between i-th training example and the top $N\alpha$ largest training samples. Intuitively, an example whose loss is above $\ell(\theta_t; z^{(N\alpha+1)})$ will be more likely selected as a component of $\mathcal{F}(\theta_t)$ and thus the inner product tends to be positive. This agrees with the existing loss-based reweighting principle. In particular, we highlight that the samples whose loss are below $\ell(\theta_t; z^{(N\alpha+1)})$ also have chance to produce a positive $v_{t,i}$ if its direction is close to the combined gradient $\nabla_\theta \mathcal{F}(\theta_t)$. In Section VI-C, we will examine how the per-sample weights obtained by IRW disrupt the loss-based ordering.

C. Stochastic Update

Please note that Eqs. (11a) and (11b) are both defined globally, meaning that we need to optimize model parameters and sample weight with full-batch data update. To scale up to large datasets, we adopt stochastic update by proposing two schemes. (i) *Global.* At the beginning of each epoch, we identify the top-$N\alpha$ samples by Eq. (11b) and compute sample weights w^* across full training data. By assuming that the weights change within each epoch can be negligible, as the same style as [21], we do not re-compute w^* for them to save computation costs. (ii) *Local.* We use local top-$B\alpha$ samples on each mini-batch for Eq. (11b), and compute batch-wise sample weights on a simplex of Δ^{B-1}. The difference of weight scale between two schemes will be absorbed into learning rates.

Complexity. The additional cost of IRW with respect to a standard training is the instance-wise backward propagations, as shown by step 5 in Algorithm 1. As is known to us, modern neural networks stores cumulative statistics in a batch rather than individual sample's gradients. A naive solution is to perform single-sample mini-batches, ideally in parallel, which remains too slow and incompatible with batch-based operators like batch normalization. We remedy this problem by using function transforms [2] which compute per-sample-gradients in an efficient way. Given this implementation, we realize that the time complexity difference between above two training schemes primarily arises from the sorting process. For the global weights in scheme (i), finding the top $N\alpha$ from N training examples based on their losses has a time complexity of $O(N \log(N\alpha))$. In contrast, for the local weights of scheme (ii), the time-complexity is $O(N \log(B\alpha))$. In practical implementation, we experimentally observe that scheme (i) is slightly faster due to the multiple graph retention steps in scheme (ii) (See Section VI-E).

V. Outliers Removal from Gradient View

The worst-case formulation, i.e., Eq (7), is vulnerable to outliers. Since we have the gradient for each training sample, we can filter out outliers by naturally utilizing such gradient information, which is more reliable than loss values, as we have claimed Section IV-A.

Inspired by [35], we develop a variant of IRW, named by IRWO, to handle potential outliers during training. Specifically, we first compute pairwise distance between every two samples' gradient within a batch. Let us define $g_i := \nabla_\theta \ell(\theta; z_i)$ where we omit the iteration index for θ for convenience. Then each entry of the distance matrix M is derived by

$$M_{ij} = 1 - \frac{\tilde{g_i}^\top \tilde{g_j}}{||\tilde{g_i}|| \cdot ||\tilde{g_j}||}, \tag{14}$$

where the decentralized gradient $\tilde{g_i} = g_i - \bar{g}$ with $\bar{g}$ as the mean of per-sample gradient. We treat each row of the distance matrix M as the new representation of each sample, which better characterizes the local density and also improves the clustering efficiency when parameter space is larger than data size, i.e., $|\theta| > |\mathcal{D}|$. We then identify outliers by searching for low-density regions. In our method, we adopt DBSCAN clustering[3] before selecting the worst-off group. Typically, training samples labeled as -1 after clustering are considered

[2] https://pytorch.org/tutorials/intermediate/per_sample_grads.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Algorithm 1 IRW algorithm for α-sized worst-case fairness

Input: Training set $\mathcal{D} = \{z_i\}_{i=1}^N$, where $z_i = (x_i, y_i)$, period τ, and group ratio α
Output: Learned model parameterized by θ

1: Initialize parameters θ
2: **for** $t = 0, 1, ..., T-1$ **do**
3: **if** Outliers exist and $T \bmod \tau = 0$ **then**
4: Perform forward passes and get $\{\ell(\theta_t; z_i)\}_{i=1}^{|\mathcal{D}|}$
5: Perform instance-wise backward propagations and get $\{g_i\}_{i=1}^{|\mathcal{D}|}$
6: Compute $\forall i\ \tilde{g}_i = g_i - \bar{g}$ where $\bar{g} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} g_i$
7: Compute distance matrix M by Eq. (14)
8: Perform clustering $\{s_i\} \leftarrow DBSCAN(M)$, where $s_i = -1$ indicates outliers
9: Update $\mathcal{D}$ by removing outliers
10: **end if**
11: Sample a mini-batch samples $\mathcal{B} \subset \mathcal{D}$
12: Perform forward passes and get $\{\ell(\theta_t; z_i)\}_{i=1}^{|\mathcal{B}|}$
13: Perform instance-wise backward propagations and get $\{g_i\}_{i=1}^{|\mathcal{B}|}$
14: Obtain $\nabla_\theta \mathcal{F}(\theta_t)$ by calculating the mean of the gradients whose losses are top-$|\mathcal{B}|\alpha$ largest
15: Update $\{v_{t,i}\}_{i=1}^B$ according to Eq. (12)
16: Update $\{w_{t,i}\}_{i=1}^B$ according to Eq. (13)
17: Update $\theta_{t+1} = \theta_t - \sum_{i=1}^B w_{t,i} g_i$
18: **end for**

outliers. Since clustering are conducted on entire training set[4], which remains expensive to acquire outliers, we choose to periodically perform this process. For instance, we can do every τ steps, as presented in summarized Algorithm 1, which describes the detailed process of local update scheme. In our experiments, τ is set as the number of batches, and thus we do outliers removal once per epoch. For the global update, i.e., scheme (i) of Section IV-C which requires to prepare all the sample weights before stochastic optimization, we can simply merge steps 14-17 into outliers removal procedure.

Related strategies. We present different methods for handling outliers in worst-case fairness studies, each operating under specific assumptions. ARL [42] addresses this problem by employing a linear adversary which helps focus on computationally-identifiable errors. DORO [34] views a certain proportion (an additional hyper-parameter) of samples which achieve largest losses as outliers, purifying the worst-off group. GraSP [35] conducts density-based clustering in gradient space where both outliers and groups are indicated by clustering labels. Very recent work GoG [39] takes the similar insight but leverages the gradient information of the learner from a graph-based perspective, which is used to help avoid assigning superior weights to noisy outliers. We follow the idea of GraSP to remove potential outliers before we compute per-sample weight, i.e., Eq. (12), since the well-prepared per-sample gradients are ready to use. The modified method is named by IRWO for simplicity. Their experimental comparison can be referred to Section VI-F.

[4]Performing clustering on a mini-batch may lead to misidentifying minority samples as outliers.

VI. Experiment

Dataset. We consider four benchmark datasets commonly used for fairness evaluation [42], [59], encompassing three binary classification tasks and one multi-class classification task. (i) UCI Adult [60]: Predicting whether an individual's annual income is above 50,000 USD. (ii) Law School [61]: Predicting the success of bar exam candidates. (iii) COMPAS [62]: Predicting recidivism for each convicted individual. (iv) CelebA [63]: Predict the hair color for face images. Our experiments follow the recent fairness studies [21], [36] which employ specific sensitive attributes for fairness evaluation; gender and race are selected for datasets (i)-(iii) and gender and young are selected for CelebA.

Metrics. During the test process, we use provided sensitive attributes to divide test set into disjoint groups. Since the worst-case fairness inherits John Rawls' equal utility principle, following the research [34], we evaluate the performance of trained models based on their overall accuracy (ACC) and worst-off group accuracy (WACC), where the model fairness can be expressed as the gap between ACC and WACC, that is

$$\Delta = |\text{ACC} - \text{WACC}|.$$

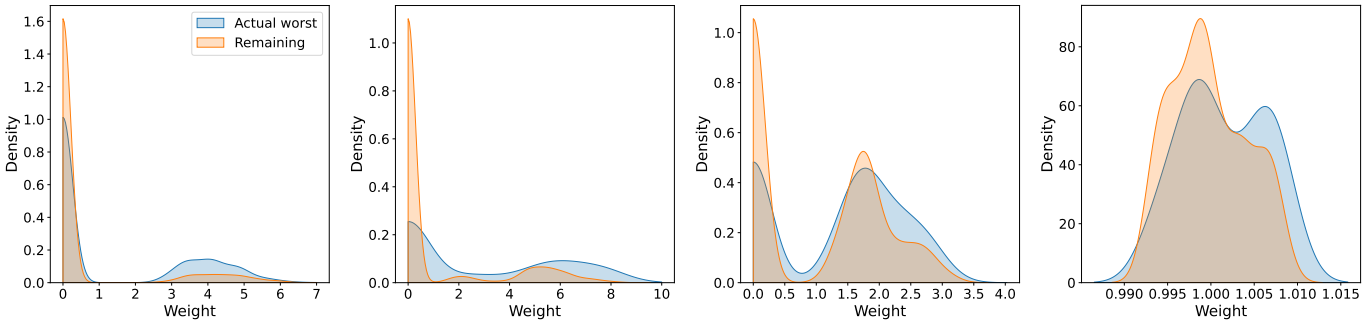
Apparently, $\Delta = 0$ indicates an ideally fair model. This serves as a standard fairness metric for the worst-case fairness problem, as established in prior work [34]. We do not exhaustively evaluate other fairness metrics such as Demographic Parity and Equal Opportunity [17], as they focus on *true positive* predictions only and are therefore not applicable to our setting.

A. Overall Comparison

In the context of α-sized worst-case fairness, the corresponding minimal group ratio α (the minimum ratio among four groups, as shown in the header row of Table I) for each dataset is given for model training. We take the following methods (i-iii) which are direct baselines for solving α-sized worst-case fairness problem, while (iv) is used as a reference which explicitly uses sensitive attributed during training. (i) DRO [40]: This method derives a convex surrogate loss which upper bounds the empirical risk of the worst-off group. Please see the details back to Section III-B1. (ii) CVaR [33]: It is known as the Conditional Value at Risk of level α, which is equivalent to minimizing the α-fraction of training points that have the highest loss [45]. (iii) BPF [21]: Blind Pareto Fairness optimizes sample weights using projected gradient descent in an adversarial manner (See Appendix B.1 of [64]), where samples with higher loss will be assigned greater weights during the training process. (iv) Reckoner [58]: This method distills fairness knowledge from low-confidence prediction samples, which aligns with findings from DRO and CVaR, although it is not explicitly dependent on α in its modeling.(v) MMPF [47]: MiniMax Pareto Fairness is an representative attributes-aware method which explicitly uses sensitive attributes to split training data and encourages the equal model accuracy across groups via

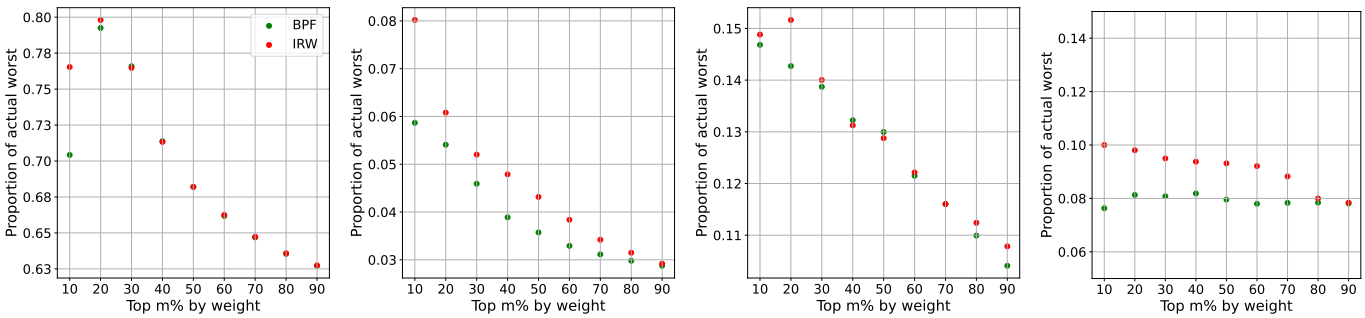
TABLE I
TEST PERFORMANCE COMPARISON ON FOUR BENCHMARK DATASETS. THE FIRST BLOCK OF METHODS ARE LEARNING W/O ANY SPECIFIED ATTRIBUTES BUT WITH THE GROUP RATIO α, WHERE THE BEST RESULTS ARE MARKED IN BOLD, AND THE SECOND BEST ARE UNDERLINED. THE LAST SECOND BLOCK METHOD, I.E., MMPF, EXPLICITLY USES ACTUAL SENSITIVE ATTRIBUTES DURING TRAINING. WE OMIT THE STANDARD VARIANCES OF Δ.

Method	UCI Adult ($\alpha = 4.78\%$)			Law School ($\alpha = 2.74\%$)			COMPAS ($\alpha = 9.46\%$)			CelebA ($\alpha = 7.23\%$)		
	ACC ↑	WACC ↑	$\Delta\downarrow$	ACC ↑	WACC ↑	$\Delta\downarrow$	ACC ↑	WACC ↑	$\Delta\downarrow$	ACC ↑	WACC ↑	$\Delta\downarrow$
DRO [40]	0.7630±0.0726	0.6965±0.0676	0.0665	0.7989±0.1791	0.5613±0.1140	0.2373	0.5492±0.0253	0.4770±0.0556	0.0722	0.5219±0.1219	0.4586±0.0945	0.0633
CVaR [33]	0.7511±0.0269	0.6992±0.0292	0.0519	0.7503±0.0501	0.5708±0.0337	0.1795	0.5343±0.0201	0.4875±0.0325	0.0468	0.6903±0.0093	0.6475±0.0598	0.0428
BPF [21]	0.8206±0.0072	0.7692±0.0052	0.0514	0.8698±0.0024	0.7592±0.0122	0.1106	0.5989±0.0144	0.5641±0.0241	0.0348	**0.9287**±0.0273	0.9031±0.0135	0.0256
Reckoner [58]	0.8395±0.0152	0.7830±0.0158	0.0515	0.8304±0.0032	0.7697±0.0057	**0.0607**	0.6455±0.0063	0.5745±0.0187	0.0710	0.7947± 0.0025	0.7752± 0.0023	0.0195
IRW (Ours)	**0.8438**±0.0074	**0.7980**±0.0105	**0.0458**	**0.8715**±0.0012	**0.7736**±0.0082	0.0979	**0.6628**±0.0068	**0.6457**±0.0187	**0.0171**	0.9243±0.0057	**0.9109**±0.0027	**0.0134**
MMPF [47]	0.8533±0.0192	0.8123±0.0920	0.0410	0.8832±0.0142	0.7870±0.0124	0.0962	0.6639±0.0025	0.6390± 0.0145	0.0248	0.9291±0.0048	0.9158±0.0052	0.0133



(a) UCI Adult (b) Law School (c) COMPAS (d) CelebA

Fig. 2. Weight distribution of training samples for our IRW. For each dataset, we record the weights of training samples at the first training epoch and use kernel density estimation to plot their distribution. The actual worst indicates the group which has the lowest accuracy on test set, and all rest are included in the remaining.



(a) UCI Adult (b) Law School (c) COMPAS (d) CelebA

Fig. 3. The proportion of actual worst-off group among top $m\%(m = 10, 20, ..., 90)$ training samples selected by reassigned weights. We compare BPF and our IRW method on four datasets.

solving a multi-objective problem. This method is an upper-bound method of Group DRO [65] in term of fairness metric Δ. Each experiment will be repeated for 10 times, and the average and standard deviation are reported.

Table I shows test ACC, WACC, and Δ of each method. We can see that: (1) Our IRW method achieves competitive or leading performance across most datasets, demonstrating the effectiveness of leveraging individual gradients to compute sample importance. Also, IRW achieves comparable or slightly worse performance than MMPF, which has utilized sensitive attributes during training. (2) BPF most often achieves the second or third best performance, demonstrating that adversarial reweighing strategy is useful in achieving α-sized fairness. It is worthwhile to point out that BPF involves a few of hyper-parameters and also heuristic controls[5], and we applied their default setup throughout the experiments. (3) Reckoner achieves comparable results to BPF on the three tabular datasets, but it performs poorly on the image classification task. One possible explanation is that Reckoner needs to generate high-dimensional noise for the image data, which makes fairness knowledge distillation unstable compared to the tabular setting. (4) CVaR uses only the α-fraction of data for model training, resulting in a significant drop in ACC and WACC, as well as in fairness results, compared to BPF, Reckoner, and our IRW. (5) DRO usually achieves the worst performance expect that it is slightly better on accuracy than CVaR for the COMPAS dataset, which is attributed to the commonly smaller α across all datasets (also observed

[5]Refer to their officially implemented code repository via https://github.com/natalialmg/BlindParetoFairness.

in experiments of [21]). Because a smaller α indicates a larger radius of the chi-squared ball in Eq. (4), leading to an unavoidable loss in model accuracy. Such unsatisfying model utility makes DRO impractical to use, for example, on the CelebA dataset. Please also refer to Fig. 5 for checking its predictions.

B. IRW Works Better than BPF

We first verify the efficacy of IRW by examining how samples from the actual worst-off group are treated during training, given only a group ratio prior. To do this, we monitor the individual weights assigned to all training samples. To align with the standard training process, we keep a base weight of 1, with weights higher or lower than 1 being increased or decreased accordingly. Other subsequent presentations regarding to IRW follow this setup. Fig. 2 depicts the weight distribution of training samples by splitting all training samples into the *actual worst*-off group and *remaining* samples. Across all datasets, we observe that while the weight distribution does not form two distinct peaks, as would be ideal, the actual worst-off group samples tend to have a relatively higher density in the range of larger weights. This indicates that IRW implicitly upweights the actual worst-off group samples, even though sensitive attributes are not accessible during model training. This observation evidences the efficacy of our IRW method.

Although can be understood by Eq. (10), BPF practically uses trainings samples' weights which do not lies on a simplex. Normalizing them may distort the weight density. To further exploit why our method IRW can outperform BPF, we compare them using the following approach. We select the top $m\%$ training samples based on their corresponding weights, and compute the proportion of actual worst-off group samples within these selected samples. Since we expect that the actual worst-off group can be assigned with higher weights, a higher proportion usually indicates a fairer model. The results are shown as Fig. 3, from which we observe that at almost every m across four datasets, our IRW achieves the higher proportion value than BPF, demonstrating that our method can better leverage prior to implicitly encourage the fitting for the actual worst-off groups. It is also worthwhile to notice that the proportion on the UCI Adult dataset is not aligned with the prior (63% vs. 4.78%) when m approaches to 100 while other three do. This is because the minority group sometimes is not necessarily the underrepresented group. Although loosing the upper bound in Eq. (5), it shows a distinction assumption that minority is always underrepresented [45].

C. Per-sample Reweight Mechanism

Figs. 2 and 3 demonstrate that IRW increases the weights of training samples from the actual worst-off group. To delve deeper, we analyze how these weights correlate with their loss values, which is central to the per-sample weight mechanism for fairness. To this end, we randomly select a batch of samples at the end of the training phase on four datasets, where the models are assumed to be close to convergence, with both weight and loss values remaining stable afterwards.

TABLE II
TEST PERFORMANCE COMPARISON ON THE UCI ADULT DATASET WHERE SPECIFIC SENSITIVE ATTRIBUTES ARE USED FOR SPLITTING TEST SET INTO DIFFERENT GROUPS. THE BEST RESULTS ARE MARKED AS BOLD, AND THE SECOND BEST ARE UNDERLINED.

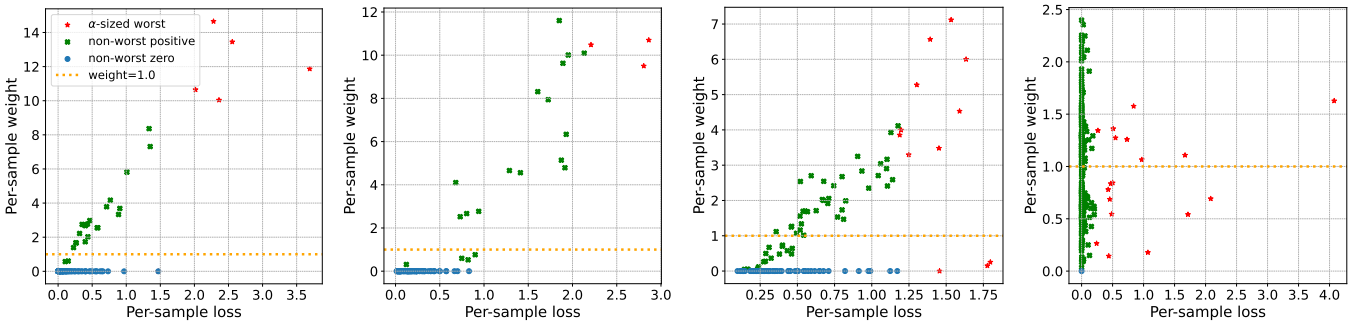
Method	F1 ↑	WF1 ↑	Δ_{F1} ↓
DRO	0.3796±0.0726	0.1581±0.0447	0.2215
CVaR	0.3999±0.1311	0.2231±0.0945	0.1768
BPF	<u>0.4841</u>±0.0517	<u>0.4079</u>±0.0631	<u>0.0762</u>
Reckoner	0.3092±0.0517	0.1579±0.0631	0.1513
IRW	**0.5749**±0.0493	**0.5024**±0.0502	**0.0725**

Fig. 4 shows the experimental results of per-sample weight versus loss, based on which we have following observations. (1) IRW typically assigns greater weights to samples with greater losses, aligning with the idea that poorly-fit samples are more likely to come from underrepresented groups. (2) The weights of α-sized worst samples are not necessarily larger than those of other samples with relatively smaller, as observed on last three datasets, which thus differs from CVaR. This occurs because the weight of an α-sized worst sample may be smaller if its gradient is not consistent with the combined gradient, as indicated by Eq. (12). (3) IRW can produce sparse weights, with zero weights often assigned to samples with smaller losses. This property eliminates the need to manually enforce a minimum update, as required in BPF, which typically demands additional effort to identify. (4) The weight-loss correlation on CelebA stands out from the others, where many training samples with very small losses have increased weights, while some of the α-sized worst samples have decreased weights. This suggests that, in our IRW approach, training samples contributing to fairness from a gradient perspective will be widely identified, even if their losses are negligible.

D. Evaluation Metric Matters

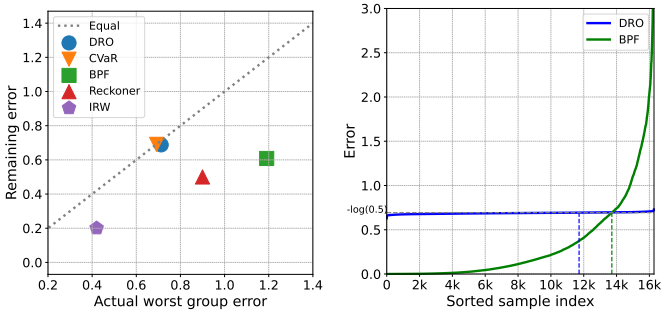
Class imbalance. F-score is particularly used in binary classification tasks where one class may be significantly underrepresented. We realize that the UCI Adult dataset tends to be dominated by true negative samples due to its imbalanced nature, with a positive to negative ratio of 1:4. Without changing each training method, we evaluate their test performance by utilizing the F-score metric. The results are shown as Table II. It is evident that with F1 being the classification metric, IRW consistently achieves the best performance, demonstrating that despite originating from group accuracy, i.e., Eq. (1), our method IRW is not biased towards accuracy metric. Furthermore, we also observe that the performance difference between IRW and BPF is larger in terms of F1 measurement, which justifies the advantage of leveraging per-sample gradient information in our reweighing strategy.

Classification error. According to Table I, we have observed that DRO and CVaR are performing worse than any other methods, even though CVaR exactly formulates the worst-case loss and DRO is also an upper bound. To understand this result, we take a closer look at their classification errors on test set, with the results on Adult as an example. Fig. 5(a) shows the



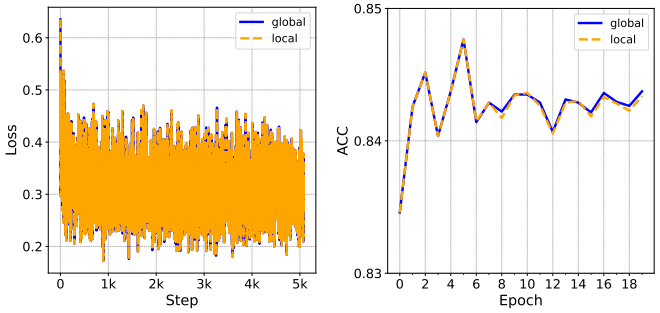
(a) UCI Adult (b) Law School (c) COMPAS (d) CelebA

Fig. 4. Per-sample weight versus their loss values on a randomly selected batch at the end of training process on four datasets. The number of α-sized worst samples (red stars) is $\text{int}(b \times \alpha)$, where $b = 128$ for (a)-(c) and $b = 256$ for (d), and α is indicated by the head row of Table I. The non-worst positive-weighted samples (green crosses) and non-worst zero-weighted samples (blue points) are the ones whose gradients yield positive or zero w_i at the iteration t, respectively. (Refer to Eq. (13)).



(a) Group-wise error (b) Instance-wise error

Fig. 5. Classification errors comparison on the UCI Adult dataset. The y-axis is truncated above 3 for better visualization.



(a) Training loss (b) Test accuracy

Fig. 6. Comparison between global and local worst-case curves on the UCI Adult dataset, where (a) is the loss versus each training step and (b) is test accuracy on each epoch.

classification error of the actual worst-off group versus remaining samples. We observe that DRO and CVaR both achieve the approximately equal errors which however increase the overall error by comparing with IRW, and BPF is the least fair in terms of error. Such results are somehow contrary to what we have seen in Table I. For example, DRO achieves a lower overall error than BPF, yet its accuracy is identified as the worst. We further explain such inconsistency on metrics by connecting prediction errors and decision making. Fig. 5(b) shows the instance-wise errors, with the x-axis representing the test sample indices, sorted in an ascending order of their errors. The y-axis is truncated above 3 for improved clarity in presentation. Since this is a binary classification on the Adult dataset, we adopt the reference error of $-\log(0.5)$ as a decision boundary (indicated by dashed vertical lines) to categorize all test samples. Samples with an error below this value are considered correctly classified, while those above it are regarded as misclassified. We observe that although the errors of misclassified samples of BPF can be large, the overall accuracy of BPF remarkably outperforms DRO, as the dashed line for BPF is substantially to the right of the dashed line for DRO. It is also worthwhile to note that the performance gap between DRO and BPF in Fig. 5(b) is smaller than the reported average value in Table I, which is attributed to the larger variance of DRO. We defer more experimental verification on other datasets to the Appendix.

E. Global vs. Local Worst-Case Fairness Training

We compare the efficiency of global and local worst-case training with the Adult dataset as an example, corresponding to schemes (i) and (ii) in Section IV-C. Their training loss curves and overall test accuracy are exhibited as Fig. 6. The results show that local worst-case training which selects α-sized worst samples among each mini-batch are aligned with the global counterpart; their training loss and test accuracy curves are close to each other, with a very tiny discrepancy after 7th epoch from Fig. 6(b). Apart from this conclusion, we observe that their loss curves fluctuate significantly within the range of 0.2 to 0.4, which is essentially shared by all worst-case reweighting strategy, as also noted by [34]. Similarly, their test accuracy fluctuates across epochs, but within a relatively smaller range of around 0.5%. Additionally, we record the wall-clock time for training; global training takes around 31 seconds while local training takes 42 seconds. By tracking the time consumption of each component during local training, we find that multiple graph-retention steps (e.g., `retain_graph=true` in Python) increase the overall time cost, a process that is not required in global training. However, local training remains an acceptable option when full-batch forward passes are not feasible in some scenarios, e.g., limited computation resources.

F. Robustness Against Outliers

We compare IRWO with foure fairness methods ARL [42], DORO [34], GraSP [35], and GoG [39], which have been used to address outliers for worst-case modeling, as discussed in Section V. For methods with specific hyper-parameters, we use their default setup to report the final results. The COMPAS and CelebA datasets are used in this group of experiments because they two are considered to contain outliers [34]. Regarding CelebA, we found that some hair color labels are incorrect—for example, some individuals are labeled with multiple hair colors. These samples were removed from the overall comparison in Section VI-A but retained for this group of experiments.

Table III summarizes the results of all robust worst-case fairness methods, from which we have the following observations. (1) IRWO achieves the best performance among all the methods compared, showing a great potential of handling outliers with gradient information. (2) Referring to the results in Table I, we find that the variants with outliers removal improve the standard counterparts; DORO significantly improves DRO and IRWO also generally outperforms IRW over accuracy. (3) ARL performs slightly worse than our IRWO, which can be attributed to its formulation that does not incorporate α. Its improved version, GoG, which aims to capture gradient similarities through a graph-based perspective, improves ARL on CelebA but fails on COMPAS—possibly due to the choice of neighborhood hops in the graph. (4) GraSP uses the same outlier removal strategy as ours but underperforms due to its reliance on estimated sensitive attributes from clustering results.

We additionally evaluate the training time of all methods on the COMPAS and CelebA dataset. The average wall-clock time across 10 trials is reported in Table IV, where the methods in the second column additional consider outliers problems. Based on the results, we make the following two observations. (1) Since our method requires two backward passes per batch update, its running cost is approximately twice that of CVaR and BPF. Reckoner has the highest running time, as it involves a complex teacher-student knowledge sharing framework. (2) For methods that handle outliers, GoG considers graph structure over individual gradients but in a same framework with ARL, and thus it leads to higher training cost. DORO is the most efficient among these methods, as it simply relies on a specified outlier ratio. Our IRWO adopts a similar strategy to GraSP, which requires clustering to identify outliers, resulting in relatively slower training.

VII. Discussion

Beyond loss values, we have found that individual gradients provide a complementary view for characterizing a sample's importance to the fairness objective. However, gradients are sometimes too complex to work with, especially in larger neural networks involving millions of billions of parameters. In such cases, the gradient representation resides in a high-dimensional space, posing challenges for storage and computation. One workaround we apply to the experiments on CelebA dataset is selecting the last few layers, which are more closely related to label information from the perspective of information bottleneck [66]. We emphasize that the effective selection of key layers for specific tasks remains an open question, while some potential approaches can be considered, such as fisher information [67] or layerwise search [68].

Resampling has been a parallel way to achieve fairness [69], [70]. Despite various resampling motivations applied in these works, we are interested in its technical connection to reweighting in the context of α-sized worst-case fairness objective. Very fundamentally, we question whether the weights derived from IRW can be treated as the probability of resampling during the training process. We provide a corollary based on the recent work in group-imbalance learning [71], which claims that resampling always exhibits lower variance than the corresponding reweighting. While this corollary holds theoretically only when stochastic gradient descent is used (Please refer to Appendix for a formal presentation), we will explore this direction in future work.

While our experiments applied two selected attributes, ensuring fairness across multiple attributes remains challenging [72] due to fairness gerrymandering [73]. This issue extends beyond the capabilities of α-sized worst-case fairness formulations, indicating a potential limitation of this approach. In addition, our fairness learning setup is motivated by data privacy concerns, which thus links our work to the privacy-preserving fairness learning studies [74], [75] that primarily focus on decentralized learning contexts. It is also worth noting that fairness concepts in the literature can be task-dependent and not necessarily limited to sensitive attributes, as seen in category-based definitions in classification [76], [77]. This inspires us to enable our methodology to operate with only class ratio bounds provided during training, broadening its potential applications in future works.

VIII. Conclusion

This paper introduces "α-sized worst-case fairness" problem where only the the minimal group ratio α is known during training, but the goal is to diminish the prejudice on under-represented subgroups defined by inaccessible demographic information. We justify the use of this group ratio prior, which can be derived from random responses ensured by the well-known local differential privacy framework. Unlike existing methods that simply prioritize training on samples with relatively larger loss, we propose an intrinsic reweighting strategy that assigns individual weights based on per-sample gradient information. This gradient-based approach also naturally helps in removing potential outliers, which have been identified as a significant risk in worst-case fairness formulations. Our experiments demonstrate the superiority of IRW and its variant, IRWO, in achieving fairness across four benchmark datasets, establishing a foundation of benchmark results for future research.

Appendix

Proposition 1. Given the minimal group proportion α, we have $\mathcal{J}_{wg}(\theta) \leq \mathcal{J}_{N\alpha}(\theta;\alpha) \leq \mathcal{J}_{dr}(\theta;\alpha)$ for all $\theta \in \Theta$.

TABLE III
TEST PERFORMANCE COMPARISON ON FOUR BENCHMARK DATASETS WHERE SPECIFIC SENSITIVE ATTRIBUTES ARE USED FOR SPLITTING TEST SET INTO DIFFERENT GROUPS. THE BEST RESULTS ARE MARKED AS BOLD, AND THE SECOND BEST ARE UNDERLINED.

Method	COMPAS			CelebA		
	ACC ↑	WACC ↑	Δ ↓	ACC ↑	WACC ↑	Δ ↓
ARL [42]	<u>0.6612</u>±0.0091	<u>0.6255</u>±0.0188	<u>0.0357</u>	0.9058±0.1038	0.8689±0.1004	0.0369
DORO [34]	0.5719±0.0241	0.5241±0.0427	0.0478	0.7714±0.0333	0.7350±0.0400	0.0364
GraSP [35]	0.6009±0.0211	0.5592±0.0481	0.0417	<u>0.9255</u>±0.0501	<u>0.8969</u>±0.0337	0.0286
GOG [39]	0.6598±0.0340	0.6142±0.0325	0.0456	0.9125±0.0501	0.8864±0.0337	<u>0.0261</u>
IRWO (Ours)	**0.6688**±0.0113	**0.6388**±0.0213	**0.0300**	**0.9307**±0.0057	**0.9209**±0.0241	**0.0098**

TABLE IV
RUNNING TIME COMPARISON OF ALL 10 METHODS.

	DRO	CVaR	BPF	Reckoner	IRW	ARL	DORO	GraSP	GoG	IRWO
COMPAS (s)	3.77	2.23	2.18	55.02	4.91	6.14	4.76	>70.0	18.85	>70.0
CelebA (m/epoch)	2.20	1.38	1.36	15.80	3.50	5.70	3.25	>35.0	14.75	>35.0

Proof. The first inequality is an obvious conclusion of Eq. (5), and we show how the second inequality holds. Let $P_{N\alpha}$ be the distribution from which the examples belong to top $N\alpha$ largest losses can be drown. Then we have $P = \alpha P_{N\alpha} + (1-\alpha)P_{res}$, where P_{res} denotes the distribution from which all the rest of examples can be drawn.

$$\begin{aligned} D_{\chi^2}(P_{N\alpha}\|P) &= \int_z \left(\frac{P_{N\alpha}(z)}{P(z)} - 1\right)^2 P(z)dz \\ &\leq \int_z \left(\frac{1}{\alpha} - 1\right)^2 P(z)dz \\ &= r \end{aligned} \tag{15}$$

Eq. (15) showed that $P_{N\alpha} \in \mathcal{B}(P, r)$. Since the sup in $\mathcal{J}_{dr}(\theta; \alpha)$ is over all $Q \in \mathcal{B}(P, r)$ according to Eq. (4), the upper bound follows. □

Theorem 2. If the feasible set defined by C is convex, the optimal solution of Eq. (10) implies that a higher loss corresponds strictly to a higher weight.

Proof. For any two selected samples z_i and z_j with corresponding weights w_i and w_j obtained by solving Eq. (10), we prove $\ell(\theta; z_i) > \ell(\theta; z_i) \Rightarrow w_i > w_j$. Let $w = [..., w_i, ..., w_j, ...]$ be the optimal solution of Eq. (10), and we have following inequality

$$\sum_{k=1}^{N} w_k \ell(\theta; z_k) > w_j \ell(\theta; z_i) + w_i \ell(\theta; z_j) + \sum_{k \neq i,j}^{N} w_k \ell(\theta; z_k), \tag{16}$$

where we swap w_i and w_j on the RHS to satisfy the simplex constraint. By canceling the identical terms and simplifying the Eq. (16), we arrive at

$$(w_i - w_j)\left(\ell(\theta; z_i) - \ell(\theta; z_j)\right) > 0.$$

Then we get $w_i > w_j$, which completes the proof. □

Corollary 1. Supposing that reweighting and resampling takes the same sample importances, the gradient variance of resampling is not greater than reweighing if SGD is used as the optimizer.

Proof. Let P be the sampling proportion by which we get training samples. The conditional variance of reweighting under SGD can be written as

$$\begin{aligned} \sigma_w^2(\theta) &= \mathbb{E}_{z\sim P}[\nabla_\theta \ell(\theta; z)]^2 - \mathbb{E}_{z\sim P}^2[\nabla_\theta \ell(\theta; z)] \\ &= \sum_{i=1}^{N} N(w_i)^2 ||g_i||^2 - ||\frac{1}{N}\sum_{i=1}^{N} w_i g_i||^2 \end{aligned} \tag{17}$$

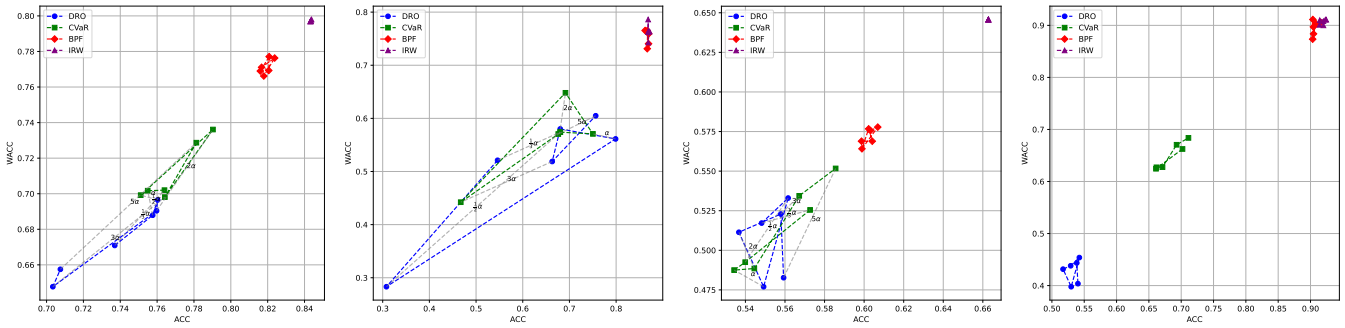
In resampling, one either repeats sampling some data points (i.e., oversampling) or removes some (i.e., undersampling) and finally gets N' samples [71]. We use Q to denote the sampling probability with which each observed data point participates in training. The conditional variance of resampling under SGD as

$$\begin{aligned} \sigma_s^2(\theta) &= \mathbb{E}_{z\sim Q}[\nabla_\theta \ell(\theta; z)]^2 - \mathbb{E}_{z\sim Q}^2[\nabla_\theta \ell(\theta; z)] \\ &= \frac{1}{N'}\sum_{i=1}^{N'} ||g_i||^2 - ||\frac{1}{N'}\sum_{i=1}^{N'} g_i||^2 \\ &= \sum_{i=1}^{N} w_i ||g_i||^2 - ||\frac{1}{N}\sum_{i=1}^{N} w_i g_i||^2 \end{aligned} \tag{18}$$

The comparison between Eqs. (17) and (18) indicates that $\sigma_w^2 \geq \sigma_s^2$. The variance of reweighting can be much larger if the importance distribution is far away from the non-uniform distribution. □

Impact of varying α. In practical scenarios, the estimated α collected from randomized response approach may not be as precise as expected. We have added performance results for various α values—$\frac{1}{3}\alpha$, $\frac{1}{2}\alpha$, α, 2α, 3α, and 5α—across all datasets. We did not include Reckoner in this group of experiments, as its formulation is not dependent on α. The results are presented in Fig. 7, where values corresponding to the same α are connected with dashed lines when none of two methods clearly dominates the others. We observe that the performance of DRO and CVaR varies significantly with changes in α, whereas our method, IRW, and the baseline BPF appear more robust.

Running time comparison. We evaluate the training time of all methods on the COMPAS and CelebA dataset. The average



(a) UCI Adult (b) Law School (c) COMPAS (d) CelebA

Fig. 7. Comparison of model performance for methods that incorporate α, evaluated over a range of α values.

wall-clock time across 10 trials is reported in Table IV, where the methods in the second column additional consider outliers problems. Based on the results, we make the following two observations. (1) Since our method requires two backward passes per batch update, its running cost is approximately twice that of CVaR and BPF. Reckoner has the highest running time, as it involves a complex teacher-student knowledge sharing framework. (2) For methods that handle outliers, GoG considers graph structure over individual gradients but in a same framework with ARL, and thus it leads to higher training cost. DORO is the most efficient among these methods, as it simply relies on a specified outlier ratio. Our IRWO adopts a similar strategy to GraSP, which requires clustering to identify outliers, resulting in relatively slower training. We acknowledge this as a limitation of this approach.

Group errors and individual errors. We present additional experimental results shown as Fig. 8 for Section VI-D. For clarity, we include only the sorted individual losses for DRO and BPF to better visualize the comparison. The three tabular datasets involve binary classification, where decision boundaries can be depicted—unlike CelebA, which is a multi-class classification task.

ACKNOWLEDGMENT

This work was supported in part by Maritime AI Research Programme (grant number: SMI-2022-MTP-06 funded by Singapore Maritime Institute) and in part by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative (No. DTC-RGC-04). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

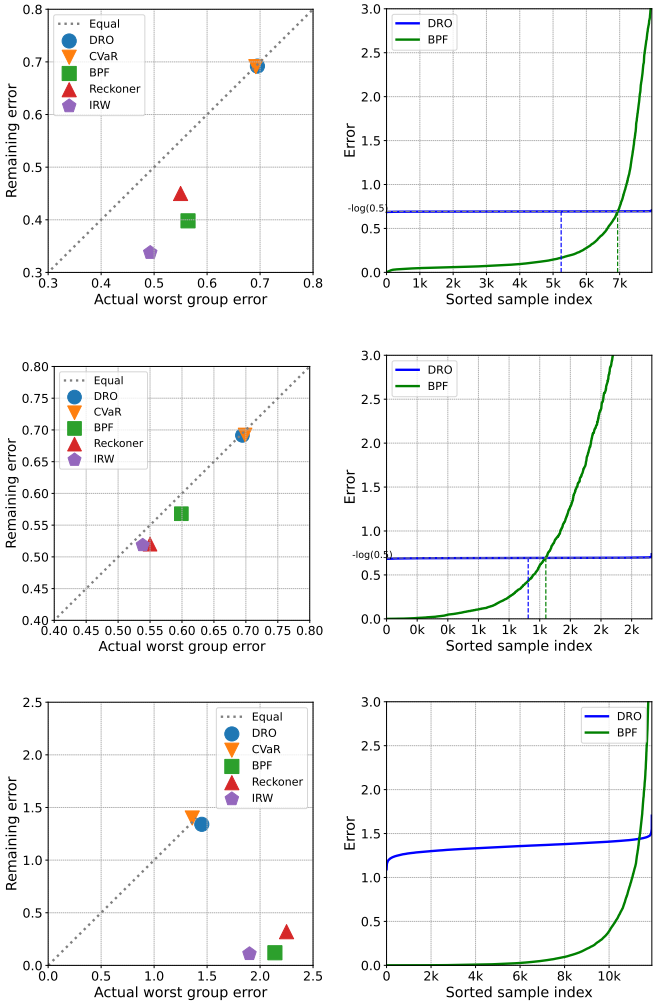


Fig. 8. Classification errors comparison. From top to bottom, each pair of subfigures corresponds to the results on a Law School, COMPAS and CelebA, respectively. Note that CelebA involves a multi-class classification which cannot provide a boundary for visualization.

REFERENCES

[1] E. Steel and J. Angwin, "On the web's cutting edge, anonymity in name only," *The Wall Street Journal*, vol. 4, 2010.

[2] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, p. eaao5580, 2018.

[3] J. L. Dieleman, C. Chen, S. W. Crosby, A. Liu, D. McCracken, I. A. Pollock, M. Sahu, G. Tsakalos, L. Dwyer-Lindgren, A. Haakenstad *et al.*, "Us health care spending by race and ethnicity, 2002-2016," *Jama*, vol. 326, no. 7, pp. 649–659, 2021.

[4] S. Nguyen, X. Fu, D. Ogawa, and Q. Zheng, "An application-oriented testing regime and multi-ship predictive modeling for vessel fuel consumption prediction," *Transportation Research Part E: Logistics and Transportation Review*, vol. 177, p. 103261, 2023.

[5] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis *et al.*, "Bias in data-driven artificial intelligence systems—an introductory

survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.

[6] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 962–970.

[7] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[8] J. M. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, ser. LIPIcs, C. H. Papadimitriou, Ed., vol. 67. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017, pp. 43:1–43:23.

[9] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International conference on machine learning*. PMLR, 2018, pp. 60–69.

[10] A. N. Carey and X. Wu, "The statistical fairness field guide: perspectives from social and formal sciences," *AI and Ethics*, vol. 3, no. 1, pp. 1–23, 2023.

[11] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[12] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[13] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[14] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig, "Controllable invariance through adversarial feature learning," *Advances in neural information processing systems*, vol. 30, 2017.

[15] Z. Zhang, S. Wang, and G. Meng, "A review on pre-processing methods for fairness in machine learning," in *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*. Springer, 2022, pp. 1185–1191.

[16] D. Guo, C. Wang, B. Wang, and H. Zha, "Learning fair representations via distance correlation minimization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 2139–2152, 2022.

[17] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[18] P. Putzel and S. Lee, "Blackbox post-processing for multiclass fairness," in *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (SafeAI 2022) co-located with the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI2022), Virtual, February, 2022*, G. Pedroza, J. Hernández-Orallo, X. C. Chen, X. Huang, H. Espinoza, M. Castillo-Effen, J. A. McDermid, R. Mallah, and S. Ó. hÉigeartaigh, Eds., vol. 3087, 2022.

[19] T. Jang, P. Shi, and X. Wang, "Group-aware threshold adaptation for fair classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6988–6995.

[20] J. Rawls, *Justice as fairness: A restatement*. Harvard University Press, 2001.

[21] N. L. Martinez, M. A. Bertran, A. Papadaki, M. Rodrigues, and G. Sapiro, "Blind pareto fairness and subgroup robustness," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7492–7501.

[22] X. Wang, J. Li, I. Tsang, and Y. S. Ong, "Towards harmless rawlsian fairness regardless of demographic prior," *Advances in Neural Information Processing Systems*, vol. 37, pp. 80 908–80 935, 2024.

[23] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[24] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala, "The price of fair pca: One extra dimension," *Advances in neural information processing systems*, vol. 31, 2018.

[25] H. Vu, T. Tran, M. Yue, and V. A. Nguyen, "Distributionally robust fair principal components via geodesic descents," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=9NVd-DMtThY

[26] J. Chi, Y. Tian, G. J. Gordon, and H. Zhao, "Understanding and mitigating accuracy disparity in regression," in *International conference on machine learning*. PMLR, 2021, pp. 1866–1876.

[27] M. Veale and R. Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data," *Big Data & Society*, vol. 4, no. 2, p. 2053951717743530, 2017.

[28] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–16.

[29] S. Yan, H.-t. Kao, and E. Ferrara, "Fair class balancing: Enhancing model fairness without observing sensitive attributes," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1715–1724.

[30] Z. Zhu, Y. Yao, J. Sun, H. Li, and Y. Liu, "Weak proxies are sufficient and preferable for fairness with missing sensitive attributes," in *International Conference on Machine Learning*. PMLR, 2023, pp. 43 258–43 288.

[31] V. Grari, S. Lamprier, and M. Detyniecki, "Fairness without the sensitive attribute via causal variational autoencoder," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L. D. Raedt, Ed. ijcai.org, 2022, pp. 696–702.

[32] L. Belli, K. Yee, U. Tantipongpipat, A. Gonzales, K. Lum, and M. Hardt, "County-level algorithmic audit of racial bias in twitter's home timeline," *arXiv preprint arXiv:2211.08667*, 2022.

[33] J. C. Duchi and H. Namkoong, "Learning models with uniform performance via distributionally robust optimization," *The Annals of Statistics*, vol. 49, no. 3, pp. 1378–1406, 2021.

[34] R. Zhai, C. Dan, Z. Kolter, and P. Ravikumar, "Doro: Distributional and outlier robust optimization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 345–12 355.

[35] Y. Zeng, K. Greenewald, K. Lee, J. Solomon, and M. Yurochkin, "Outlier-robust group inference via gradient space clustering," *arXiv preprint arXiv:2210.06759*, 2022.

[36] J. Chai and X. Wang, "Self-supervised fair representation learning without demographics," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 100–27 113, 2022.

[37] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[38] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*. PMLR, 2017, pp. 1885–1894.

[39] Y. Luo, Z. Li, Q. Liu, and J. Zhu, "Fairness without demographics through learning graph of gradients," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD 2025, Toronto, ON, Canada, August 3-7, 2025*, Y. Sun, F. Chierichetti, H. W. Lauw, C. Perlich, W. H. Tok, and A. Tomkins, Eds. ACM, 2025, pp. 918–926.

[40] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1929–1938.

[41] H. Heidari, M. Loi, K. P. Gummadi, and A. Krause, "A moral framework for understanding fair ml through economic models of equality of opportunity," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 181–190.

[42] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, "Fairness without demographics through adversarially reweighted learning," *Advances in neural information processing systems*, vol. 33, pp. 728–740, 2020.

[43] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annual review of statistics and its application*, vol. 8, no. 1, pp. 141–163, 2021.

[44] S. Gupta, V. Kovatchev, M. De-Arteaga, and M. Lease, "Fairly accurate: Optimizing accuracy parity in fair target-group detection," *arXiv preprint arXiv:2407.11933*, 2024.

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

[46] H. Zhao, A. Coston, T. Adel, and G. J. Gordon, "Conditional learning of fair representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=Hkekl0NFPr

[47] N. Martinez, M. Bertran, and G. Sapiro, "Minimax pareto fairness: A multi objective perspective," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6755–6764.

[48] B. Chaudhary, A. Pandey, D. Bhatt, and D. Tiwari, "Practical bias mitigation through proxy sensitive attribute label generation," *arXiv preprint arXiv:2312.15994*, 2023.

[49] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE international conference on data mining workshops*. IEEE, 2009, pp. 13–18.

[50] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 7, pp. 1445–1459, 2012.

[51] H. Wang, "Fairness metrics for recommender systems," in *icWCSN 2022: 9th international Conference on Wireless Communication and Sensor Networks, Dalian, China, January 11 - 13, 2022*. ACM, 2022, pp. 89–92. [Online]. Available: https://doi.org/10.1145/3514105.3514120

[52] V. Monjezi, A. Trivedi, V. Kreinovich, and S. Tizpaz-Niari, "Fairness testing through extreme value theory," *arXiv preprint arXiv:2501.11597*, 2025.

[53] B. Petrongolo, "The gender gap in employment and wages," *Nature Human Behaviour*, vol. 3, no. 4, pp. 316–318, 2019.

[54] H. Namkoong and J. C. Duchi, "Variance-based regularization with convex objectives," *Advances in neural information processing systems*, vol. 30, 2017.

[55] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, 2011.

[56] T. K. Nayak, "A review of rigorous randomized response methods for protecting respondent's privacy and data confidentiality," *Methodology and Applications of Statistics: A Volume in Honor of CR Rao on the Occasion of his 100th Birthday*, pp. 319–341, 2022.

[57] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International conference on machine learning*. PMLR, 2018, pp. 4334–4343.

[58] H. Ni, L. Han, T. Chen, S. Sadiq, and G. Demartini, "Fairness without sensitive attributes via knowledge sharing," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1897–1906.

[59] J. Chai, T. Jang, and X. Wang, "Fairness without demographics through knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 152–19 164, 2022.

[60] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.

[61] L. F. Wightman, "Lsac national longitudinal bar passage study. lsac research report series." 1998.

[62] M. Barenstein, "Propublica's compas data revisited," *arXiv preprint arXiv:1906.04711*, 2019.

[63] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[64] R. S. Chen, B. Lucier, Y. Singer, and V. Syrgkanis, "Robust optimization for non-convex objectives," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

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

[66] K. Kawaguchi, Z. Deng, X. Ji, and J. Huang, "How does information bottleneck help deep learning?" in *International Conference on Machine Learning*. PMLR, 2023, pp. 16 049–16 096.

[67] A. Soen and K. Sun, "On the variance of the fisher information for deep learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5708–5719, 2021.

[68] J. Li, Y. Pan, Y. Lyu, Y. Yao, Y. Sui, and I. W. Tsang, "Earning extra performance from restrictive feedbacks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11 753–11 765, 2023.

[69] T. Zhang, T. Zhu, J. Li, M. Han, W. Zhou, and S. Y. Philip, "Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 4, pp. 1763–1774, 2020.

[70] Y. Romano, S. Bates, and E. Candes, "Achieving equalized odds by resampling sensitive attributes," *Advances in neural information processing systems*, vol. 33, pp. 361–371, 2020.

[71] J. An, L. Ying, and Y. Zhu, "Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[72] H. Tian, B. Liu, T. Zhu, W. Zhou, and S. Y. Philip, "Multifair: Model fairness with multiple sensitive attributes," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

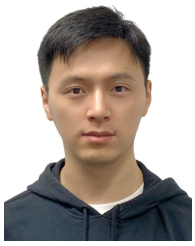
[73] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International conference on machine learning*. PMLR, 2018, pp. 2564–2572.

[74] G. Yu, X. Wang, C. Sun, Q. Wang, P. Yu, W. Ni, and R. P. Liu, "Ironforge: an open, secure, fair, decentralized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[75] Y. Shi, H. Yu, and C. Leung, "Towards fairness-aware federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

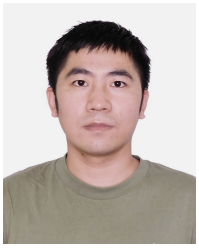
[76] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *International conference on machine learning*. PMLR, 2021, pp. 11 492–11 501.

[77] X.-X. Wei and H. Huang, "Balanced federated semisupervised learning with fairness-aware pseudo-labeling," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.



Jing Li is a Research Scientist at the Centre for Frontier AI Research (CFAR), A*STAR, Singapore. He received his Ph.D. in Information Technology from the University of Technology Sydney, Australia, in 2023, and his B.Eng. and M.Eng. degrees in Computer Science and Technology from Northwestern Polytechnical University, Xi'an, China, in 2015 and 2018, respectively. His current research focuses on trustworthy machine learning.

Yinghua Yao is a Research Scientist at the Centre for Frontier AI Research (CFAR), A*STAR, Singapore. She received a joint Ph.D. from the University of Technology Sydney (UTS) and the Southern University of Science and Technology (SUSTech), and a B.E. degree in Computer Science and Technology from SUSTech, Shenzhen, China, in 2018. Her research focuses on generative models.



Yuangang Pan is working as a Senior Research Scientist at A*STAR Centre for Frontier AI Research. He completed his Ph.D. degree in Computer Science in Mar 2020 from University of Technology Sydney (UTS), Australia. Before joining A*STAR, he was a postdoctoral research associate at the Australian Artificial Intelligence Institute at UTS. He has authored or co-authored papers in various top journals, such as JMLR, IEEE TPAMI, IEEE TNNLS, IEEE TKDE, and ACM TOIS. His research interests include Deep Clustering, Deep Generative Learning, and Robust Ranking Aggregation.

Xuanqian Wang received her M.S. degree in Computer Science and Technology from Beihang University (BUAA), Beijing, China. Her research interests include fairness, computer vision, and few-shot learning.

Ivor W. Tsang is the Director of A*STAR Centre for Frontier AI Research. He is an Adjunct Professor at School of Computer Science and Engineering (SCSE), Nanyang Technological University, Singapore. His research interests include transfer learning, deep generative models, learning with weakly supervision, Big Data analytics for data with extremely high dimensions in features, samples and labels. In 2013, he was the recipient of the ARC Future Fellowship for his outstanding research on Big Data analytics and large-scale machine learning. In 2019, his JMLR paper Towards ultrahigh dimensional feature selection for Big Data was the recipient of the International Consortium of Chinese Mathematicians Best Paper Award. In 2020, he was recognized as the AI 2000 AAAI/IJCAI Most Influential Scholar in Australia for his outstanding contributions to the field between 2009 and 2019. His research on transfer learning granted him the Best Student Paper Award at CVPR 2010 and the 2014 IEEE TMM Prize Paper Award. He is an IEEE Fellow, and serves on the Editorial Board for the JMLR, MLJ, JAIR, IEEE TPAMI, IEEE TAI, IEEE TBD, and IEEE TETCI. He serves/served as a AC or Senior AC for NeurIPS, ICML, AAAI and IJCAI, and the steering committee of ACML.

Xiuju Fu received the Ph.D. degree from Nanyang Technological University, Singapore, in 2003. She is currently a Senior Research Scientist and group manager with the Institute of High Performance Computing, A*STAR. She has been a Principal Investigator (PI) and co-PI of several grant projects. She has coauthored dozens of papers in conference proceedings and journals. Her research interests include complex network analytics, modeling and simulation, and data mining. In recent years, she is active in developing big data intelligence and modeling and simulation approaches for applications in maritime traffic safety and operation efficiency, supply chain and public health areas.