On the Degradation of Underrepresented Groups When Mitigating Bias

No Author Given

No Institute Given

Abstract. Machine learning models may exhibit suboptimal performance for specific groups due to the model design and the inherent data bias. Some fairness interventions are motivated by improving the performance of disadvantaged groups, with a trade-off in overall performance. However, in this paper, we find that applying such fairness interventions can lead to a deterioration in classifier performance across all demographic categories. Motivated by the skewness of the distribution, we introduce the *group skewness*, which indicates the inherent data bias, to characterize this phenomenon. We prove that the accuracy decreases within the skewed groups while fairness improves. We corroborate our analysis with extensive experiments on both real and synthetic datasets. Building on these insights, we propose that by collecting more features that are informative for disadvantaged groups, the group skewness can be mitigated without compromising the model performance.

Keywords: Algorithmic Fairness · Pareto Frontier · Machine Learning.

1 Introduction

Fairness has become a critical consideration in algorithmic decision-making, particularly in domains with significant real-world implications such as healthcare, recruitment, and criminal justice. Current machine learning models often demonstrate suboptimal performance for certain underprivileged subgroups, such as black and female subgroups, thus leading to unfairness. Ihe unfairness issue can be attributable to the model's design and inherent biases present in the data [34]. Thus, numerous approaches have been proposed to address bias and discrimination stemming from various sources to improve fairness [21, 33, 40].

An intuitive approach is to improve the performance of the underrepresented groups to ensure fairness, even at the cost of overall model performance. Many fairness interventions are driven, at least in part, by this idea [22, 37, 10, 9]. For example, some works [9, 22] seek to reweight the samples from the underrepresented for better treatment, while others [10] focus on rectifying imbalanced data distributions that disproportionately affect these groups.

However, we revisit this idea and find that this intuitive idea is flawed: interventions targeting the improvement of the underrepresented groups' performance are incapable of fully rectifying bias [43, 13, 35]. Instead, it suggests that fairness interventions may result in the degradation of performance across all

demographic groups, as the accuracy-fairness trade-off is almost inevitable [42, 36]. Although such interventions may improve fairness, they often do so at the cost of classifier performance. Besides, it is known that the attempts to balance fairness by compromising the performance of advantaged groups have faced considerable criticism [12, 8]. However, rather than merely reducing the performance of the advantaged groups, these interventions may result in the degradation of all groups, which is much more unbearable. We manifested this degradation of all groups with a toy example, as depicted in Figure 1. As we improve fairness from Point A to Point B, the error for the underrepresented groups decreases, and the error disparity reduces. However, further improvements in fairness from Point B to Point C lead to an increase in error for the underrepresented groups. For the blue curve below, though, there is no degradation of the underrepresented groups. Consequently, the central question here is:

How can we characterize the degradation of underrepresented groups when mitigating bias?

Drawing inspiration from the notion of skewness within distributions [17], we refer to the answer as the *group skewness*, which is initially introduced in [26]. Group skewness delineates the scenario where the optimal classifier for one demographic group consistently yields inferior performance compared to other groups, which indicates the inherent biases within the dataset. This concept not only aids in identifying skewed groups but also sheds light on the conditions leading to the performance degradation of the underrepresented groups. By characterizing the degradation, we can achieve several objectives: (i) recognizing the limitations of fairness interventions targeting underrepresented groups; (ii) understanding the limitations of fairness interventions when applied in decision-making; and (iii) informing the development of feature collection strategies that foster fairness in downstream tasks.

In this paper, we begin by demonstrating that the overall accuracy-fairness Pareto frontier observed in error plots across demographic groups is a subset of the subgroup accuracy-fairness Pareto frontier. The latter requires classifiers to maintain non-dominance in accuracy across all groups while preserving fairness. Additionally, we establish that accuracy experiences an initial rise followed by a decline within skewed groups as fairness improves, whereas accuracy within those groups continues to increase for group-balanced distributions. These findings are supported by extensive experiments conducted on various real-world datasets. Based on these insights, we propose that by collecting more features that are informative for underrepresented groups, the group skewness can be mitigated and the degradation can be eliminated without compromising the model performance.

Our contributions are as follows:

- **Characterizing degradation of the underrepresented groups.** We focus on the phenomenon where accuracy decreases of the underrepresented groups when mitigating bias, which indicates that several fairness interventions cannot fully correct bias.. Previous studies have touched upon this

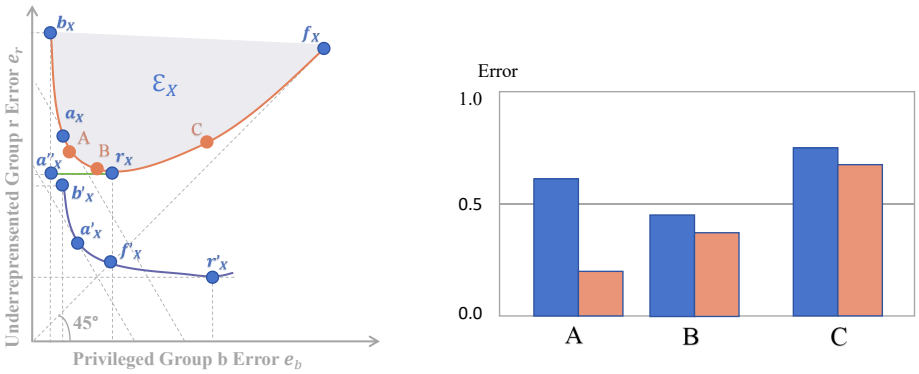

Fig. 1. Pareto Frontier We illustrate the relationships among the Pareto Frontier in this paper. In the figure on the left side, the shaded area represents all the error pairs achieved by all possible classifiers. b_X, r_X, a_X, and f_X (depicted as blue dots), as defined in this paper, represent the endpoints of the Pareto frontiers (orange curve, blue curve, and green curve). The orange curve is r-skewed, with both b_X and r_X located above the 45-degree curve. On this curve, subgroup r experiences degradation when improving fairness beyond r_X. The blue curve is group-balanced, with b_X and r_X positioned on opposite sides of the 45-degree curve, and the error for subgroup r continues to decrease as fairness improves. The green curve is also r-skewed; however, sensitive attributes (or highly related proxy sensitive attributes) are incorporated into the input, resulting in $r''_X = b''_X = a''_X$. Here, the error for subgroup r remains constant as fairness improves. Points A, B, and C lie on the Pareto Frontier. Applying fairness interventions to achieve optimal performance often leads to degradation in accuracy, as depicted in the figure on the right side.

topic, but empirically or intuitively [43, 13]. To our knowledge, we are the first to theoretically characterize this phenomenon and conducted extensive experiments to benchmark this phenomenon in the existing datasets.

– **Proposing a data-collection method to address this issue.** We propose a strategy to mitigate group skewness without compromising model performance by collecting features that are more informative for underrepresented groups. We prove that as the number of informative features for the underrepresented group goes to infinity, the degradation of the underrepresented groups can be eliminated.
– **Numerical experiments on real-world datasets.** To lend empirical support to our theoretical findings, we conduct extensive experiments on real-world datasets. Our results provide actionable insights for future research and practical implementations.

Remark 1. Our paper is inspired by a series of works in the field, including those by [16, 36, 28]. We center our analysis on the Pareto frontier, which delineates the optimal performance that a classifier can achieve. For a given data distribution, its Pareto frontier is fixed and represents the overall accuracy and fairness that no model can surpass. It is substantiated by the findings of [36]. In that pa-

per, the author demonstrates that current state-of-the-art fairness interventions are effective enough to mitigate bias. These interventions have propelled model performance to nearly reach the optimal performance on the Pareto frontier.

2 Related Work

Fair machine learning (FairML) methods have substantially developed in recent years [33, 40, 15, 24, 32, 39], and Research into the trade-off between fairness and accuracy has recently gained prominence [26, 13, 42, 30, 27]. Recent works focusing on the trade-off often have distinct emphases tailored to specific scenarios. For instance, [43] investigated the accuracy-fairness trade-off in computer vision, while [41] studied it in fair regression. Besides, some works have explored the trade-off under scenarios involving distribution shift and optimization with privacy concerns [38, 30, 27, 18]. Our work primarily aligns with theoretical analyses. Accordingly, we categorize existing approaches into three types: metric-based, data-based, and distribution-based.

Metric-based Approaches Metric-based approaches obtain general conclusions through mathematical analysis of the definitions of algorithmic fairness and accuracy. For example, [42] quantified the lower bound of unfairness by simply transforming the expression of the fairness definition. [6, 25, 13] discussed the incompatibility between different fairness criteria and concluded that all those fairness criteria cannot be achieved simultaneously. The maximization of both accuracy and fairness is also proved incompatible [19]. However, those papers did not provide a detailed analysis of the accuracy-fairness trade-off.

Data-based Approaches Progress in the field has witnessed substantial developments in data-based approaches, with a strong emphasis on leveraging observational data. [23] first examined the nuanced interplay between accuracy and fairness, crafting an optimal classifier contingent upon the proportion of instances. Meanwhile, [11] quantified unfairness via a comprehensive bias-variance decomposition, and [31] delved into the geometric analysis of this trade-off with discrete data sources. However, these works fall short in the existence of the Pareto frontier of the model performance, and the impact on different demographic groups remains largely unexplored.

Distribution-based Approaches These studies frequently assume that data distributions across different sensitive groups are known and accessible [42, 28, 7, 36]. [28] derived the decision boundary by intuitively using the true positive rate of the selected classifier on sensitive attributes to measure unfairness. [7] explored how fairness constraints on the training set can affect generalization performance when test set distributions differ. On the other hand, [16] assumed separate classifiers for different groups and derived the Chernoff bound to characterize the trade-off. However, these studies often focus more on the existence of the

accuracy-fairness trade-off under different fairness interventions rather than their impact across groups.

In our work, we adopt a distribution-based analysis approach and consider the Pareto frontier of all possible classifiers. However, unlike prior works, our paper presents two distinctive features. Firstly, we focus on the phenomenon where accuracy decreases across all demographic groups as fairness improves, a nuanced observation compared to conventional trade-off analyses. We are the first, to the best of our knowledge, to theoretically characterize this phenomenon, indicating that several fairness interventions cannot fully correct bias. Moreover, we propose a data-collection method to address this issue with a theoretical guarantee. The most related work to us is [26]. However, unlike their research, we delve into the overall accuracy-fairness Pareto frontier rather than the subgroup accuracy-fairness Pareto frontier, which requires classifiers to maintain non-dominance in accuracy across all groups while preserving fairness. Therefore, our analysis accommodates various fairness interventions. Besides, we empirically characterize this phenomenon on real-world datasets, which provides insights for future research.

3 Preliminaries

In this section, we introduce notation and discuss several types of Pareto frontiers relevant to our study.

3.1 Notation

Each individual is characterized by a covariate vector X drawn from the set $\mathcal{X}$, a label Y from the finite set $\mathcal{Y}$, and a sensitive attribute A from the finite set $\mathcal{A}$. The predicted value $\hat{Y}$ also belongs to $\mathcal{Y}$. In our investigation, we focus on binary classification with binary sensitive attributes. However, our analysis can be generalized to accommodate multiple sensitive attributes, which we defer to the future work. Consequently, $\mathcal{Y} = \{0, 1\}$ and $\mathcal{A} = \{0, 1\}$. For example, in the context of job hiring, $Y = 1$ denotes a high-quality job applicant, $A = 1$ denotes a male applicant, and $\hat{Y} = 1$ denotes that the applicant is hired. The vector X encompasses non-sensitive attributes such as resumes and references.

We evaluate our model value here using the loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where we employ $\ell(\hat{y}, y) = \mathbf{1}(\hat{y} \neq y)$ throughout this paper. We can further extend this definition to individuals within each group:

Definition 1. *For any algorithm $a \in \mathcal{A}_X$ and group $A \in \mathcal{A}$, the group A error is:*

$$e_A(a) := \mathbb{E}_{D \sim a(X)}[\ell(D, Y) \mid A]. \tag{1}$$

While other loss functions may assign different weights to subgroup errors, our results hold for arbitrary loss under certain conditions, as demonstrated in [26]. In the following sections, we denote $e_g(a)$ as the *subgroup error*, and $e(a)$ as the *overall error*. We adopt *Overall Accuracy Equality* as the fairness criterion and quantify unfairness as follows:

Definition 2. *For any algorithm $a \in \mathcal{A}_X$ and group $A \in \mathcal{A}$, the group A error is:*

$$U(a) = |e_{A=1}(a) - e_{A=0}(a)| \tag{2}$$

We note that our results can be partially extended to other fairness criteria, such as *Equalized Odds* and *Statistical Parity*. A comprehensive examination of these criteria and experimental results will be presented in Section 4.

3.2 Pareto Frontier

People always prefer algorithms with higher accuracy and fairness. Hence, we characterize the set of preferences through the *overall accuracy-fairness Pareto dominance*, and define the *overall accuracy-fairness frontier* given fixed Xas :

Definition 3. *For any algorithm a and b, we say a overall accuracy-fairness dominates b satisfying $e(a) \leq e(b)$ and $U(a) \leq U(b)$, with at least one inequality strict. This dominance relationship is denoted as $a > FAob$. The overall accuracy-fairness Pareto frontier (FA_o) given X is defined as:*

$$FA_o \equiv \{e(a) : no\ a' \in \mathcal{A}_{\mathcal{X}}\ such\ that\ a' >_{FA_o} a\} \tag{3}$$

Similarly, we establish the concepts of *subgroup accuracy dominance* and the *subgroup accuracy frontier*:

Definition 4. *For any algorithm a and b, we say a subgroup accuracy dominates b satisfying $e_{A=1}(a) \leq e_{A=1}(b)$ and $e_{A=0}(a) \leq e_{A=0}(b)$, with at least one inequality strict. The dominance is denoted as $a >_{A_g} b$. The subgroup accuracy Pareto frontier (A_g) given X is*

$$A_g \equiv \{e(a) : no\ a' \in \mathcal{A}_{\mathcal{X}}\ such\ that\ a' >_{A_g} a\} \tag{4}$$

The distinction between these two frontiers is clear: FA_o represents the target for which fairness interventions are applied, while A_g elucidates the optimal trade-off between different group errors across all potential classifiers. The latter aids in revealing potential degradation for the underrepresented groups. In this paper, we aim to elucidate the relationship between these two frontiers. Additionally, we define the concept of the *underrepresented group* within our context:

Definition 5. *Group r is the underrepresented group when $e_r \geq e_b$ at a_X where a_X represents the optimal performance for the whole data distribution that $a_X \equiv \underset{(e_r, e_b)}{\text{argmin}}\ e$*

It is noted that the notion of the *underrepresented group* differs from the commonly used concept of the *minority group*, which often carries ethical or political connotations. In this paper, the designation of the underrepresented group is solely based on computational analysis, devoid of ethical or political implications. For instance, while females may be recognized as the minority group, males could be identified as the underrepresented group in the UCI-Adult dataset when accuracy parity serves as the fairness criterion.

4 Potential Degradation of Underrepresented Groups

In this section, we characterize the potential degradation of underrepresented groups when improving fairness. Firstly, we introduce the key result of group skewness introduced in [26]. We then investigate the connections among various types of frontiers and deduce the prerequisites for the potential degradation of underrepresented groups. The extension to other fairness criteria will be briefly discussed.

4.1 Subgroup Accuracy-Fairness Pareto Frontier

To bridge the gap between FA_o and A_g, we aim to identify their shared characteristics. However, before that, we define the *feasible set* of subgroup error pairs as those pairs representing the evaluations of certain algorithms, formulated as $\mathcal{E}X \equiv (eA = 1(a), e_{A=0}(a)) : a \in \mathcal{A}_X$. It is easy to recognize that the subgroup accuracy frontier A_g comprises all subgroup error pairs that remain undominated by group accuracy within the feasible set. Subsequently, we introduce the *subgroup accuracy-fairness Pareto dominance* and the *subgroup accuracy-fairness frontier* as follows:

Definition 6. *For any algorithm a and b, we say a subgroup accuracy-fairness Pareto dominates b satisfying $e_{A=1}(a) \leq e_{A=1}(b)$, $e_{A=0}(a) \leq e_{A=0}(b)$ and $U(a) \leq U(b)$, with at least one inequality strict. The dominance is denoted as $a >_{FA_g} b$. The subgroup accuracy-fairness frontier (FA_g) given X is:*

$$FA_g \equiv \{(e_{A=1}(a), e_{A=0}) \in \mathcal{E}_X : no\ a' \in \mathcal{A}_{\mathcal{X}}\ such\ that\ a' >_{FA_g} a\} \tag{5}$$

We proceed to define the group skewness of the data distribution, as presented in [26]:

Definition 7. *Given the Covariate vector X, we characterize the data distribution as:*

- *r-skewed if $e_r < e_b$ at r_X and $e_r \leq e_b$ at b_X*
- *b-skewed if $e_b < e_r$ at b_X and $e_b \leq e_r$ at r_X*
- *group-balanced otherwise*

where $\{r, b\}$ represents the sensitive attributes. $r_X \equiv \underset{(e_r,e_b)}{\operatorname{argmin}}\ e_r$, $b_X \equiv \underset{(e_r,e_b)}{\operatorname{argmin}}\ e_b$, and $(e_r, e_b) \in \mathcal{E}_X$

r_X and b_X represent the error pair from the optimal Bayes classifier for each subgroup, respectively, Computing them poses an NP-hard problem; however, we can approximate them by training classifiers for each demographic group. One of the key findings by [26] is to discover the relation between the group skewness and the shape of the subgroup accuracy-fairness frontier FA_g, formally stated as follows:

Theorem 1. *The subgroup accuracy-fairness frontier FA_g is the lower boundary of the feasible set $\mathcal{E}_X$ between*

- *r_X and b_X if X is group-balanced*
- *g_X and f_X if X is g-skewed*

where the lower boundary between two points represents this part of the set lies between the two points and the lower curve connecting the two, and f_X represents the error pair of the fairest classifier that $f_X \equiv \underset{(e_r,e_b)}{\operatorname{argmin}}\, U(a)$.

Full proof can be seen in the Appendix. This theorem indicates that inherent discrimination in the dataset shapes the subgroup accuracy-fairness frontier. It shows that on the frontier FA_g, the skewed subgroup error declines at first, followed by a rise as the fairness improves. However, when the data distribution is group-balanced, the subgroup error continuously declines as the fairness improves. Given that the Pareto frontier is a property of the data distribution, it motivates us to propose the guidelines for downstream feature collection to push the frontier in the following section.

4.2 Degradation of the Underrepresented Groups

To characterize the potential degradation of the underrepresented groups, understanding the connections among various Pareto frontiers is pivotal. In the following lemma, we provide the relations among those Pareto frontiers.

Lemma 1. *The overall accuracy-fairness Pareto frontier FA_o and the subgroup accuracy Pareto frontier A_g are subsets of the subgroup accuracy-fairness Pareto frontier FA_g that $FA_o \subset FA_g, A_g \subset FA_g$.*

This lemma can be readily derived from their respective definitions. We demonstrate the connections using a simplified example, as depicted in Figure 1. We assume a distribution that is r-skewed, with r_X positioned to the lower right of b_X. The curve between b_X and r_X signifies the subgroup accuracy Pareto frontier A_g, while the curve between a_X and f_X represents the overall accuracy-fairness Pareto frontier FA_o. It is evident that both of these frontiers are subsets of the subgroup accuracy-fairness Pareto frontier, which is delineated by the curve between b_X and f_X.

Now we characterize the potential degradation of the underrepresented groups. In Figure 1, we can find that the error for the subgroup r decreases at first, but increases as fairness continues improving.

Theorem 2. *The overall accuracy-fairness frontier FA_o is the lower boundary of the feasible set $\mathcal{E}_X$ between*

- *When improving fairness, the subgroup r error does not increase initially. Upon reaching the minimum, the error does not decrease. The subgroup g error continuously increases if and only if X is r-skewed.*

– *When improving fairness, the subgroup r error non-increases and the subgroup b error increases if and only if X is b-skewed or group-balanced.*

non-decreasing is synonymous with increasing, except when $r_X = b_X$. This theorem provides a comprehensive understanding of the potential degradation of the underrepresented groups. For instance, in the UCI-Adult dataset where males are underrepresented and the dataset is male-skewed, the error for the male group decreases at first and then increases when improving fairness. Conversely, in the COMPAS dataset where females are underrepresented and the dataset is group-balanced, the error for the female group continuously decreases as fairness is enhanced. Since the Pareto frontier remains fixed given the data distribution, it prompts us to explore the features that affect the prediction, which serves as motivation to propose a feature collection method to eliminate the degradation of the underrepresented groups.

Remark 2. For alternative fairness criteria, such as Equalized Odds and Demographic Parity, adjustments to the measurement of unfairness are necessary, thus the subgroup accuracy-fairness Pareto frontier FA_g cannot be simply manifested by a curve. Consequently, our findings can only be partially extrapolated to these scenarios. For instance, we extend the notion of group balance to generalized group balance, wherein $F_X \subset FA_g$, with F_X representing the fairness-optimal set defined as $F_X = e \in \mathcal{E}_X : e = \underset{(e_r, e_b)}{\operatorname{argmin}} U(a)$. We can establish that there is no degradation across groups under this condition. While our findings may be constrained in the scope of generalized group balance, we conducted extensive experiments involving diverse fairness constraints on real-world datasets. These experiments demonstrate that the degradation of underrepresented groups persists as fairness improves.

Remark 3. As previously discussed in the first section 1, fairness interventions aimed at enhancing the performance of the most disadvantaged groups may not fully rectify bias. Particularly, in the case of reweighting methods, which adjust the weights of demographic groups during training, continuously improving fairness on the Pareto frontier is infeasible. This limitation arises due to the necessity for non-negative weights for privileged groups. For instance, considering a data distribution skewed towards subgroup r, as illustrated in Figure 1, the classifier at r_X mirrors the same Bayes optimal classifier for the underrepresented group. To achieve fairness beyond r_X, it becomes necessary to introduce negative weights for subgroup b, which is often impossible for those methods. Consequently, these reweighting techniques fall short of achieving complete fairness on the Pareto frontier.

5 Eliminate the Degradation through Informative Feature Collection

In this section, we aim to address the issue of potential degradation of the underrepresented groups by collecting features that are more informative for those

subgroups. Conceptually, we can liken the input data to an information source and the models to channels. By collecting more features, we can enhance the performance of the new optimal Bayes classifier compared to its predecessor, thereby pushing the Pareto frontier without any compromise [14]. Our focus here is on devising data collection methodologies that collect more informative features for the underrepresented groups, with the overarching goal of mitigating their degradation while maintaining overall accuracy.

Before delving into our proposed methodology, we first establish a formal definition of the informative feature:

Definition 8. *Feature S is informative for group r , i.e. for the subgroup error e_r at r_X and e'_r at $r_{X\cup S}$, $e_r > e'_r$. Feature S is more informative for group r than group b, i.e. for the subgroup error (e_r, e_b) at r_X and (e'_r, e'_b) at $r_{X\cup S}$, $e_r - e'_r < e_b - e'_b$.*

Informative features play a pivotal role in classification tasks, often exhibiting disparate impact on different subgroups. For instance, as discussed in [29], the healthcare costs are more informative for White patients than for Black patients when predicting their healthcare needs. The reason for the observed degradation in performance among underrepresented groups is that their optimal Bayes classifier consistently outputs inferior performance, compared to other subgroups. Hence we propose to collect more informative features for the underrepresented groups.

Claim. As the number of informative features for the underrepresented subgroup approaches infinity, the degradation of the underrepresented groups can be mitigated without compromise.

The proof is provided in the appendix. This claim can be readily understood by referring to Figure 1. It shows that subgroup r is the underrepresented group. When we collect more informative features for subgroup r, the error for both the subgroup r decreases. However, the rate of decrease in e_r exceeds that of the 45-degree line, indicating swifter mitigation of degradation while addressing bias. Collecting sensitive attributes (or features related to sensitive attributes) can also aid in eliminating degradation, though, the distribution remains skewed toward the group, with $r_X = b_X = a_X$. In Figure 1, a''_X denotes the error pairs corresponding to the optimal performance achieved after incorporating sensitive attributes as inputs. The green curve illustrates that the error for the underrepresented group r remains constant as we improve fairness. However, given that the utilization of sensitive attributes is prohibited by law [3], this approach may not be applicable in practical scenarios.

6 Experiments

In this section, we illustrate that existing real-world datasets may exhibit either group-skewed or group-balanced characteristics, and to elucidate how group

skewness correlates with degradation of the underrepresented groups. Our analysis reveals that the UCI-Adult dataset demonstrates male-skewed and white-skewed characteristics, with a notable increase in error for the male or black group under higher fairness requirements. Conversely, the COMPAS dataset exhibits group-balanced properties, with error decreasing for underrepresented groups as fairness improves. Furthermore, we enhance the datasets by collecting informative features for the underrepresented group, subsequently, mitigating the skewness and eliminating the degradation.

6.1 Potential Degradation of the Underrepresented Groups

Setup We conducted experiments utilizing the UCI-Adult dataset [4], the COMPAS dataset [2], and the German Credit dataset [20], as these datasets are widely used in fairness interventions and have readily available codes. We employed the Reduction fairness intervention method [1], implemented using the IBM AIF360 library [5]. Fairness requirements (2) were deliberately violated at various levels by adjusting the tolerance within the IBM AIF360 library. A random forest served as the baseline classifier, although the choice of machine learning models did not significantly impact the results. To mitigate error variance, we conducted 10 iterations for each violation and computed the expected errors for each subgroup at different levels. Additionally, we explore degradation under other fairness constraints and we choose *Demographic Parity* and *Equalized Odds* as the fairness criteria. The measurement of unfairness and the experimental setup align with those outlined in [36]. Further details regarding data preprocessing, model training, and additional experimental findings are provided in the Appendix.

Results The results can be seen in Figure 2, 3. We trained a classifier to approximate the Bayes optimal classifier for each subgroup and plot the Pareto frontier of all error pairs. The privileged group here is male or white, with the sensitive attributes being gender or race, respectively. Our result shows that the UCI-Adult dataset is male-skewed and white-skewed, while the COMPAS dataset is group-balanced. We then utilized the baseline classifier for Reduction and applied it to the training set, evaluating its performance on the test set. We then present our results on the UCI-Adult dataset and the COMPAS dataset with race as the sensitive attribute in the main text. Further details regarding the sensitive attribute gender and the German Credit dataset are provided. For the UCI-Adult dataset, the error for the white group increases as fairness improves, while for the COMPAS dataset, the error for the underrepresented group decreases. It is worth noting, as mentioned in [36], that state-of-the-art fairness interventions have achieved nearly optimal fairness-accuracy curves. Therefore, it can be inferred that while enhancing fairness, the UCI-Adult dataset suffers degradation in the underrepresented groups, whereas COMPAS does not. Interestingly, our results demonstrate the degradation of the underrepresented groups in both the UCI-Adult dataset and the COMPAS dataset under other fairness constraints. This highlights the need for further exploration of the conditions contributing to this degradation.

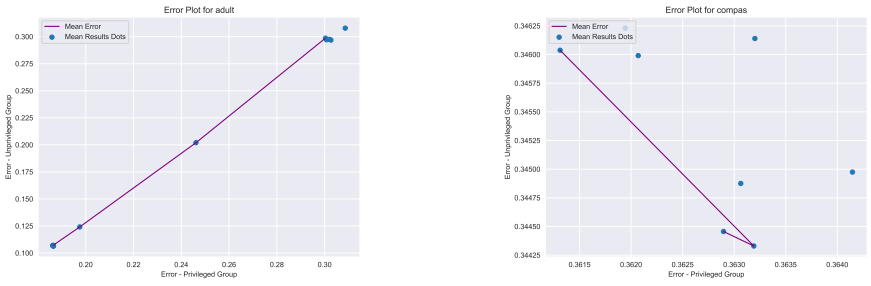

Fig. 2. Error plot of the UCI-Adult dataset and the COMPAS dataset, under the criteria of Overall Accuracy Parity.

In the following subsection, we will demonstrate how group skewness can be mitigated and degradation can be eliminated by incorporating more informative features for underrepresented groups.

6.2 Collecting Informative Features for Underrepresented Groups

Setup We conducted experiments on the COMPAS dataset, employing existing fairness interventions. Initially, to construct the skewed datasets, we selected ['priors_count', 'stay_discrete'] as the base features. Subsequently, we augmented these features with ['c_charge_degree', 'priors_count'] as informative features for underrepresented groups. A random forest served as the baseline classifier, with n_estimators set to 17. We conducted 10 iterations and computed expectations to mitigate error variance. Further details regarding data preprocessing, model training, and additional experimental findings are provided in the Appendix.

Results With the initial features, the distribution is male-skewed and white-skewed, with the degradation of the underrepresented groups. Upon inclusion of informative features tailored for underrepresented groups, the skewness of the distribution is mitigated. Consequently, the degradation observed in the underrepresented groups diminishes, with an overall reduction in prediction error. This underscores the efficacy of our feature collection approach without compromising model performance.

7 Conclusions

Fairness has emerged as a critical consideration in algorithmic decision-making, particularly in domains such as healthcare and criminal justice. The inherent bias present in machine learning models often results in suboptimal performance for

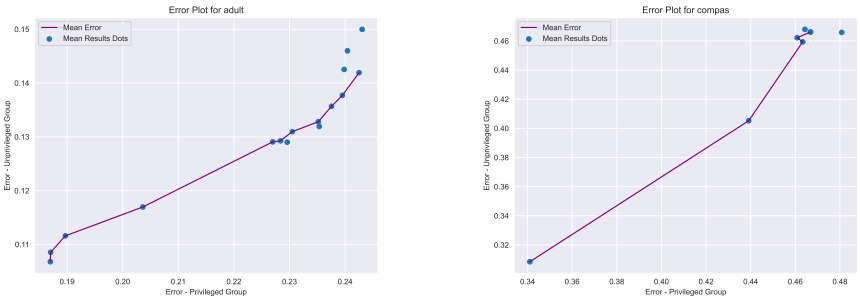

Fig. 3. Error plot of the UCI-Adult dataset and the COMPAS dataset, under the criteria of Equalized Odds.

certain demographic groups, highlighting the necessity for interventions to enhance fairness. While numerous approaches strive to bolster the performance of the most disadvantaged groups to promote fairness, our work reveals a fundamental limitation in this intuitive strategy. Despite efforts to enhance fairness by targeting disadvantaged groups, our findings demonstrate that such methods cannot entirely rectify bias. We introduce the concept of group skewness, which delineates situations where the optimal classifier for one subgroup still yields inferior performance compared to other groups. Moreover, we establish that under certain conditions, accuracy may decrease across all demographic groups as fairness improves. We propose a novel strategy to mitigate group skewness by collecting features that are more informative for underrepresented groups. Through empirical experiments conducted on various real-world datasets, we validate our findings and provide insights for future research endeavors. Our theoretical framework and empirical findings offer valuable insights for policymakers seeking to promote fairness while preserving the performance of underrepresented groups.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification (2018)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine Bias (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Barocas, S., Selbst, A.D.: Big data's disparate impact. California Law Review **104**(3), 671–732 (2016). https://doi.org/10.15779/z38bg31, <Go to ISI>://WOS:000384887500002
4. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996), DOI: https://doi.org/10.24432/C5XW20
5. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards,

J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias (Oct 2018), https://arxiv.org/abs/1810.01943
6. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**(1), 3–44 (2021). https://doi.org/10.1177/0049124118782533, https://doi.org/10.1177/0049124118782533
7. Blum, A., Stangl, K.: Recovering from biased data: Can fairness constraints improve accuracy? CoRR **abs/1912.01094** (2019), http://arxiv.org/abs/1912.01094
8. Brown, C.: Giving up levelling down. Economics & Philosophy **19**(1), 111–134 (2003)
9. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: 2009 IEEE International Conference on Data Mining Workshops. pp. 13–18 (2009). https://doi.org/10.1109/ICDMW.2009.83
10. Chakraborty, J., Majumder, S., Menzies, T.: Bias in machine learning software: Why? how? what to do? In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. p. 429–440. ESEC/FSE 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3468264.3468537, https://doi.org/10.1145/3468264.3468537
11. Chen, I.Y., Johansson, F.D., Sontag, D.: Why is my classifier discriminatory? In: 32nd Conference on Neural Information Processing Systems (NIPS). Advances in Neural Information Processing Systems, vol. 31 (2018), <Go to ISI>://WOS:000461823303053
12. Christiano, T., Braynen, W.: Inequality, injustice and levelling down. Ratio **21**(4), 392–420 (2008)
13. Corbett-Davies, S., Gaebler, J.D., Nilforoshan, H., Shroff, R., Goel, S.: The measure and mismeasure of fairness. The Journal of Machine Learning Research **24**(1), 14730–14846 (2023)
14. Cover, T.M., Thomas, J.A.: Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, USA (2006)
15. Deng, Z., Zhang, J., Zhang, L., Ye, T., Coley, Y., Su, W.J., Zou, J.: Fifa: Making fairness more generalizable in classifiers trained on imbalanced data. arXiv preprint arXiv:2206.02792 (2022)
16. Dutta, S., Wei, D., Yueksel, H., Chen, P.Y., Liu, S., Varshney, K.R.: Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org (2020)
17. Groeneveld, R.A., Meeden, G.: Measuring skewness and kurtosis. Journal of the Royal Statistical Society. Series D (The Statistician) **33**(4), 391–399 (1984), http://www.jstor.org/stable/2987742
18. Gultchin, L., Cohen-Addad, V., Giffard-Roisin, S., Kanade, V., Mallmann-Trenn, F.: Beyond impossibility: Balancing sufficiency, separation and accuracy. arXiv preprint arXiv:2205.12327 (2022)
19. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 3323–3331. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
20. Hofmann, H.: Statlog (German Credit Data). UCI Machine Learning Repository (1994), DOI: https://doi.org/10.24432/C5NC77

21. Hort, M., Chen, Z., Zhang, J.M., Harman, M., Sarro, F.: Bias mitigation for machine learning classifiers: A comprehensive survey. ACM Journal on Responsible Computing (2023)
22. Jiang, H., Nachum, O.: Identifying and correcting label bias in machine learning (2019)
23. Kamiran, F., Calders, T.: Data pre-processing techniques for classification without discrimination. Knowledge and Information Systems **33** (10 2011). https://doi.org/10.1007/s10115-011-0463-8
24. Kang, M., Li, L., Weber, M., Liu, Y., Zhang, C., Li, B.: Certifying some distributional fairness with subpopulation decomposition. Advances in Neural Information Processing Systems **35**, 31045–31058 (2022)
25. Kusner, M.J., Loftus, J.R., Russell, C., Silva, R.: Counterfactual fairness (2018)
26. Liang, A., Lu, J., Mu, X.: Algorithm design: A fairness-accuracy frontier. arXiv preprint arXiv:2112.09975 (2021)
27. Lowy, A.: Differentially Private and Fair Optimization for Machine Learning: Tight Error Bounds and Efficient Algorithms. Ph.D. thesis, University of Southern California (2023)
28. Menon, A.K., Williamson, R.C.: The cost of fairness in binary classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 107–118. PMLR (23–24 Feb 2018), https://proceedings.mlr.press/v81/menon18a.html
29. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
30. Pham, T.H., Zhang, X., Zhang, P.: Fairness and accuracy under domain generalization. arXiv preprint arXiv:2301.13323 (2023)
31. Pinzón, C., Palamidessi, C., Piantanida, P., Valencia, F.: On the impossibility of non-trivial accuracy in presence of fairness constraints. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conferencere on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. pp. 7993–8000. AAAI Press (2022), https://ojs.aaai.org/index.php/AAAI/article/view/20770
32. Qi, T., Wu, F., Wu, C., Lyu, L., Xu, T., Liao, H., Yang, Z., Huang, Y., Xie, X.: Fairvfl: A fair vertical federated learning framework with contrastive adversarial learning. Advances in Neural Information Processing Systems **35**, 7852–7865 (2022)
33. Shui, C., Xu, G., Chen, Q., Li, J., Ling, C.X., Arbel, T., Wang, B., Gagné, C.: On learning fairness and accuracy on multiple subgroups. Advances in Neural Information Processing Systems **35**, 34121–34135 (2022)
34. Singh, A., Kempe, D., Joachims, T.: Fairness in ranking under uncertainty. Advances in Neural Information Processing Systems **34**, 11896–11908 (2021)
35. Sühr, T., Hilgard, S., Lakkaraju, H.: Does fair ranking improve minority outcomes? understanding the interplay of human and algorithmic biases in online hiring. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 989–999 (2021)
36. Wang, H., He, L., Gao, R., Calmon, F.P.: Aleatoric and epistemic discrimination in classification. arXiv preprint arXiv:2301.11781 (2023)
37. Wang, H., Ustun, B., Calmon, F.P.: Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine

Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 6618–6627. PMLR (2019), http://proceedings.mlr.press/v97/wang19l.html
38. Wick, M., Tristan, J.B., et al.: Unlocking fairness: a trade-off revisited. Advances in neural information processing systems **32** (2019)
39. Zhang, F., Kuang, K., Chen, L., Liu, Y., Wu, C., Xiao, J.: Fairness-aware contrastive learning with partially annotated sensitive attributes. In: The Eleventh International Conference on Learning Representations (2022)
40. Zhang, G., Zhang, Y., Zhang, Y., Fan, W., Li, Q., Liu, S., Chang, S.: Fairness reprogramming. Advances in Neural Information Processing Systems **35**, 34347–34362 (2022)
41. Zhao, H.: Costs and benefits of fair regression (2021)
42. Zhao, H., Gordon, G.J.: Inherent tradeoffs in learning fair representations (2022)
43. Zietlow, D., Lohaus, M., Balakrishnan, G., Kleindessner, M., Locatello, F., Schölkopf, B., Russell, C.: Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers (2022)

1 Omitted Proofs

1.1 Proof of Theorem 1

As illustrated in Figure 1, this theorem naturally deduced when the lower boundary of the feasible set $\mathcal{E}_X$ is convex: With the random-guess classifier, which outputs results following a uniform distribution across all possible outputs, we can achieve accuracy parity. Consequently, according to its definition, the subgroup accuracy-fairness Pareto Frontier FA_g will always intersect with the 45-degree line. Therefore, the theorem only stipulates two conditions that are stated in the main text.

Now we prove that the feasible set $\mathcal{E}_X$ is convex. We pick two classifiers $a_1, a_2 \in \mathcal{E}_X$. For both subgroups r and b, we obtain the flattened confusion matrix that:

$$C(a) \equiv \begin{pmatrix} \mathrm{TPR}_{A=r}(a), \mathrm{FPR}_{A=r}(a), \mathrm{TNR}_{A=r}(a), \mathrm{FNR}_{A=r}(a) \\ \mathrm{TPR}_{A=b}(a), \mathrm{FPR}_{A=b}(a), \mathrm{TNR}_{A=b}(a), \mathrm{FNR}_{A=b}(a) \end{pmatrix} \in [0,1]^8$$

. Therefore, we can obtain that: $\lambda * C(a_1) + \lambda * C(a_2) = C(\lambda a_1) + C((1-\lambda)a_2)$

Here we define a new classifier $a = \lambda a_1 + (1-\lambda)a_2$, with λ as the probability of implementing the classifier a_1. This classifier is well-defined since $C(\lambda a_1 + (1-\lambda)a_2)$ is fixed as the number of samples goes to infinity. Because the feasible set $\mathcal{E}_X$ contains all possible classifiers implemented on X, there exists a classifier a such that $a = \lambda a_1 + (1-\lambda)a_2$. Then we can split the expectation that: $C(a) = C(\lambda a_1 + (1-\lambda)a_2) = C(\lambda a_1) + C((1-\lambda)a_2)$.

Therefore, flattened confusion matrix set $\mathcal{C} = \{C(a) \mid a \in \mathcal{A}\}$ is convex. Since the error for each sensitive group can be computed by the linear combination of the elements from the $\mathcal{C}$. Since the new convex set is still convex after affine transformation, the feasible set $\mathcal{E}_X$ is convex.

1.2 Proof of Lemma 1

Indeed, since unfairness can be straightforwardly measured by the distance between the error pair and the 45-degree line, this lemma can be directly derived from the definition.

1.3 Proof of Theorem 2

We begin by proving that the error pair e for the optimal overall performance must lie on the curve between r_X and b_X. Let's denote the optimal accuracy as Acc_{opt}. If e resides outside the curve, then $Acc_{opt} < \min\{Acc(r_X), Acc(b_X)\}$, which contradicts its definition.

Therefore, the flattened confusion matrix set $\mathcal{C} = \{C(a) \mid a \in \mathcal{A}\}$ is convex. The error for each subgroup can be computed by a linear combination of the elements from $\mathcal{C}$. Additionally, since the new set remains convex after affine transformation, the feasible set $\mathcal{E}_X$ is convex.

1.4 Proof of the claim

According to the definition of informative features, through iteratively collecting features, we can continuously improve the optimal performance for the under-represented group. As we can always ignore the new feature and still take the original model for prediction, the set of the original classifier set is the subset of the set of classifiers with the newly added features. Therefore, any classifier on the new Pareto frontier cannot be overall accuracy-fairness dominated by the classifier on the original Pareto frontier. Therefore, by collecting more features, we can improve the performance of the new optimal Bayes classifier without any compromise [14],

2 Details on the Experimental Results

2.1 Data Preprocessing

We mostly follow the data preprocessing implemented in [36]. The main difference between our setups is that we don't include the sensitive attributes as our input, since it causes the unwilling skewness without degradation.

Adult We utilize sex (female or male) and race (white or black) as the sensitive attribute and income ($>50K or <=50$K) as the prediction target. The input features include hours-per-week, education-num, age, marital status, and relationship status (husband or wife). We group age into 12 disjoint intervals: [0, 20), [20, 25), ..., [65, 70), [>70) and hours-per-week into 14 disjoint intervals: [0, 10), [10, 15), ..., [65, 70), [>70)

COMPAS Race (African-American or Caucasian) and sex (female or male) are used as the sensitive attribute, and is_recid (recid. or no recid.) is the target for prediction. Input features consist of age, c_charge_degree, priors_count, c_jail_in, and c_jail_out. We derive the length_of_stay feature by subtracting c_jail_in from c_jail_out. Entries with inconsistent arrest information or missing information are removed. Traffic offenses are excluded. We quantize length_of_stay every 30 days and let 0 be a separate category. Age and length_of_stay are quantized similarly to the Adult dataset.

German Credit Age (below or above 25 years old) serves as the sensitive attribute, and the credit column (indicating whether the loan was a good decision) is the target for prediction. Input features include loan duration in months, credit amount, age, number of existing credits at this bank, credit history, savings, and length of present employment. Credit amount and loan duration are grouped into intervals: : [0, 5000), [5000, 10000), [>10000). We also group the duration of the loan into two categories: under 36 months and over 36 months.

2.2 Details of Model Training

The hyperparameters used are as follows: for the Adult dataset, Random Forest with n_estimators=15, min_samples_leaf=3, criterion = log_loss, and bootstrap = False; for the COMPAS dataset, Random Forest with n_estimators=17 originally. When we conduct experiments on the feature collection methods, we set n_estimators=85 for a less informative feature set. For the German Credit dataset, Random Forest with n_estimators=100, min_samples_split =2, min_samples_leaf=1.

As for the Reduction method, we use the AIF360 implementation of ExponentiatedGradientReduction. The allowed fairness constraint violation $\mathcal{E}$ varies for different datasets. For example, for the Adult dataset, $\mathcal{E}$ ranges from {0.001, 0.01, 0.05, 0.2, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 1, 2, 4, 8, 16}, while for COMPAS, it originally ranges from {0.0001, 0.001, 0.1, 0.6, 0.7, 0.9, 1, 5}. When we conduct experiments on the feature collection methods, we use the range {0.0001, 0.0005, 0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}. For the German Credit dataset, it ranges from {0.0001, 0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 4, 6, 8, 10, 50, 80, 100}.

2.3 Additional Experimental Results

In this section, we present our additional experimental results to further support our findings. We present our results here on the UCI-Adult dataset and the COMPAS dataset with gender as the sensitive attribute. Those results are consistent with our analysis that underrepresented group suffers degradation when fairness keeps improving. We also present our results here on the German Credit dataset. Our observations show that the German Credit dataset is {age<25}-skewed, and suffers degradation in underrepresented groups. Compared to this, there is no degradation when applying other fairness constraints like Equailized Odds and Demographic Parity. The results for our feature collection method are also illustrated here, which demonstrate the efficacy of our method.

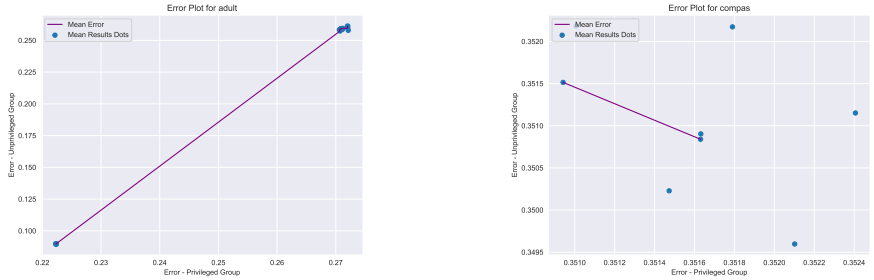

Fig. 4. Error plot of the UCI-Adult dataset and the COMPAS dataset, under the criteria of Accuracy Parity with gender as the sensitive attribute.

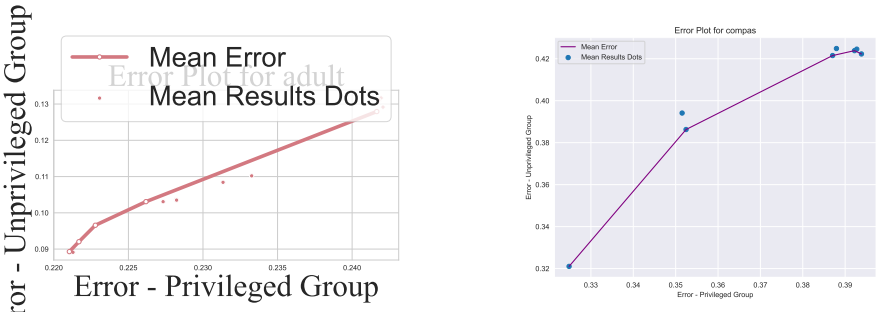

Fig. 5. Error plot of the UCI-Adult dataset and the COMPAS dataset, under the criteria of Equalized Odds with gender as the sensitive attribute.

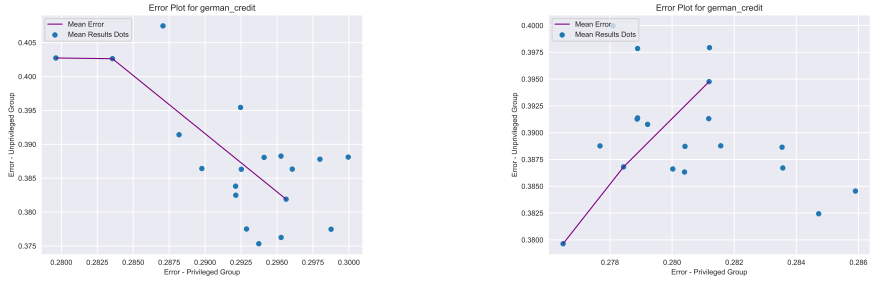

Fig. 6. Error plot of the German Credit dataset, under the criteria of Equalized Odds and Demographic Parity.

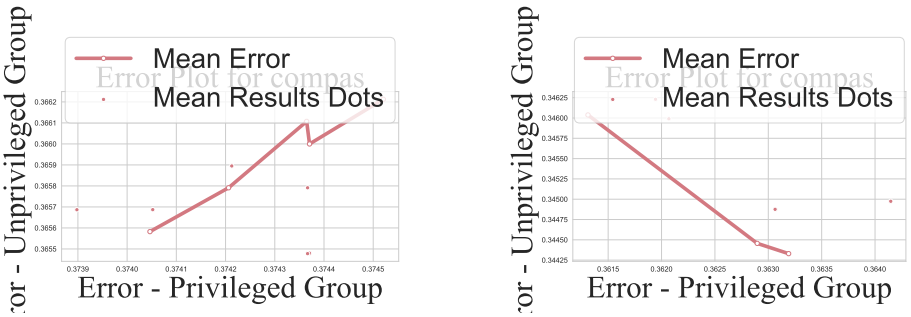

Fig. 7. Error plot of the German Credit dataset, under the criteria of Accuracy Parity, compared w/wo the informative features for the underrepresented groups.