Achieving Fairness through Constrained Recourse

Shuang Wang
Electrical and Computer Engineering
Clemson University
Clemson, SC, USA
swang8@clemson.edu

Yongkai Wu
Electrical and Computer Engineering
Clemson University
Clemson, SC, USA
yongkaw@clemson.edu

***Abstract*—Data-driven decision-making systems are progressively deployed in high-risk scenarios, raising significant social concerns about their potential to perpetuate inequalities related to demographic characteristics. Recent research efforts have focused on ensuring equal decision-making for individuals, primarily through model adjustments and data modification techniques. However, these approaches often rely on distance-based formulation, overlooking the practical aspects of achieving equity. To address this gap, our study introduces a novel method that leverages actionable recourse to reflect the feasibility of attaining fairness in decision-making. This method leverages constrained optimization to achieve fairness within limited budgets, thereby balancing equity with practical constraints. We present experimental results that demonstrate the superiority of our approach over traditional distance-based methods. These results underscore our method's potential in ensuring equitable decisions and maintaining feasibility and efficiency in real-world applications.**

***Index Terms*—machine learning, algorithmic fairness**

I. Introduction

The recent advance of machine learning has enabled unprecedented advancements in various sectors, ranging from healthcare, job placement, recommendation, and criminal justice. However, as these technologies become deeply integrated into critical decision-making processes, lack of fairness and equity has emerged as central concerns. Machine learning models, reflecting the patterns in the training data, have shown tendencies to perpetuate and even exacerbate existing societal inequities. This is particularly evident in cases where algorithmic decisions disproportionately affect individuals characterized by race, gender, or other demographic factors. Fairness-aware machine learning [1]–[11] seeks to address the ethical, social, and legal implications of decision systems, particularly their propensity to inherit and amplify existing prejudice, or even introduce new biases against certain demographic groups. In the fair machine learning literature, researchers have put forth various techniques for fair predictions, including model tweaking [1]–[4], data modification [5]–[8], and decision adjustment [9]–[11].

Against this backdrop, algorithmic recourse emerges as a vital tool in recent decision-making domains, aiming at providing individuals who have experienced adverse algorithmic predictions with a means to convert these unfavorable outcomes into favorable ones [12]–[14]. Notably, through recourse methods like actionable recourse [15], feasible and actionable counterfactual explanations (FACE) [16], and counterfactual reasoning-based explanations [17], those unfavorable predictions can be flipped with pre-defined costs, making recommendations to the original data by a series of actions. Equalizing Recourse [18] steps into this gap by introducing a regularization method using the boundary distance within a recourse framework, aiming to guarantee balanced penalties for unfavorable outcomes among different groups.

Existing practices in fairness for machine learning models often fall short of addressing the issue of biased decision-making [19]. To illustrate one common challenge, let us consider a running example. Suppose a model is utilized for loan approval, and its decisions are based on several factors, such as income, account balance, residential address, and marital status. While conventional fair decision-making methods focus on data adjustment or action recommendations to minimize discrimination, they typically do not consider the varying degrees of influence each modification has on the decision-making process. For instance, increasing income and account balances can substantially improve the chances of loan approval, whereas changes in residential address or marital status might have a negligible impact. Existing approaches that rely solely on the distance from the decision boundary fail to capture the actual effort required to change a decision, ignoring the importance of key features [1], [20], [21]. This oversight can lead to impractical solutions toward fair decisions, as models might focus on applicants closer to the decision threshold and overlook those who could improve their chances with minor changes. Such a narrow focus on boundary proximity does not guarantee fair outcomes and often leads to inefficient resource allocation if the fair model is deployed. This running example highlights the need for more nuanced and holistic approaches in fair machine learning practices to ensure equitable and effective decision-making.

Our motivation is driven by pursuing a method that recognizes the importance of key features and accurately captures the effort required for prediction reversal, thereby promoting fair outcomes with minimal actionable modifications. Additionally, we expect our strategy to guarantee the fair global distribution of resources beyond local adjustments. Therefore, we are motivated to propose a novel framework to model the actionable recourse and search for minimal modifications with constrained optimization. Our contributions are summarized as follows: i) Our study uncovers significant limitations in

the prevalent practice of using decision boundary distance as a primary criterion to reverse individual predictions in fairness-oriented machine learning models. ii) We develop a resource allocation method to fulfill fairness through globally constrained optimization rather than local adjustments, ensuring fairness among different groups throughout the decision-reversal process.

II. Preliminaries

A. Notions and Settings

In this manuscript, we utilize the following notions to describe the problem and solutions. A dataset $\mathcal{D}$ has m instances and w features, where $m, w \in \mathbb{N}^+$. Following the common i.i.d. assumption, each instance can be described as a vector $\boldsymbol{x}$ =$[x_1, x_2, ..., x_w] \subseteq X_1 \cup X_2 \ ... \cup X_w \subseteq \mathbb{R}^m$ with a ground truth label $y \in \{0, 1\}$, where $y = 1$ indicates a favorable outcome and $y = 0$ indicates an unfavorable outcome. We train a binary classification model denoted as $f(\boldsymbol{x})$. It is critically essential to emphasize that the binary classifier $f(\boldsymbol{x})$'s coefficients remain fixed after training with the original dataset. It reflects real-world scenarios where it is often impractical or impossible to alter an institution's established decision-making algorithms. By making minimal modifications to the original dataset, an algorithmic recourse method helps to flip unfavorable predictions to favorable ones with corresponding flipping costs. Notably, these flips can be achieved without changing the decision-making algorithm $f(\boldsymbol{x})$, which is consistent with the fixed model practice.

B. Fairness Notions and Metrics

A dataset $\mathcal{D}$ contains the sensitive attribute S, the decision label Y, and additional attributes within X so that $\mathcal{D} = \{S, X, Y\}$. The sensitive attribute S serves as a critical role in indicating the majority and minority groups. The sensitive attribute and decision label are binary, where $S = 1$ represents the majority group and $S = 0$ represents the minority group, respectively, with $S \in \{1, 0\}$ covering the entire scope of attribute S. Similarly, $Y = 1$ denotes favorable outcomes, while $Y = 0$ denotes unfavorable outcomes.

Definition 1: **Statistical Parity.** Given observational data $\mathbf{P}(S, X, Y)$, if it exhibits statistical parity between S and Y, denoted as $S \perp Y$, then the data is considered fair.

Based on statistical parity, numerous fairness metrics have been rigorously suggested, including predictive parity, opportunity equality, equalized odds, and other measurements [22], [23]. In this work, we adopt and meticulously apply the risk difference, $\mathbf{P}(Y = 1|S = 1) - \mathbf{P}(Y = 1|S = 0)$, denoted as $\mathcal{Q}$, to quantify the strength of bias [24].

C. Algorithmic Recourse

Cost-based algorithmic recourse methods, distinct from the boundary distance-based methods, are developed to measure specific flipping costs for individuals requiring prediction reversals [15]. Consider an individual $\boldsymbol{x}$ for whom a binary decision model makes an unfavorable prediction, indicated as $f(\boldsymbol{x}) = 0$. After applying a recourse method, we identify an action $\boldsymbol{a}$ with a minimal flipping cost, effectively turning this unfavorable prediction into a favorable one, i.e., $f(\boldsymbol{x} + \boldsymbol{a}) = 1$. Here, $\boldsymbol{a}$ represents feasible actions within realistic and ethical bounds that can be applied to an original data point $\boldsymbol{x}$'s undesirable prediction, with "minimal flipping cost" denoting the least modification needed on actionable features. Actions must stay within the dataset's observed feature ranges and cannot modify immutable features like gender, race, and other features that cannot be changed in real-world scenarios. To quantify the cost associated with prediction reversal, we utilize the flipping cost function $cost(\boldsymbol{a})$ from an algorithmic recourse methodology [15]. The most prevalent representation of the cost function is,

$$\begin{aligned} \min \quad & cost(\boldsymbol{a}; \boldsymbol{x}) \\ \text{s.t.} \quad & f(\boldsymbol{x} + \boldsymbol{a}) = 1, \\ \text{where} \quad & f(\boldsymbol{x}) = 0, \boldsymbol{a} \in \mathcal{A}(\boldsymbol{x}) \end{aligned} \tag{1}$$

Here, $cost(\boldsymbol{a}; \boldsymbol{x}) \rightarrow \mathbb{R}^+$ is a quantitative function for precisely calculating the flipping cost associated with a reversal prediction. Furthermore, this function rigorously adheres to two essential principles: i). $cost(0; \boldsymbol{x}) = 0$, no action means no cost; ii). $cost(\boldsymbol{a}; \boldsymbol{x}) \leq cost(\boldsymbol{a} + \boldsymbol{\tau}; \boldsymbol{x})$, more actions bring about more costs, where $\boldsymbol{\tau}$ represents the additional actions taken. Additionally, $\mathcal{A}(\boldsymbol{x})$ denotes a set of feasible actions that $\boldsymbol{x}$ can take, generated by a recourse strategy. It is crucial to note that $\mathcal{A}(\boldsymbol{x})$ plays a key role in determining the actions that an individual can reasonably take to reverse the decision in their favor without changing the decision-making algorithm $f(\boldsymbol{x})$. The optimal action $\boldsymbol{a}$ is determined by the dual criteria of minimizing the cost function $cost(\boldsymbol{a}; \boldsymbol{x})$ while concurrently shifting the prediction outcome from $f(\boldsymbol{x}) = 0$ to $f(\boldsymbol{x} + \boldsymbol{a}) = 1$. We collect flipping costs for those with unfavorable predictions, expressing them as a vector that $\boldsymbol{C} = [c_1, c_2, ..., c_n] \subseteq \mathbb{R}^n$, where $c_{\mathrm{i}} > 0$, and n represents the number of unfavorable predictions. In summary, given a decision-making model $f(\boldsymbol{x})$ and instances $\boldsymbol{x}$ with unfavorable predictions, the recourse method identifies feasible actions $\boldsymbol{a}$, which can be applied to flip the undesirable predictions, generating a corresponding flipping cost for each instance, resulting in a cost vector $\boldsymbol{C}$ representing the flipping costs for all undesirable predicted instances.

III. Methodology

A. Achieving Fairness via Global Optimization (AVIATOR)

With a dataset $\mathcal{D} = \{S, X, Y\}$, we denote the rejected *i-th* individual as (s_i, x_i, y_i), which can be extended to (s_i^+, s_i^-, x_i, y_i) using the one-hot encoding strategy for generality. In this context, s_i^+ and s_i^- are mutually exclusive since they belong to the majority group and minority group separately. Also, s_i^+ and s_i^- satisfy the properties that $s_i^+ + s_i^- = 1$ and $s_i^+ \neq s_i^-$ so that we can obtain the total number of the majority group as $s_t^+ = \sum_{i=1}^n s_i^+$, and the total number of the minority group as $s_t^- = \sum_{i=1}^n s_i^-$.

Next, we generate a vector $\boldsymbol{U} = [u_1, u_2, ..., u_n] \subseteq \mathbb{R}^n$, where each $u_i \in \{0, 1\}$ is strategically employed as an optimization variable. Specifically, u_i with a binary value indicates

whether the *i-th* individual is flipped. An initial unfavorable outcome is denoted by $y_i = 0$, and after flipping ($u_i = 1$), a favorable outcome is represented by $\hat{y}_i = y_i + u_i = 1$. In general, $\hat{Y}$ represents the new predictions after the flipping steps and the range of $\hat{Y}$ is $\{0, 1\}$. The risk difference quantifies the updating disparity of the probabilities with favorable outcomes between the majority and minority groups, defined as $\mathcal{Q} = \mathbf{P}(\widehat{Y} = 1|S = 1) - \mathbf{P}(\widehat{Y} = 1|S = 0)$. This measure can be interpreted as a representation of bias strength after the allocation process. Thus,

$$\begin{aligned} \mathcal{Q}(u_i) &= \mathbf{P}(\widehat{Y} = 1|S = 1) - \mathbf{P}(\widehat{Y} = 1|S = 0) \\ &= \sum_{i=1}^{n} \left(\frac{\hat{y}_i \cdot s_i^+}{s_t^+} - \frac{\hat{y}_i \cdot s_i^-}{s_t^-} \right) \end{aligned} \tag{2}$$

The flipping number for the majority group $\mathcal{N}^+(u_i)$ is,

$$\mathcal{N}^+(u_i) = \sum_{i=1}^{n} (\hat{y}_i \cdot s_i^+) - \sum_{i=1}^{n} (y_i \cdot s_i^+) \quad = \sum_{i=1}^{n} s_i^+ \cdot u_i \tag{3}$$

Similarly, the flipping number for the minority group is $\mathcal{N}^-(u_i) = \sum_{i=1}^{n} s_i^- \cdot u_i$. Additionally, the total flipping number $\mathcal{N}_t(u_i)$ is,

$$\mathcal{N}_t(u_i) = \sum_{i=1}^{n} s_i^+ \cdot u_i + \sum_{i=1}^{n} s_i^- \cdot u_i \quad = \sum_{i=1}^{n} u_i \tag{4}$$

As a result, the total flipping cost $\mathcal{C}_t(u_i)$ is,

$$\mathcal{C}_t(u_i) = \boldsymbol{C} \cdot \boldsymbol{U} = \sum_{i=1}^{n} c_i \cdot u_i \tag{5}$$

Therefore, our objective integrates two core principles from a global perspective. Firstly, we aim to maximize the number of individuals required to flip unfavorable predictions given a fixed-cost budget. Simultaneously, we aim to fulfill group-level fairness, ensuring that both majority and minority subgroups have a fair opportunity to make favorable predictions after recourse. These two principles are pursued concurrently, demonstrating a comprehensive strategy rather than isolated adjustments. To fulfill these principles, we develop an integer programming approach to find an optimal flipping strategy that maximizes fairness with a fixed-cost budget. Considering the constraints imposed by a limited budget b, we formulate an objective function $\mathcal{L}$ that makes an effort to accomplish these goals. Specifically, $\mathcal{L}$ is designed to minimize $\mathcal{Q}(u_i)$ (Eq. (2)) in order to mitigate discrimination while concurrently maximizing $\mathcal{N}_t(u_i)$ (Eq. (3)) to increase the number of flips for unfavorable predictions. This approach can be expressed as the following formulation,

$$\begin{aligned} \mathcal{L}(u_i) \;&=\; \min_{u_i \in \{0,1\}} \; (\mathcal{Q}(u_i))^2 \;-\; \lambda \cdot \mathcal{N}_t(u_i) \\ s.t. \; &\sum_{i=1}^{n} c_i \cdot u_i \;\leq\; b, \; \lambda \;\geq\; 0 \end{aligned} \tag{6}$$

Algorithm 1 AVATAR

Input: $R^- = \{s_i^-, x_i, y_i, c_i\}_{i=1}^{p}$, $M = \{s_j, x_j, y_j, c_j\}_{j=1}^{n}$, $k = 0$, $\tilde{\mathcal{D}} \leftarrow \emptyset$, ε.
Output: $\tilde{\mathcal{D}}$

for $i = 1$ **to** p **do**
  if $\mathcal{Q}(\cdot) > \varepsilon$ **then**
    $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{s_i^-, x_i, y_i, c_i\}$
    $y_i \leftarrow 1$
    pop $R^- = \{s_i, x_i, y_i, c_i\}_{i=i}$
    pop $M_j = \{s_j, x_j, y_j, c_j\}_{j=i}$
  else
    while M **do**
      $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{s_j, x_j, y_j, c_j\}_{k=0}$
      $y_k \leftarrow 1$
      pop $M_j = \{s_j, x_j, y_j, c_j\}_{j=k}$
      $k \leftarrow k + 1$ ▷ Increment k to track the next instance in M
      if $\mathcal{Q}(\cdot) > \varepsilon$ **then**
        break
return $\tilde{\mathcal{D}}$

IV. Experiment

We utilize two datasets for our experiments: the German Credit Dataset [25] and the Adult Census Dataset [26]. We identify instances with unfavorable predictions from established Logistic Regression [27] and XGBoost [28] as binary decision models. We employ two algorithmic recourse methods to flip these unfavorable predictions: the actionable-recourse method [15] and the interventional tree method [29]. Both methods are applied to flip the unfavorable predictions from the Logistic Regression and XGBoost models, generating corresponding flipping costs and detailing the necessary modifications on the original datasets. Our implementation is developed based on Gurobi (Ver. 9.0.0).

A. Experimental Datasets

The German Credit Dataset targets predicting an individual's credit risk, either good or bad, based on multiple attributes. We treat "Gender" as a sensitive attribute with binary values. Here, "male" refers to the majority group, and "female" refers to the minority group, denoted as $S = 1$ and $S = 0$, respectively. The initial bias value for this German Credit dataset is 7.48%.
The Adult Census Dataset commonly used in fairness-aware machine learning, is another dataset chosen in this experiment. We consider "sex" as a sensitive attribute with a binary value, where the value 1 refers to "Male" and the value 0 refers to "Female", denoted as $S = 1$ and $S = 0$, respectively. Additionally, the task is to predict whether an individual has a high income (">50K") or a low income ("≤50K") in a financial year. The initial bias value for this dataset is 19.53%.

B. Baselines

1) Boundary Distance Evaluation Approach (MASAGE): We present MASAGE as a comparative framework to highlight the limitation of the boundary distance-based method in

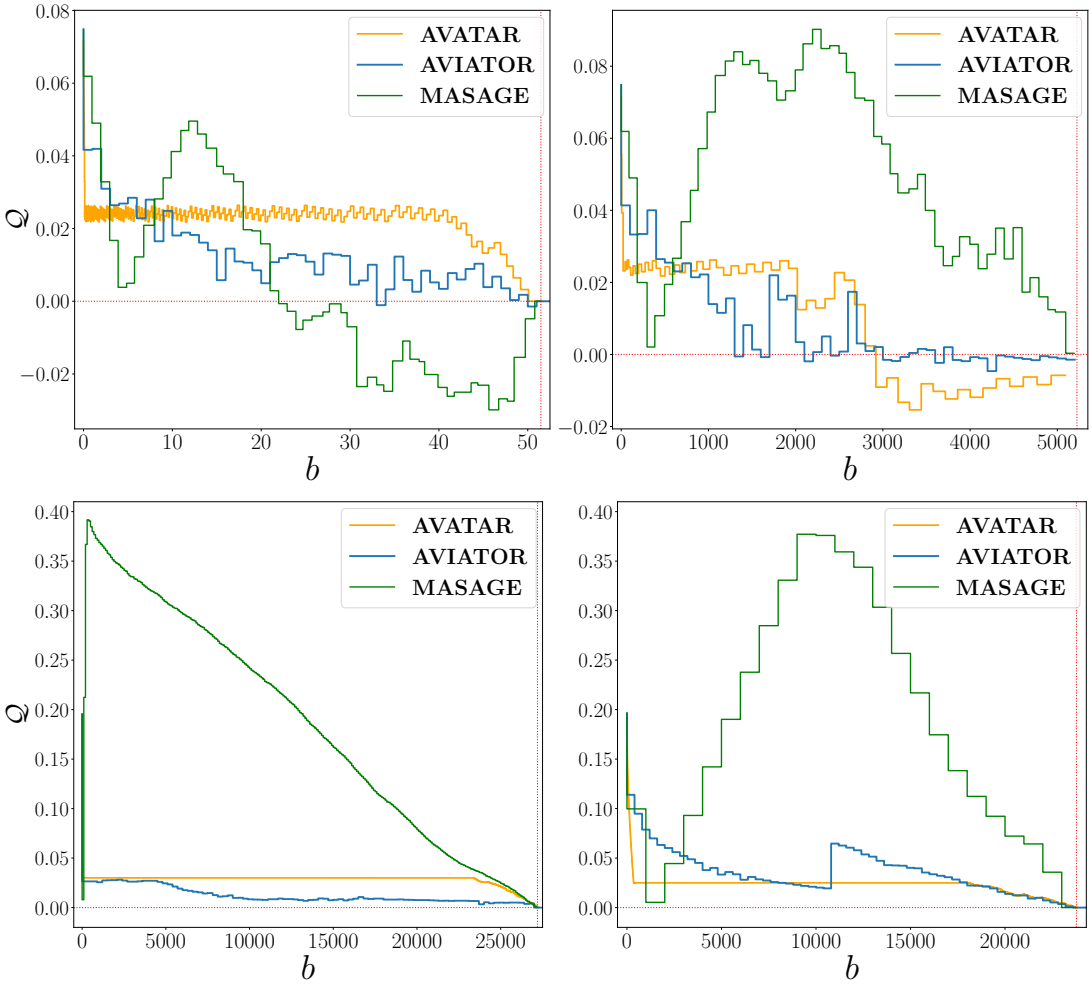

Fig. 1. Comparison of fairness for German Credit (above) and Adult (below) datasets. With the left and right corresponding to the Logistic Regression and XGBoost models, respectively.

fairness. MASAGE, inspired by the work of Kamiran and Calders [30], prioritizes instances based on their distances to the decision boundary, with a particular focus on the minority group. Here, given an established decision-making model and a data point, the boundary distance is defined as the absolute difference between its predicted outcome and a prediction threshold. The MASAGE baseline works in the following way. Evaluation via Boundary Distance: Instances in the majority group are arranged into a sequence based on ascending boundary distances, and similarly, instances in the minority group are independently subjected to the same procedure. The idea behind this arrangement is that the instances nearer the decision boundary may have the outcomes reversed more easily, indicating fewer efforts required for flipping action; Prioritizing Minority Group: MASAGE starts by prioritizing the minority group, making sequential flipping based on their boundary distances to reduce bias, ultimately promoting a fairer representation in the decision-making process; Managing Remaining Instances: After flipping almost all instances from the minority group, the remaining instances, which may include a few from the minority group and all from the majority group, are re-ranked and then flipped by their boundary distances with increasing order.

The reasons for selecting MASAGE as a baseline are threefold: i) We aim to illustrate the limitation of the boundary distance metric with the MASAGE baseline because it particularly utilizes boundary distance as the primary criterion for its flipping operation; ii) The straightforward boundary distance metric makes MASAGE an ideal baseline for demonstrating the advantages of our proposed method; iii) The MASAGE prioritizes the minority group, providing a distinctive comparative perspective, which allows us to indicate how our method more effectively achieves fairness across groups.

2) Localized Fairness Adjustments (AVATAR): As seen in Algorithm 1, we detail the AVATAR baseline, a local adjustment strategy for decision reversal.

This method has three modules. Input: R^- represents the individuals of the minority group that are not favored by the decision model, which has p instances. M contains all individuals with unfavorable predictions, including majority and

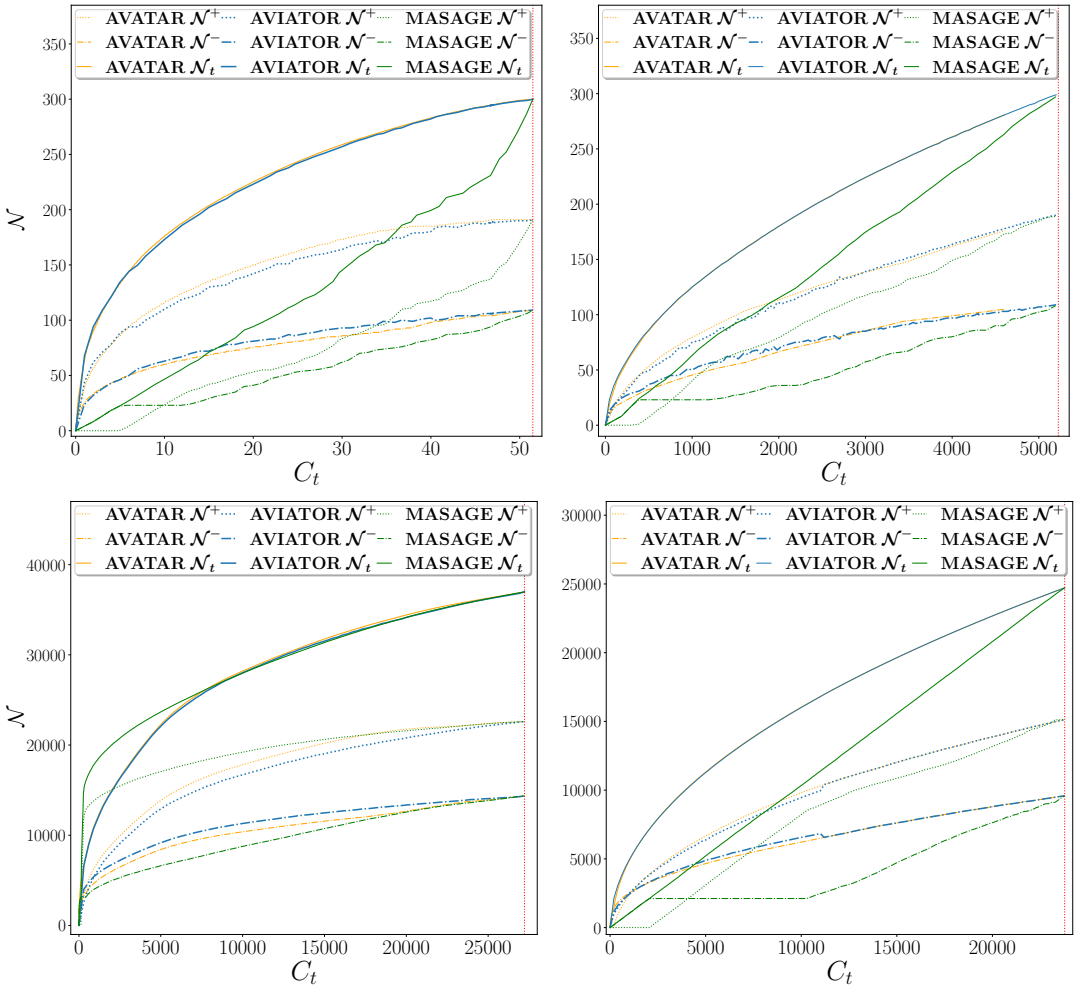

Fig. 2. Comparison of total flipping number for German Credit (above) and Adult (below) datasets of glocal view. With the left and right corresponding to the Logistic Regression and XGBoost models.

minority groups, totaling n instances. R^- and M are sorted increasingly based on the flipping costs from the recourse method. The iterator variable k is initialized to 0 to track the instance in M. $\widetilde{D}$ is an empty set and stores the flipped instances. And the ϵ is the bias threshold value; Main Loop: For each individual in R^- (Line 1), prioritizing the minority group if the bias measurement $\mathcal{Q}$ exceeds a threshold ϵ (Line 2). The individual who meets the condition is added to $\tilde{\mathcal{D}}$ with the flipping of its unfavorable prediction (Line 2-4). Then, remove this instance from R^- and the corresponding instance in M (Lines 5-6). Otherwise, while M is not empty (Lines 7-8), the algorithm selects the individual in M tracked by k, flips its prediction, and pops the corresponding instance in M (Lines 9-11). Then, the tracker k is incremented (Line 12). This process continues until $\mathcal{Q}$ exceeds ϵ, or M is empty (Lines 13-14); Output: The algorithm returns $\tilde{\mathcal{D}}$, a set of flipped all individuals (Line 15).

The reasons for selecting AVATAR as a baseline are twofold: i). As a suitable contrast to our global optimization approach, AVATAR's reliance on a local adjustment strategy highlights the shortcomings in resource allocation; ii). The AVATAR's iterative approach cycles focus on majority and minority groups, harmoniously integrating cost-efficiency, fairness, and locality in the allocation process.

C. Experimental Results And Analysis

1) Comparison of Fairness: A comprehensive representation of this comparison is shown in Fig. 1, which illustrates the fairness trends of our study methods.

MASAGE. Results: After a sharp initial fall, the bias curve of the MASAGE baseline soon recovers. This pattern is consistently observed across two datasets, with the Adult dataset exemplifying the trend prominently; Analysis: MASAGE first prioritizes the minority group instances closest to the decision boundary, leading to a rapid decrease in bias. As it starts addressing the remaining instances, which consist of a few instances from the minority group farther from the decision boundary and a large number of instances from the majority group close to the boundary, the bias begins to increase. This rise is due to prioritizing majority group instances near the boundary after the closest minority group instances have been flipped. The bias stabilizes at zero when all instances are

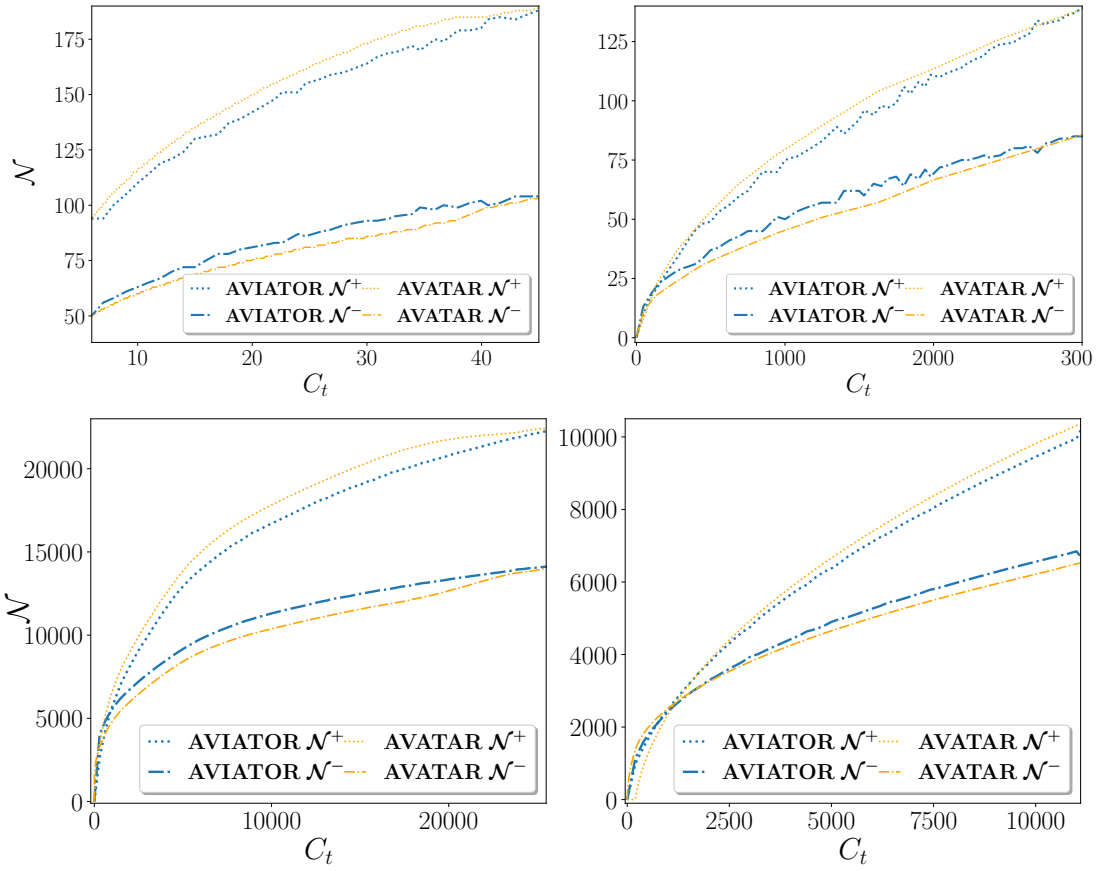

Fig. 3. Comparison of flipping number for German Credit (above) and Adult (below) datasets of local view. With the left and right corresponding to the Logistic Regression and XGBoost models.

flipped. Because of MASAGE's strong focus on the minority group, it has a propensity to reintroduce bias swiftly after it has been decreased.

AVATAR. Results: The bias curve of the AVATAR baseline exhibits a sharp reduction around ϵ, which is set to 0.025 in Algorithm 1, holding this value around until the bias value falls to 0. AVATAR performs more fairly than the MASAGE baseline; Analysis: AVATAR is a cost-sensitive approach, which starts by flipping instances from the minority group to reduce bias. When the bias drops below ϵ, it shifts its focus to the majority group, targeting instances with the lowest flipping costs. If the bias surpasses ϵ, the focus reverts to the minority group. This iterative approach between groups enables a stable bias trajectory, contrasting with the behavior observed in the MASAGE baseline.

AVIATOR. Results: Although AVIATOR's bias value fluctuates as the allocation continues, they are constantly lower than MASAGE and AVATAR; Analysis: AVIATOR aims to maximize the utilization of the budget globally while providing both groups a fair opportunity to reverse undesirable outcomes. AVIATOR can provide a lower bias value with closely matched total flipping numbers by fine-tuning the hyperparameter λ in Eq. (6) and consistently generate smaller bias values throughout the process than MASAGE and AVATAR. MASAGE with the boundary distance metric fails to capture accurately the actual effort required to flip a prediction, thereby revealing its limitation in individual prediction reversal. Moreover, maintaining fairness during the allocation process is challenging for this distance-based strategy, resulting in a more biased output. Additionally, despite using the adjustment via ϵ to modify fairness, the cost-sensitive AVATAR cannot reach the level of fairness attained by our AVIATOR method.

2) Comparison of Flipping Number: To illustrate the advantages of our AVIATOR solution, we will provide global (Fig. 2) and local (Fig. 3) perspectives of the flipping numbers under the same cost constraint.

MASAGE. Results: Overall, MASAGE flips less under the same budget b than the other two approaches. However, MASAGE performs remarkably well for $\mathcal{C}_t \leq 8000$ in the Adult dataset; Analysis: The MASAGE strategy has an inherent prioritization towards flipping instances within the minority group, which may induce a delay when addressing instances from the majority group.

AVATAR & AVIATOR. Results: We tune the hyperparameter λ in Eq.(6) to achieve the total flipping number of the AVIATOR near that of the AVATAR. According to Fig. 3, the AVIATOR approach demonstrates a preference for the minority group in the 10-45 and 0-3000 ranges of the German

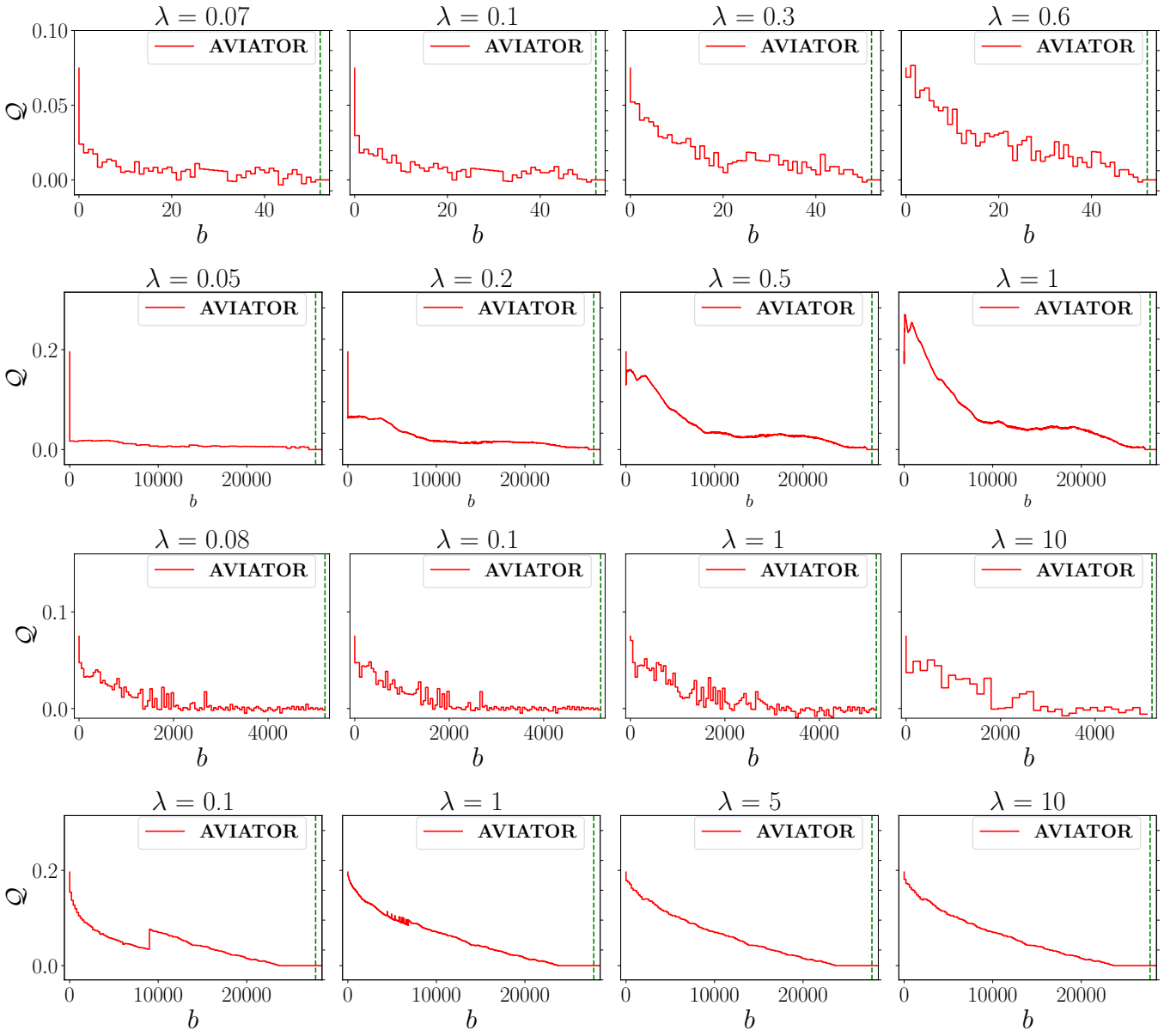

Fig. 4. The impact of λ for the AVIATOR method on different datasets and models, displayed in sequence from top to bottom: first and second images show the German Credit and Adult datasets with the Logistic Regression model, respectively; third and fourth images illustrate the German Credit and Adult datasets with the XGBoost model.

Credit dataset. Similar findings can be observed in the Adult dataset, where the AVIATOR prefers the minority group in the 1500–12000 and 400–25000 ranges; Analysis: Although the total flipping number of AVIATOR and AVATAR are relatively similar, we notice that AVIATOR flips a higher proportion of instances from minority groups at particular $\mathcal{C}_t$ levels. While AVATAR works cyclically to ensure fairness, it may not always be the most effective. In certain iterations, some minority instances may remain unflipped if they do not meet the criteria for flipping. Conversely, AVIATOR employs a continual strategy that considers all instances without giving the minority group special priority. AVIATOR can allocate resources flexibly by not segregating the instances into separate cycles. In contrast to AVATAR's cyclical approach, this implies that the minority instances can be flipped earlier in the process if it has a relatively low flipping cost. This experiment reveals that MASAGE, with boundary distance metric, fails to maximize the number of flips under resource constraints, thus not reflecting the effort needed to flip instances. AVATAR, while cost-sensitive, is not the most effective due to its local adjustments, leading to delays in flipping and an undue bias towards minority groups. By flipping more instances within the same budgetary constraints, precisely estimating the effort needed for reversal, and maximizing the utilization of

resources, our method outperforms these approaches.

3) The Impact of Hyperparameter λ on $\mathcal{Q}$: As delineated by Eq. (6), the λ emerges as a pivotal determinant in the AVIATOR framework, controlling the optimization equilibrium between the flipping number and the bias metric $\mathcal{Q}$. This interplay of λ in optimization reveals that an increase in λ makes the system more inclined to maximize flipping instances, potentially at the expense of adequate bias reduction. On the other hand, a more conservative λ recalibrates the system's emphasis toward pronounced bias reduction, demonstrating discernment in instance selection for flipping, prioritizing significant bias mitigation. This trade-off balance indicates that an increase in λ correlates with a noticeable rise in bias across both datasets, empirically supported by Fig. 4.

V. Conclusion

Our study provides critical insights into the limitations of traditional fairness approaches in machine learning, particularly those relying heavily on the concept of decision boundary distance. We have demonstrated that this conventional method fails to address the practical aspects of altering model predictions, especially in terms of feasibility and cost-effectiveness. Our research proposes an alternative approach that incorporates cost functions to assess and implement actionable changes, thereby balancing fairness with practical applicability. This approach not only enhances the fairness of the outcomes but also ensures that the modifications are feasible and within reasonable budgetary constraints. Ultimately, this balanced approach will contribute to building trust in machine learning systems and ensuring their beneficial impact on society.

References

[1] 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, WWW 2017, Perth, Australia, April 3-7, 2017*. ACM, 2017, pp. 1171–1180.

[2] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011*. IEEE Computer Society, 2011, pp. 643–650.

[3] Y. Wu, L. Zhang, and X. Wu, "On convexity and bounds of fairness-aware classification," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 2019, pp. 3356–3362.

[4] X. Jiang, Y. Dai, and Y. Wu, "Fair selection through kernel density estimation," in *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*. IEEE, 2023, pp. 1–8.

[5] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*. ACM, 2008, pp. 560–568.

[6] Y. Wu and X. Wu, "Using loglinear model for discrimination discovery and prevention," in *2016 IEEE International Conference on Data Science and Advanced Analytics*. IEEE, 2016, pp. 110–119.

[7] L. Zhang, Y. Wu, and X. Wu, "A causal framework for discovering and removing direct and indirect discrimination," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 2017.

[8] ——, "Achieving non-discrimination in data release," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1335–1344.

[9] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016.

[10] Y. Wu, L. Zhang, and X. Wu, "Counterfactual fairness: Unidentification, bound and algorithm," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. ijcai.org, 2019, pp. 1438–1444.

[11] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in prediction," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. ijcai.org, 2018, pp. 3097–3103.

[12] A. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: Contrastive explanations and consequential recommendations," *ACM Comput. Surv.*, no. 5, pp. 95:1–95:29, 2023.

[13] A. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: from counterfactual explanations to interventions," in *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*. ACM, 2021, pp. 353–362.

[14] A. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: definitions, formulations, solutions, and prospects," *CoRR*, 2020.

[15] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. ACM, 2019, pp. 10–19.

[16] R. Poyiadzi, K. Sokol, R. Santos-Rodríguez, T. D. Bie, and P. A. Flach, "FACE: feasible and actionable counterfactual explanations," in *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*. ACM, 2020, pp. 344–350.

[17] C. Fernandez, F. J. Provost, and X. Han, "Explaining data-driven decisions made by AI systems: The counterfactual approach," *CoRR*, 2020. [Online]. Available: https://arxiv.org/abs/2001.07417

[18] V. Gupta, P. Nokhiz, C. D. Roy, and S. Venkatasubramanian, "Equalizing recourse across groups," *CoRR*, 2019. [Online]. Available: http://arxiv.org/abs/1909.03166

[19] M. Wan, D. Zha, N. Liu, and N. Zou, "In-processing modeling techniques for machine learning fairness: A survey," *ACM Trans. Knowl. Discov. Data*, no. 3, pp. 35:1–35:27, 2023.

[20] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research. PMLR, 2017, pp. 962–970.

[21] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. ACM, 2020, pp. 607–617.

[22] P. Garg, J. D. Villasenor, and V. Foggo, "Fairness metrics: A comparative analysis," in *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*. IEEE, 2020, pp. 3662–3666.

[23] E. del Barrio, P. Gordaliza, and J. Loubes, "Review of mathematical frameworks for fairness in machine learning," *CoRR*, 2020. [Online]. Available: https://arxiv.org/abs/2005.13755

[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, no. 6, pp. 115:1–115:35, 2022.

[25] H. Hofmann, "Statlog (german credit data)," https://doi.org/10.24432/C5NC77, Nov. 1994.

[26] B. G. Becker and R. Kohavi, "Adult," https://doi.org/10.24432/C5XW20, Apr. 1996.

[27] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. Springer, 1989.

[28] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 2016, pp. 785–794.

[29] J. Klaise, A. V. Looveren, G. Vacanti, and A. Coca, "Alibi explain: Algorithms for explaining machine learning models," *J. Mach. Learn. Res.*, pp. 181:1–181:7, 2021. [Online]. Available: http://jmlr.org/papers/v22/21-0017.html

[30] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd international conference on computer, control and communication*, 2009.