Minimax AUC Fairness: Efficient Algorithm with Provable Convergence

Zhenhuan Yang[1*], **Yan Lok Ko**[2], **Kush R. Varshney**[3], **Yiming Ying**[2†]

[1]Etsy, Inc, Brooklyn, New York, USA
[2]University at Albany, State University of New York, Albany, New York, USA
[3]IBM Research, Yorktown Heights, New York, USA
zhenhuan.yang@hotmail.com, yko@albany.edu, krvarshn@us.ibm.com, yying@albany.edu

Abstract

The use of machine learning models in consequential decision making often exacerbates societal inequity, in particular yielding disparate impact on members of marginalized groups defined by race and gender. The area under the ROC curve (AUC) is widely used to evaluate the performance of a scoring function in machine learning, but is studied in algorithmic fairness less than other performance metrics. Due to the pairwise nature of the AUC, defining an AUC-based group fairness metric is pairwise-dependent and may involve both intra-group and inter-group AUCs. Importantly, considering only one category of AUCs is not sufficient to mitigate unfairness in AUC optimization. In this paper, we propose a minimax learning and bias mitigation framework that incorporates both intra-group and inter-group AUCs while maintaining utility. Based on this Rawlsian framework, we design an efficient stochastic optimization algorithm and prove its convergence to the minimum group-level AUC. We conduct numerical experiments on both synthetic and real-world datasets to validate the effectiveness of the minimax framework and the proposed optimization algorithm.

Introduction

Recent years have witnessed an increasing recognition that allocation decisions unfavorable to people from vulnerable groups (defined by sensitive attributes such as race, gender, and age) worsen with the use of machine learning. Alongside, a burgeoning set of mitigation algorithms has been developed (Agarwal et al. 2018; Calders, Kamiran, and Pechenizkiy 2009; Calmon et al. 2017; Chouldechova and Roth 2020; Donini et al. 2018; Hardt, Price, and Srebro 2016; Lohia et al. 2019; Pleiss et al. 2017; Zafar et al. 2017). Recognizing that they are only a small sliver of all actions that may be taken when viewing the program of justice holistically, the aim of bias mitigation is to ensure that the output of a classifier is not dependent on sensitive attributes. Most existing work focuses on statistical fairness metrics composed of entries of the classifier's confusion matrix.

On another front of machine learning, the area under the ROC curve (AUC) (Hanley and McNeil 1982) is one of the most widely used performance metrics in classification tasks with class-imbalance and when the relative costs of false positives and false negatives are difficult to pin down, and in bipartite ranking tasks. Learning a scoring function by maximizing its AUC — instead of the accuracy — () (Cortes and Mohri 2003; Gao et al. 2013; Ying, Wen, and Lyu 2016; Liu et al. 2020; Lei and Ying 2021; Yang and Ying 2022; Zhao et al. 2011; Yang et al. 2021a). However, there is not yet much work on AUC-related fairness in machine learning.

Properly defining group-level AUC fairness leads to two categories of metrics, depending on the specific groups to which the positive and negative examples belong. The first type, *intra-group* AUC, constrains both positive and negative examples to the same group. The second type, *inter-group* AUC,[1] computes the metric with positive and negative examples being from different groups. On one hand, by only focusing on intra-group AUC, one does not fully account for all possible disparate impacts.[2] Indeed, as observed by Kallus and Zhou (2019), similar intra-group AUCs may still lead to a disparate impact where positive examples of one group are misranked below negative examples of another group. On the other hand, we also witness from the synthetic experiment in Figure 1 that solely relying on inter-group AUC fairness can overlook unfairness with respect to the intra-group AUC. These two observations strongly suggest that to mitigate disparate impact when the performance metric is the AUC score, one should consider *both* inter- and intra-group AUC fairness during the learning process.

In this paper, we follow the Rawlsian principle of maximin welfare for distributive justice (Rawls 2001) and formalize our fairness goal as follows:

Find a scoring function that maximizes the minimum of inter-group AUC and intra-group AUC.

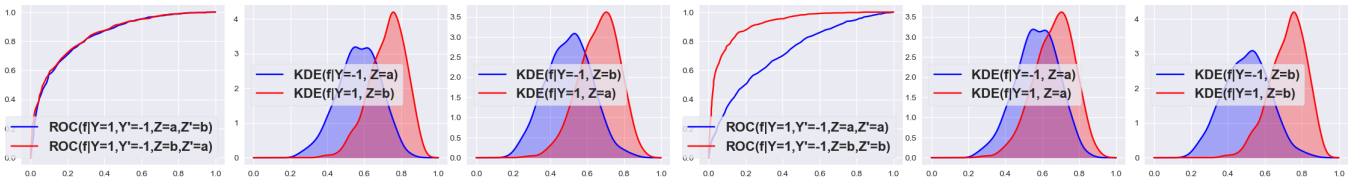
Unlike usual discrimination-aware approaches that put constraints on the norm of fairness metrics, the maximin principle does not introduce unnecessary harm (Ustun, Liu, and Parkes 2019; Martinez, Bertran, and Sapiro 2020). Hence it is more natural for our initial purpose of learning via AUC maximization.

[*]This work was mostly done while Zhenhuan Yang was a student at University at Albany, SUNY.

[†]Corresponding Author.

[1]Kallus and Zhou (2019) originally name this xAUC.

[2]We use the term disparate impact in the general sense of inequality in outcomes across groups, not in the specific sense of the ratio of selection rates across groups.



(a) *Inter-group ROC curves and Positive/negative KDE comparison* (b) *Intra-group ROC curves and Positive/negative KDE comparison*

Figure 1: Illustration of the discrepancy between inter-group AUC and intra-group AUC. In this example, inter-group AUC is in a fair situation and intra-group AUC is in an unfair situation. Here f denotes the synthetic scoring function sampled from Gaussian distribution and KDE denote the kernel density estimation of f, Y denotes the class label and Z denotes the protected attribute (see Appendix E for more details). In part (a), gaps of positive peaks and negative peaks are similar when interchanging groups. In part (b), the gap of the positive peak and the negative peak is small and the overlap is large when $Z = Z' = a$, which indicates probability of positive sample being misranked than negative sample is higher than $Z = Z' = b$.

Our main contributions are summarized as follows.

1. We justify the necessity of simultaneously achieving fairness in terms of intra- and inter-group AUCs. We then propose a minimax learning framework under the Rawlsian principle, collecting the objectives of both into one.
2. We propose a stochastic algorithm that updates the model parameter by gradient descent and the group weight by mirror ascent. We then prove the maximum re-weighted group error is guaranteed to be minimized by our algorithm in the nonconvex-concave setting.
3. We conduct numerical experiments on real-world datasets. We demonstrate the effectiveness of the minimax AUC fairness framework by observing substantial utility improvement over the notion of equality of error rates, and also fairness improvement over frameworks dealing with intra-group or inter-group AUC alone.

Related Work

Algorithmic fairness has received much attention recently, especially in the context of classification. There are three main strategies for bias mitigation: pre-processing the training data (Dwork et al. 2012; Feldman et al. 2015; Calmon et al. 2017), enforcing fairness directly during the training step (known as in-processing) (Kamishima, Akaho, and Sakuma 2011; Agarwal et al. 2018; Donini et al. 2018), and post-processing the classifications (Hardt, Price, and Srebro 2016; Pleiss et al. 2017; Lohia et al. 2019). Our work fits in the middle category of in-processing.

AUC-Related Fairness Metrics. Several works have attempted to address unfairness concerns in AUC-related problems (See Appendix A for a complete discussion). Dixon et al. (2018) propose the Pinned AUC fairness metric, which works by resampling the data such that each of the two groups make up 50% of the data, and then calculating the AUC on the resampled dataset. Beutel et al. (2019) propose ranking pairwise fairness definitions and a methodology that regularizes the training objective through a term that measures the correlation between the residual of selected and unselected people. Kallus and Zhou (2019) observe inter-group AUC (xAUC) unfairness in the COMPAS dataset and propose a post-processing approach to achieve xAUC similarity. Narasimhan et al. (2020) propose a cross-group fairness metric for general pairwise ranking problems and propose to maximize AUC under cross-group fairness constraints. Vogel, Bellet, and Clémençon (2021) define a fairness metric in terms of the ROC itself and propose a regularization method to achieve fairness. It is worth mentioning that all of the above works only focus on closing the gap of either intra-group metric or inter-group metric, but ignore the interplay between them. Furthermore, forcing unrealistic small group differences could harm the overall AUC maximization target.

Minimax Principle in Algorithmic Fairness. The Rawlsian principle has inspired several recent works to develop minimax frameworks to mitigate the disparate impact of machine learning models during training. In particular, Mohri, Sivek, and Suresh (2019) apply the agnostic federated learning framework in the minimax group fairness context. Martinez, Bertran, and Sapiro (2020) view the minimax fairness framework as a multi-objective function and pursue the Pareto front solution. Lahoti et al. (2020) study fairness without demographics and update group weights via adversarial learning. Shekhar et al. (2021) propose an adaptive sampling algorithm based on the principle of optimism to update group weights. Diana et al. (2021) utilize minimax group errors as a fairness constraint upper bound for further optimization.

The above minimax fairness frameworks all focus on classification and none has considered the AUC metric. Thus, the work herein is unique in applying the Rawlsian principle to AUC problems and considering both inter- and intra-group AUC simultaneously.

The Minimax Pairwise-Group Fairness Framework For Optimizing AUC

Let X and Y be two random variables. Here $Y \in \mathcal{Y} = \{\pm 1\}$ denotes the binary output label and $X \in \mathcal{X}$ denotes the input features where $\mathcal{X}$ is a closed and bounded domain in $\mathbb{R}^d$. Define a scoring function $f_\theta : \mathcal{X} \to \mathbb{R}$, where θ is the model parameter taking values in Θ. The AUC of f_θ measures the probability that it correctly ranks a positive example above a negative example, i.e.,

$$\mathrm{AUC}(f_\theta) = \mathbb{E}[\mathbb{I}[f_\theta(X) > f_\theta(X')]|Y = 1, Y' = -1], \quad (1)$$

where $\mathbb{I}$ is the indicator function taking value 1 when the event is true, and 0 otherwise. Observe that the definition (1) depends on a pair of examples (X, Y) and (X', Y').

In the context of fairness, we consider a third random variable $Z \in \mathcal{Z}$ to denote the sensitive group attribute, such as gender. For notational simplicity, we restrict our current interest to the case of two groups, i.e., $\mathcal{Z} = \{a, b\}$. However, our results can easily be extended to the general setting of multiple groups. The distribution $\mathcal{D}$ of the triplet (X, Y, Z) can be expressed as a mixture of the distributions of $X, Y|Z = z$. We define the group-level AUC as

$$\mathrm{AUC}_{z,z'}(f_\theta) = \mathbb{E}[\mathbb{I}[f_\theta(X) > f_\theta(X')]|Y=1, Y'=-1, Z=z, Z'=z']. \quad (2)$$

It is worth mentioning the group-level AUC naturally enjoys a pairwise dependence w.r.t. groups Z, Z'. Following the terminology in Beutel et al. (2019), we name such group-level AUC as *intra-group* AUC when $z = z'$, and *inter-group* AUC when $z \neq z'$. Previous work has aimed for fairness in these definitions separately (Beutel et al. 2019; Kallus and Zhou 2019; Narasimhan et al. 2020), i.e. either

$$\mathrm{AUC}_{a,a}(f_\theta) = \mathrm{AUC}_{b,b}(f_\theta), \text{ or}$$
$$\mathrm{AUC}_{a,b}(f_\theta) = \mathrm{AUC}_{b,a}(f_\theta).$$

However, if one only requires fair treatment in terms of intra-group (resp. inter-group) AUCs, the model may still suffer from unfair treatment in terms of inter-group (resp. intra-group) AUCs as argued by Kallus and Zhou (2019) and Figure 1. To address this issue, one naive solution is to require fairness in both of them so that $\mathrm{AUC}_{a,a}(f_\theta) \approx \mathrm{AUC}_{b,b}(f_\theta) = \kappa_1$ and $\mathrm{AUC}_{a,b}(f_\theta) \approx \mathrm{AUC}_{b,a}(f_\theta) = \kappa_2$. However, the potential discrepancy of $\kappa_1 \neq \kappa_2$ could also lead to unfairness. We elaborate in the following example.

Consider finding qualified candidates with gender as the sensitive attribute. In this scenario, AUC measures the probability of a qualified candidate being ranked higher than an unqualified candidate. Assuming the fair intra-group AUC is approximately 80%, which indicates the chance of a qualified female ranking higher than an unqualified female is more or less the same 80% as the male. Further assuming the fair inter-group AUC is approximately 60%, which means the chance of a qualified female ranking higher than an unqualified male is more or less the same 60% as the reverse case. Nevertheless, this leads to the unfair situation that a qualified female ranking higher than an unqualified male (60%) is lower than herself ranking higher than another unqualified female (80%)! Moreover, if male candidates are the majority of the applicant pool, her qualification is likely to be overwhelmed.

The above example demonstrates that disparity between intra-group AUCs and inter-group AUCs should not be allowed. Ideally speaking, group-level AUC fairness should be cast as

$$\mathrm{AUC}_{a,a}(f_\theta) = \mathrm{AUC}_{a,b}(f_\theta) = \mathrm{AUC}_{b,a}(f_\theta) = \mathrm{AUC}_{b,b}(f_\theta). \quad (3)$$

In the fair hiring example, the above identity can be interpreted as

The chance of a qualified candidate from any gender ranking higher than an unqualified candidate from any gender is the same.

While the above discussion naturally leads to a regularization or a constrained optimization fairness scheme, we adapt the minimax fairness scheme (Martinez, Bertran, and Sapiro 2020; Diana et al. 2021). To explain our choice, we first decompose the overall AUC as a mixture of the intra-group and the inter-group AUCs.

$$\mathrm{AUC}(f_\theta) = \sum_{z \in \mathcal{Z}} \mathbb{P}[Z = z, Z' = z] \, \mathrm{AUC}_{z,z}(f_\theta) + \sum_{z \neq z'} \mathbb{P}[Z = z, Z' = z'] \, \mathrm{AUC}_{z,z'}(f_\theta), \quad (4)$$

where $\mathbb{P}[Z, Z']$ is the prior distribution of a pair of sensitive attributes. Recall that the target is to optimize AUC, or to maximize the probability of a qualified candidate ranking higher than an unqualifed one in the example. Eq. (4) indicates that maximizing intra-group or inter-group AUC does not conflict with the interest of this target. Therefore, maximizing the minimum group-level AUC will have the potential of not significantly sacrificing the overall AUC while boosting the similarity on Eq. (3). Based on the above discussion, we are in position to formalize the learning problem in a maximin sense.

$$\max_{\theta \in \Theta} \min_{z,z' \in \mathcal{Z}} \mathrm{AUC}_{z,z'}(f_\theta). \quad (5)$$

Eq. (5) is intractable for two reasons. First, the true distribution $\mathcal{D}$ is unknown. In practice, we only have access to a sample $S = \{S_1, \cdots, S_n\}$ drawn from $\mathcal{D}$. For convenience, we denote $S_i \in S$ by $\mathbf{x}_i^{z+}$ (resp. $\mathbf{x}_i^{z-}$) if it is coming from group z with positive (resp. negative) label, we further denote $S^{z+} = \{\mathbf{x}_i^{z+} | i \in [n]\}$ and $S^{z-} = \{\mathbf{x}_i^{z-} | i \in [n]\}$ as two subsets of such examples. We can therefore define the empirical group-level AUC as:

$$\widehat{\mathrm{AUC}}_{z,z'}(f) = \widehat{\mathrm{AUC}}(f; S^{z+}, S^{z'-}) = \frac{1}{n^{z+} n^{z'-}} \sum_{i=1}^{n^{z+}} \sum_{j=1}^{n^{z'-}} \mathbb{I}[f_\theta(\mathbf{x}_i^{z+}) > f_\theta(\mathbf{x}_j^{z'-})],$$

where n^{z+} and $n^{z'-}$ denote the number of samples from $S^{z+}, S^{z'-}$, respectively. Furthermore, we replace the non-differentiable indicator function $\mathbb{I}$ with some (sub)-differentiable and non-increasing surrogate loss function ℓ. Let $\hat{R}^\ell_{z,z'} = \hat{R}^\ell(\cdot; S^{z+}, S^{z'-})$ denote the ℓ-surrogated empirical risk for AUC with group z, z'. And let $\hat{R}^\ell(\cdot) = \hat{R}^\ell(\cdot; S) = (\hat{R}^\ell(\cdot; S^{z+}, S^{z'-}))_{z,z' \in \{a,b\}}$. The problem therefore becomes minimizing the largest group-level risk, which can be formulated as

$$\min_{\theta \in \Theta} \max_{z,z' \in \mathcal{Z}} \hat{R}^\ell_{z,z'}(\theta). \quad (6)$$

The above problem is again non-differntiable. We introduce an additional variable $\lambda \in \mathbb{R}^4$ and relax the problem as a

Algorithm 1: *MinimaxFairAUC*

1: **Inputs:** Training set S with label Y and protected attribute Z, model f_θ, number of iterations T, batch size m, learning rates $\{\eta_\theta, \eta_\lambda\}$
2: Initialize $\theta_0 \in \Theta$ and $\lambda_0 \in \Lambda$ with $\lambda_{z,z'} = \frac{n^{z+}n^{z'-}}{n^+n^-}$ for all $z, z' \in \mathcal{Z}$
3: **for** $t = 1$ to $T-1$ **do**
4: $\quad B_t = \texttt{StratifiedSampler}_m(S; Y, Z)$
5: $\quad \theta_t = \theta_{t-1} - \eta_\theta \lambda_{t-1}^\top \nabla \hat{R}^\ell(\theta_{t-1}; B_t)$
6: $\quad \gamma_t = \lambda_{t-1} \exp(\eta_\lambda \hat{R}^\ell(\theta_{t-1}; B_t))$
7: $\quad \lambda_t = \gamma_t / \|\gamma_t\|_1$
8: **end for**
9: **Outputs:** $\theta_\tau \sim \texttt{Unif}(\{\theta_t\}_{t=1}^T)$

zero-sum game between two players θ and λ. At each round, player θ intends to find a better position to minimizes the weighted risk, while player λ assigns weights to each group that maximizes the weighted risk. Formally speaking, we arrive at the following minimax problem

$$\min_{\theta\in\Theta} \max_{\lambda\in\Lambda} F(\theta, \lambda) = \lambda^\top \hat{R}^\ell(\theta) = \sum_{z,z'\in\mathcal{Z}} \lambda_{z,z'} \hat{R}^\ell_{z,z'}(\theta), \quad (7)$$

where $\Lambda = \{\lambda \in \mathbb{R}^4 | \sum_{z,z'\in\mathcal{Z}} \lambda_{z,z'} = 1, \lambda_{z,z'} \geq 0\}$ is a 2×2-dimensional simplex.

Efficient Optimization Algorithm with Convergence Guarantee

In this section, we introduce a stochastic gradient method summarized in Algorithm 1 to solve the minimax optimization problem (7).

Let us illustrate Algorithm 1 in detail. At Line 4, the `StratifiedSampler` operator randomly sub-samples mini-batch by dividing the training set S into strata based on the label and the group attribute. We note that this subsampling scheme will later guarantee the stochastic gradients being unbiased estimators of the full sample gradients, which is essential for the stochastic optimization analysis. We summarize this result in the following proposition.

Proposition 1. *For any fixed $\theta \in \Theta$, let $B \subset S$ be given by `StratifiedSampler`. The following statement holds*

$$\mathbb{E}_B[\hat{R}^\ell(\theta; B)] = \hat{R}^\ell(\theta; S), \mathbb{E}_B[\nabla \hat{R}^\ell(\theta; B)] = \nabla \hat{R}^\ell(\theta; S).$$

The detailed description of `StratifiedSampler` and the proof of Proposition 1 are deferred to Appendix B. While a uniform sampler over the full dataset S also can be shown to construct unbiased estimators, it can miss certain groups or labels during training, especially when the dataset is imbalanced in groups or labels (Shekhar et al. 2021).

At Line 5, the player θ performs stochastic gradient descent with learning rate η_θ. Here we slightly abuse the notation and use λ_t to denote the t-th iteration of $\lambda \in \mathbb{R}^4$. At Line 6-7, the player λ performs the well-known exponential weight updates. We note that the exponential weight update is equivalent to mirror ascent when the mirror map $\Phi : \Lambda \to \mathbb{R}$ is given by the negative entropy (Bubeck et al. 2015), i.e. $\Phi(\lambda) = \sum_{z,z'\in\mathcal{Z}} \lambda_{z,z'} \log(\lambda_{z,z'})$. The exact update is given by

$$\nabla\Phi(\gamma_t) = \nabla\Phi(\lambda_{t-1}) + \eta_\lambda \hat{R}^\ell(\theta_{t-1}; B_t),$$
$$\lambda_t = \arg\min_{\lambda\in\Lambda} D_\Phi(\lambda \parallel \gamma_t),$$

where $D_\Phi(\lambda \parallel \lambda') = \Phi(\lambda) - \Phi(\lambda') - \nabla\Phi(\lambda')^\top(\lambda - \lambda')$ is the Bregman divergence associated with Φ. Therefore, Algorithm 1 can be viewed as a stochastic gradient descent mirror ascent method. Next we introduce some standard assumptions for our convergence analysis.

Assumption 1. For any $\theta \in \Theta$ and $\lambda \in \Lambda$, the gradients of F are bounded by G_θ and G_λ respectively, i.e. $\|\lambda^\top \nabla \hat{R}^\ell(\theta; S)\|_2 \leq G_\theta$, and $\|\hat{R}^\ell(\theta; S)\|_\infty \leq G_\lambda$.

Assumption 2. The objective F is L_θ and L_λ smooth respectively, i.e. $\|\lambda^\top \nabla \hat{R}^\ell(\theta; S) - \lambda^\top \nabla \hat{R}^\ell(\theta'; S)\|_2 \leq L_\theta \|\theta - \theta'\|_2$ and $\|\hat{R}^\ell(\theta; S) - \hat{R}^\ell(\theta'; S)\|_\infty \leq L_\lambda \|\lambda - \lambda'\|_1$ for any $\theta, \theta' \in \Theta$ and $\lambda, \lambda' \in \Lambda$.

Assumption 3. For any fixed $\theta \in \Theta, \lambda \in \Lambda$ and randomly sampled pair $\xi = \{(x, y, z), (x', y', z')\} \subset S$, the variances of the stochastic gradients of the function $F(\cdot, \cdot; \xi)$ are bounded by σ_θ^2 and σ_λ^2 respectively, i.e.

$$\mathbb{E}_\xi[\|\lambda^\top \nabla \hat{R}^\ell(\theta; \xi) - \lambda^\top \nabla \hat{R}^\ell(\theta; S)\|_2^2] \leq \sigma_\theta^2$$

and

$$\mathbb{E}_\xi[\|\hat{R}^\ell(\theta; \xi) - \hat{R}^\ell(\theta; S)\|_\infty^2] \leq \sigma_\lambda^2.$$

To ease the notation, we denote $G = \max\{G_\theta, G_\lambda\}$, $L = \max\{L_\theta, L_\lambda\}$ and $\sigma = \max\{\sigma_\theta, \sigma_\lambda\}$. Now let $P(\theta) = \max_{\lambda\in\Lambda} F(\theta, \lambda)$. Its Moreau envelope $P_{1/2L}$ is defined as

$$P_{1/2L}(\omega) = \min_\theta P(\theta) + L\|\theta - \omega\|_2^2.$$

With the above setup, we are in position to present the convergence of Algorithm 1.

Theorem 2 (Informal)**.** *Suppose Assumption 1, 2 and 3 hold true. Then the output θ_τ of Algorithm 1 satisfies*

$$\mathbb{E}[\|\nabla P_{1/2L}(\theta_\tau)\|_2 \leq \epsilon(T, \eta_\theta, \eta_\lambda), \quad (8)$$

where $\epsilon(T, \eta_\theta, \eta_\lambda)$ is an absolute constant. In particular, to achieve some small $\epsilon = \epsilon(T, \eta_\theta, \eta_\lambda)$, one choose $\eta_\theta = \Theta(\epsilon^4)$, $\eta_\lambda = \Theta(\epsilon^2)$ and $T = \mathcal{O}(\epsilon^{-8})$. Furthermore, there exists $\hat{\theta} \in \Theta$ such that $\mathbb{E}[\|\hat{\theta} - \theta_\tau\|_2] \leq \epsilon/2L$ and it satisfies

$$\mathbb{E}[\min_{\xi\in\partial P(\hat{\theta})} \|\xi\|_2] \leq \epsilon. \quad (9)$$

The exact statement and detailed proof are deferred to Appendix C. The proof of Theorem 2 is non-trivial and consists of several intermediate lemmas. The main idea is to derive a telescoping upper bound on the error term

$$\Delta_t = \mathbb{E}_{B_t}[\max_\lambda F(\theta_t, \lambda) - F(\theta_t, \lambda_t)]$$

by utilizing the concavity of $F(\theta_t, \cdot)$ and the Pythagorean inequality w.r.t. the Bregman divergence D_Φ. It is worth noting that this result even holds for general nonconvex-concave problems over the simplex, which is of interest in its own right. We leave the study of the generalization error as future work. We end this section with several remarks on the implications of Theorem 2 and the comparison of Algorithm 1 with other minimax fairness algorithms.

Remark 1. Eq. (8) characterizes the local gradient convergence of the output θ_τ on the Moreau envelope $P_{1/2L}$. Such a convergence bound can be transferred to the objective at concern P in (9) (Davis and Drusvyatskiy 2019). We call P the primal objective as it has been taken out of the maximization on the dual variable $\lambda \in \Lambda$. In the ideal case, the group weights λ only represent the maximum group risk. Therefore the convergence on the primal objective is exactly seeking the (local) minimum of the maximum group risk, defined as the original problem (6). Furthermore, the primal objective also provides guidance on the stopping criterion in empirical evaluations, which is known to be vague for general minimax optimization. That is, to stop when the maximum group-level AUC risk is saturated.

Remark 2. One of the main focuses of minimax fairness algorithms is how to update the model parameter and the group weights. The most closely related algorithm by Diana et al. (2021) also applies the exponential weights update on the group weights. However, the model update in that algorithm requires the exact solution on the weighted group risks at each iteration. This is time-consuming, aside from being intractable in most cases. On the contrary, Algorithm 1 is iteration-efficient as it alternately updates the model parameter and the group weight based on the stochastic mini-batch. The stochastic gradient method in Martinez, Bertran, and Sapiro (2020) relieves the intractability, yet still require a computationally-expensive inner loop to update the model parameter. Moreover, since the algorithm is heuristic, the authors provide no convergence analysis. Mohri, Sivek, and Suresh (2019) also apply an efficient stochastic optimization algorithm by updating the group weights by stochastic gradient ascent. However, such an update is sub-optimal compared to our stochastic mirror ascent on simplex (Beck and Teboulle 2003). Furthermore, their convergence relies on a convexity assumption while our results hold for the general nonconvex model class. Finally, it is worth noting the above minimax fairness frameworks all focus on classification and confusion matrix-based metrics; our algorithm is the first for group-level AUCs.

Empirical Evaluations

In this section, we evaluate the performance of Algorithm 1 in terms of utility and fairness.

Datasets Information & Feature Engineering. We evaluate our algorithms on four datasets that have been commonly used in the fair machine learning literature (Zafar et al. 2017; Donini et al. 2018). We apply one-hot encoding to all categorical attributes and normalize all numerical attributes with zero mean and unit variance. The summary statistics of the datasets are given in Appendix E.

- The `Adult` dataset is based on US census data and consists in predicting whether income exceeds $\$50K$ a year. The sensitive attribute is the gender of the individual, i.e. female ($Z = a$) or male ($Z = b$).
- The `Bank` dataset consists in predicting whether a client will subscribe to a term deposit. The sensitive attribute is the age of the individual: $Z = a$ when the age is less than 25 or over 60 and $Z = b$ otherwise.
- The `Compas` dataset consists in predicting recidivism of convicts in the US. The sensitive variable is the race of the individual, precisely $Z = a$ if the individual is categorized as non Caucasian and $Z = b$ if the individual is categorized as Caucasian.
- The `Default` dataset (Yeh and Lien 2009) investigates customers' default payments. The goal is to predict whether a customer will face the default situation in the next month or not. The sensitive attribute is the gender of the individual, i.e. female ($Z = a$) or male ($Z = b$).

Choice of Models and Loss Functions. To parameterize the family of scoring functions, we used a simple fully-connected neural network of 2 hidden layers with ReLU activation and batch normalization. The detail of the network is deferred in Appendix E. The surrogate loss function ℓ is chosen as the logistic loss, i.e. $\ell : s \mapsto \log(1 + \exp(-s))$ since it is statistically consistent (Gao and Zhou 2015) of the original 0/1 loss.

Baselines. We compare our framework with three in-processing baselines that have been proposed to 1) achieve fair group-level AUC scores; or 2) address unfair impact in terms of the Rawlsian principle.

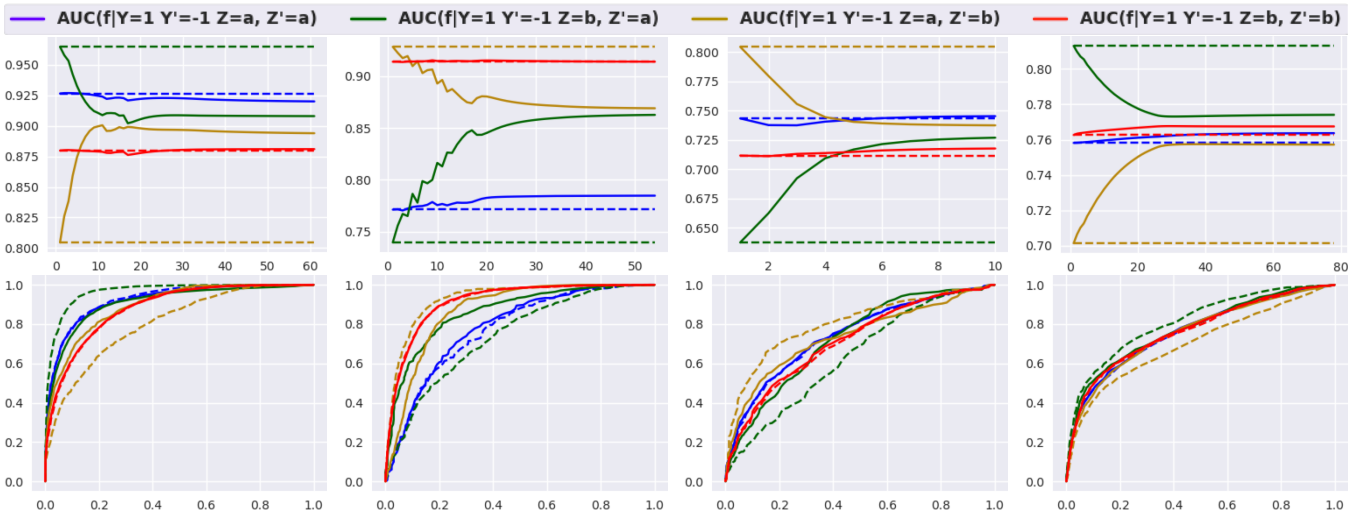
- The `AUCMax` algorithm conducts AUC maximization on the full dataset without differentiating any groups. It updates the model parameter by mini-batch SGD.
- The `MinimaxFair` algorithm by Diana et al. (2021) aims to minimize the maximum group-level misclassification error. We replace the intractible model update by 10 epochs of SGD.
- There are regularization methods (Beutel et al. 2019) or constrained optimization methods (Narasimhan et al. 2020) targeting inter-group AUC fairness. We pick the one by Vogel, Bellet, and Clémençon (2021) as a representative since the authors considered the same datasets. Vogel, Bellet, and Clémençon learn fair AUC scores by regularization on the difference of inter-group AUCs, which we refer to as `InterFairAUC`.
- We also consider AUC maximization under the constraint of (3) as an alternative to our AUC fairness criterion under the name `EqualAUC`. See Appendix D for details.

Implementation Details[3] We partition the datasets to training, validation and testing in the ratio 60%:20%:20%. The batch size $|B|$, initial stepsizes $\eta_0^\theta, \eta_0^\lambda$ and other hyper-parameters are chosen based on the validation set. For Algorithm 1, early stopping is implemented based on the maximum group loss over the validation set. We repeat each experiment in 25 runs across different random seeds and report the average result on the testing set. More details are given in Appendix E.

Results on Real Datasets

We first investigate the convergence property of Algorithm 1. We initialize the model parameters θ_0 as the one trained from `AUCMax` to better illustrate the improvement over

[3] https://github.com/zhenhuan-yang/MinimaxFairAUC.



(a) *Adult* (b) *Bank* (c) *COMPAS* (d) *Default*

Figure 2: Convergence plots on training set (upper half) and ROC plots on test set (lower half) of Algorithm 1 (solid curves) versus AUCMax (dashed curves). For convergence plots, the x-axis indicates the number of epochs, and the y-axis indicates the AUC score. For ROC plots, the x-axis indicates the FPR and the y-axis indicates the TPR.

the baseline. As shown in Figure 2 upper half, the inter-group AUC unfairness is prevalent in all four datasets, and intra-group AUC unfairness exists in Adult and Bank. When there is only inter-group AUC unfairness, Algorithm 1 mainly lifts the lowest inter-group AUC score. This can potentially benefit intra-group AUC (cf. Compas and Default.) When both intra-group and inter-group AUC unfairness exist, Algorithm 1 alternately lifts the lowest AUCs from inter-group then from intra-group, leading to a smaller gap in both of them separately. We next investigate the generalization/test performance of Algorithm 1. In Figure 2 lower half, the ROC curves of Algorithm 1 narrows the gaps of AUCMax towards the middle.

We next develop a quantitative understanding of Algorithm 1 on its utility and fairness performance against other baselines in Table 1. MinimaxFair struggles on both metrics. This is due to the objective differences: MinimaxFair is aiming at classification and the disparity in accuracy. Therefore it may overlook the disparity caused by group-level AUCs. InterFairAUC does not perform too well on fairness as expected. This is because the regularization only focuses on the difference of inter-group AUCs and eventually the unfairness of intra-group AUCs will dominate the min/max ratio, especially on Adult and Bank where original gaps of intra-group AUCs are large. It only increase the overall AUC on Bank because the dataset is imbalanced. EqualAUC achieves competitive min/max ratios on all datasets. However it suffers from a large utility drop compared to the non-fairness intended baseline AUCMax. On the contrary, Algorithm 1 preserves the largest overall AUC scores. It achieves over 99% on Adult and Bank compared to AUCMax, and achieves even higher score on Compas and Default. This phenomenon is consistent with our argument in (5) that the proposed minimax objective does not conflict with AUC maximization. The further improvement may due to the initialization via AUCMax. Using fairness aware retraining to boost utility has been observed in the literature recently, especially when the motivation is seeking minimax fairness or its analog (Globus-Harris, Kearns, and Roth 2022). In the meantime, for the fairness measurement, Algorithm 1 provides a competitive min/max ratio versus EqualAUC. Even on Bank, EqualAUC is not significantly more fair (p value > 0.01) than Algorithm 1 while Algorithm 1 has a better overall AUC metric (p value < 0.001). On Compas, Algorithm 1 even reaches the best overall score and the ratio simultaneously. Therefore it can be concluded that Algorithm 1 has advantages over EqualAUC in the sense of better trade-offs between overall AUC and fairness.

Synthetic Datasets Experiments

In this subsection, we design synthetic datasets to further understand the effectiveness of minimax AUC Algorithm 1 versus the unfair baseline. In particular, we generate two dimensional data points from different Gaussian distributions conditioned on the class label and the group attribute. See Appendix E for more details. The generated data points are shown in the left most plot in Figure 3. We generate the same number of samples for each label and group partition so that the priors in (4) can be ignored and all group-level AUCs contribute the same towards the overall AUC. Furthermore, we intentionally design the overlap between the positive and negative samples for group a so that the scoring function faces difficulty differentiating them. As we see from the middle left plot in Figure 3, the baseline AUCMax learns almost nothing from group a as the positive and negative KDEs are largely overlapped. Yet this fact does not stop the algorithm from maximizing the overall AUC as it keeps the negative KDE from group b low. Therefore all three $AUC_{a,b}, AUC_{b,a}, AUC_{b,b}$ are optimized except $AUC_{a,a}$.

Algorithm \ Metric	Adult		Bank		Compas		Default	
	Overall	Min/Max	Overall	Min/Max	Overall	Min/Max	Overall	Min/Max
AUCMax	$\mathbf{.902 \pm .002}$	$.823 \pm .005$	$\underline{.910 \pm .002}$	$.780 \pm .018$	$.732 \pm .004$	$.779 \pm .041$	$.763 \pm .005$	$.871 \pm .017$
MinimaxFair	$.894 \pm .007$	$.905 \pm .010$	$.885 \pm .003$	$.827 \pm .004$	$.730 \pm .001$	$.913 \pm .029$	$.753 \pm .002$	$.909 \pm .021$
InterFairAUC	$.894 \pm .004$	$.950 \pm .003$	$\mathbf{.912 \pm .001}$	$.836 \pm .018$	$\underline{.738 \pm .003}$	$.939 \pm .014$	$\underline{.763 \pm .003}$	$.952 \pm .024$
EqualAUC	$.886 \pm .003$	$\underline{.953 \pm .004}$	$.866 \pm .005$	$\mathbf{.891 \pm .025}$	$.731 \pm .003$	$\underline{.956 \pm .012}$	$.761 \pm .002$	$\mathbf{.972 \pm .020}$
Algorithm 1	$\underline{.901 \pm .004}$	$\mathbf{.953 \pm .002}$	$.907 \pm .004$	$\underline{.858 \pm .014}$	$\mathbf{.741 \pm .004}$	$\mathbf{.961 \pm .012}$	$\mathbf{.767 \pm .002}$	$\underline{.968 \pm .013}$

Table 1: Comparison of Algorithm 1 versus baselines. 'Overall' is the AUC score on the full dataset, measuring the utility. 'Min/Max' is the minimum group-level AUC score over the maximum one, measuring the fairness. The numbers are reported as 'Mean ± Standard Deviation'. Best results at each column are highlighted in bold. Second best are highlighted in underline.

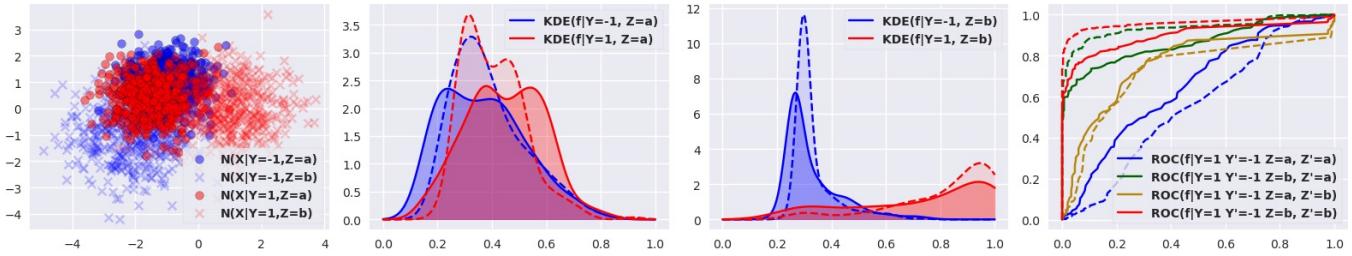
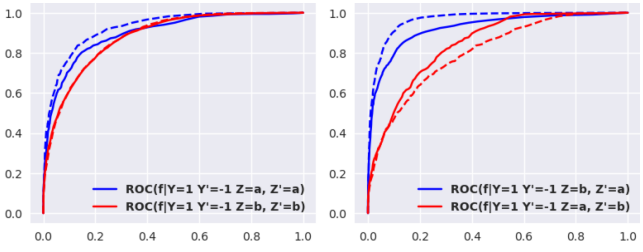
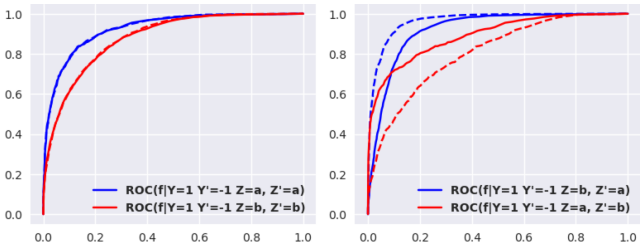


Figure 3: Experiments on synthetic datasets (left) of Algorithm 1 (solid curves) versus AUCMax (dashed curves).



(a) *Intra-group AUCs only*



(b) *Inter-group AUCs only*

Figure 4: Ablation study of Algorithm 1 (solid curves) versus AUCMax (dashed curves) on Adult. Algorithm 1 solves Eq. (7) on the 2-dimensional simplex Λ with $z = z'$ (top) or $z \neq z'$ (bottom) only.

However this is severely unfair for the positive sample from group a. Algorithm 1 mitigates this issue by separating the positive and negative KDEs from group a to the correct direction. It also pushes the negative KDE from group b to be lower so that positives from group a maintain higher scores than negatives from group b. This is based on the sacrifice of the positives' scores from group b. We consider such small trade-off as acceptable since the separation between these two conditional distributions is still visually clear.

Ablation Studies

In this subsection, we apply Algorithm 1 on the maximin AUC problem (5) with only intra-group AUCs or inter-group AUCs. We only report the results on Adult due to space limits. See Appendix E for results on other datasets. In Figure 4(a), when intra-group AUCs are the only objective, Algorithm 1 slightly mitigates the intra-group AUC unfairness, yet it can potentially help mitigate the inter-group AUC unfairness as well. However, this improvement is very limited compared to lower half in Figure 2(a), where the unfairness is almost eliminated. In Figure 4(b), when inter-group AUCs are the only objective, inter-group fairness is more or less achieved with crossing of two curves, yet it is not guaranteed to mitigate the unfairness within intra-group AUCs.

Conclusion

In this paper, we propose a minimax learning framework for AUC maximization with fairness concern. Our framework addresses both intra-group and inter-group AUC unfairness as well as the discrepancy in between. Based on this framework, we design Algorithm 1: an efficient algorithm with stochastic gradient descent on the model and mirror ascent on the group weights. We provide a non-trivial analysis and show that Algorithm 1 converges to the optimal solution in terms of the minimum group-level AUC in the nonconvex-concave setting. We conduct numerical experiments on both synthetic and real-world datasets to validate its utility and the fairness performance. One future direction is to consider fairness metrics involving group-level partial AUCs (Narasimhan and Agarwal 2017; Yang et al. 2021b).

Acknowledgements

This work is supported by SUNY-IBM AI Research Alliance and NSF grants (IIS-2103450, IIS-2110546 and DMS-2110836).

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach To Fair Classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Beck, A.; and Teboulle, M. 2003. Mirror Descent And Nonlinear Projected Subgradient Methods For Convex Optimization. *Operations Research Letters*, 31(3): 167–175.

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Wei, L.; Wu, Y.; Heldt, L.; Zhao, Z.; Hong, L.; Chi, E. H.; et al. 2019. Fairness In Recommendation Ranking Through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2212–2220.

Bubeck, S.; et al. 2015. Convex Optimization: Algorithms And Complexity. *Foundations and Trends® in Machine Learning*, 8(3-4): 231–357.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building Classifiers With Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 13–18. IEEE.

Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; and Varshney, K. R. 2017. Optimized Pre-processing For Discrimination Prevention. *Advances in neural information processing systems*, 30.

Chouldechova, A.; and Roth, A. 2020. A Snapshot Of The Frontiers Of Fairness In Machine Learning. *Communications of the ACM*, 63(5): 82–89.

Cortes, C.; and Mohri, M. 2003. AUC Optimization vs. Error Rate Minimization. In *NIPS*.

Davis, D.; and Drusvyatskiy, D. 2019. Stochastic Model-based Minimization Of Weakly Convex Functions. *SIAM Journal on Optimization*, 29(1): 207–239.

Diana, E.; Gill, W.; Kearns, M.; Kenthapadi, K.; and Roth, A. 2021. Minimax Group Fairness: Algorithms And Experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 66–76.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring And Mitigating Unintended Bias In Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J. S.; and Pontil, M. 2018. Empirical Risk Minimization Under Fairness Constraints. *Advances in Neural Information Processing Systems*, 31.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying And Removing Disparate Impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Gao, W.; Jin, R.; Zhu, S.; and Zhou, Z.-H. 2013. One-pass AUC Optimization. In *International conference on machine learning*, 906–914.

Gao, W.; and Zhou, Z.-H. 2015. On The Consistency Of AUC Pairwise Optimization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Globus-Harris, I.; Kearns, M.; and Roth, A. 2022. An Algorithmic Framework For Bias Bounties. *arXiv preprint arXiv:2201.10408*.

Hanley, J. A.; and McNeil, B. J. 1982. The Meaning And Use Of The Area Under A Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143(1): 29–36.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality Of Opportunity In Supervised Learning. *Advances in neural information processing systems*, 29.

Kallus, N.; and Zhou, A. 2019. The Fairness Of Risk Scores Beyond Classification: Bipartite Ranking And The Xauc Metric. *Advances in neural information processing systems*, 32.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware Learning Through Regularization Approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 643–650. IEEE.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. 2020. Fairness Without Demographics Through Adversarially Reweighted Learning. *Advances in neural information processing systems*, 33: 728–740.

Lei, Y.; and Ying, Y. 2021. Stochastic Proximal AUC Maximization. *J. Mach. Learn. Res.*, 22(61): 1–45.

Liu, M.; Yuan, Z.; Ying, Y.; and Yang, T. 2020. Stochastic AUC Maximization With Deep Neural Networks. In *8th International Conference on Learning Representations (ICLR)*.

Lohia, P. K.; Ramamurthy, K. N.; Bhide, M.; Saha, D.; Varshney, K. R.; and Puri, R. 2019. Bias Mitigation Post-processing For Individual And Group Fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, 2847–2851. IEEE.

Martinez, N.; Bertran, M.; and Sapiro, G. 2020. Minimax Pareto Fairness: A Multi Objective Perspective. In *International Conference on Machine Learning*, 6755–6764. PMLR.

Mohri, M.; Sivek, G.; and Suresh, A. T. 2019. Agnostic Federated Learning. In *International Conference on Machine Learning*, 4615–4625. PMLR.

Narasimhan, H.; and Agarwal, S. 2017. Support Vector Algorithms For Optimizing The Partial Area Under The ROC Curve. *Neural computation*, 29(7): 1919–1963.

Narasimhan, H.; Cotter, A.; Gupta, M.; and Wang, S. 2020. Pairwise Fairness For Ranking And Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(04), 5248–5255.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness And Calibration. *Advances in neural information processing systems*, 30.

Rawls, J. 2001. *Justice As Fairness: A Restatement*. Harvard University Press.

Shekhar, S.; Fields, G.; Ghavamzadeh, M.; and Javidi, T. 2021. Adaptive Sampling For Minimax Fair Classification. *Advances in Neural Information Processing Systems*, 34: 24535–24544.

Ustun, B.; Liu, Y.; and Parkes, D. 2019. Fairness Without Harm: Decoupled Classifiers With Preference Guarantees. In *International Conference on Machine Learning*, 6373–6382. PMLR.

Vogel, R.; Bellet, A.; and Clémençon, S. 2021. Learning Fair Scoring Functions: Bipartite Ranking Under ROC-based Fairness Constraints. In *International Conference on Artificial Intelligence and Statistics*, 784–792. PMLR.

Yang, T.; and Ying, Y. 2022. AUC Maximization In The Era Of Big Data And AI: A Survey. *ACM Computing Surveys (CSUR)*.

Yang, Z.; Xu, Q.; Bao, S.; Cao, X.; and Huang, Q. 2021a. Learning with Multiclass AUC: Theory And Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, Z.; Xu, Q.; Bao, S.; He, Y.; Cao, X.; and Huang, Q. 2021b. When All We Need Is A Piece Of The Pie: A Generic Framework For Optimizing Two-way Partial AUC. In *International Conference on Machine Learning*, 11820–11829. PMLR.

Yeh, I.-C.; and Lien, C.-h. 2009. The Comparisons Of Data Mining Techniques For The Predictive Accuracy Of Probability Of Default Of Credit Card Clients. *Expert systems with applications*, 36(2): 2473–2480.

Ying, Y.; Wen, L.; and Lyu, S. 2016. Stochastic Online AUC Maximization. *Advances in neural information processing systems*, 29.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zhao, P.; Hoi, S. C. H.; Jin, R.; and Yang, T. 2011. Online AUC Maximization. In *ICML*, 233–240.