Learning Fair Scoring Functions: Bipartite Ranking under ROC-based Fairness Constraints

Robin Vogel
LTCI, Télécom Paris,
Institut Polytechnique de Paris, France

Aurélien Bellet
INRIA, France

Stephan Clémençon
LTCI, Télécom Paris,
Institut Polytechnique de Paris, France

Abstract

Many applications of AI involve *scoring* individuals using a learned function of their attributes. These predictive risk scores are then used to take decisions based on whether the score exceeds a certain threshold, which may vary depending on the context. The level of delegation granted to such systems in critical applications like credit lending and medical diagnosis will heavily depend on how questions of *fairness* can be answered. In this paper, we study fairness for the problem of learning scoring functions from binary labeled data, a classic learning task known as *bipartite ranking*. We argue that the functional nature of the ROC curve, the gold standard measure of ranking accuracy in this context, leads to several ways of formulating fairness constraints. We introduce general families of fairness definitions based on the AUC and on ROC curves, and show that our ROC-based constraints can be instantiated such that classifiers obtained by thresholding the scoring function satisfy classification fairness for a desired range of thresholds. We establish generalization bounds for scoring functions learned under such constraints, design practical learning algorithms and show the relevance our approach with numerical experiments on real and synthetic data.

1 INTRODUCTION

With the availability of data at ever finer granularity and the development of technological bricks to efficiently store and process this data, the infatuation with machine learning (ML) and artificial intelligence (AI) is spreading to nearly all fields (science, transportation, energy, medicine, security, banking, insurance, commerce...). Expectations are high. There is no denying the opportunities, and we can rightfully hope for an increasing number of successful deployments in the near future. However, AI will keep its promises only if certain issues are addressed. In particular, ML systems that make significant decisions for humans, regarding for instance credit lending in the banking sector (Chen, 2018), diagnosis in medicine (Deo, 2015) or recidivism prediction in criminal justice (Rudin et al., 2018), should guarantee that they do not penalize certain groups of individuals.

Hence, stimulated by the societal expectations, notions of *fairness* in ML as well guarantees that they can be fulfilled by models trained under appropriate constraints have recently been the subject of a good deal of attention in the literature, see *e.g.* (Dwork et al., 2012; Kleinberg et al., 2017) among others. Fairness constraints are generally modeled by means of a (qualitative) *sensitive variable*, indicating membership to a certain group (*e.g.*, ethnicity, gender). The vast majority of the work dedicated to algorithmic fairness in ML focuses on binary classification. In this context, fairness constraints force classifiers to have similar true positive rates (or false positive rates) across sensitive groups. For instance, Hardt et al. (2016); Pleiss et al. (2017) propose to modify a pre-trained classifier in order to fulfill such constraints without deteriorating too much the classification performance. Other work incorporates fairness constraints in the learning stage (see *e.g.*, Agarwal et al., 2018; Woodworth et al., 2017; Zafar et al., 2017a,b, 2019; Menon and Williamson, 2018; Bechavod and Ligett, 2017). In addition to algorithms, statistical guarantees (in the form of generalization bounds) are crucial for fair ML, as they ensure that the desired fairness constraint will be met at deployment. Such learning guarantees have been established by Donini et al. (2018) for the case of fair classification.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Many real-world problems are however not concerned with learning a binary classifier but rather aim to learn a *scoring function*. This statistical learning problem is known as *bipartite ranking* and covers in particular tasks such as credit scoring in banking, pathology scoring in medicine or recidivism scoring in criminal justice, for which fairness is a major concern (Kallus and Zhou, 2019). While it can be formulated in the same probabilistic framework as binary classification, bipartite ranking is not a local learning problem: the goal is not to guess whether a binary label Y is positive or negative from an input observation X but to rank any collection of observations $X_1, \ldots, X_n$ by means of a scoring function $s : \mathcal{X} \to \mathbb{R}$ so that observations with positive labels are ranked higher with large probability. Due to the global nature of the task, evaluating the performance is itself a challenge. The gold standard measure, the ROC curve, is functional: it is the PP-plot of the false positive rate (FPR) *vs* the true positive rate (TPR), and the higher the curve, the more accurate the ranking induced by s. Sup-norm optimization of the ROC curve has been investigated by Clémençon and Vayatis (2009, 2010), while most of the literature focuses on the maximization of scalar summaries of the ROC curve such as the AUC criterion (Agarwal et al., 2005; Clémençon et al., 2008; Zhao et al., 2011) or alternative measures (Rudin, 2006; Clémençon and Vayatis, 2007; Menon and Williamson, 2016).

A key advantage of learning a scoring function over learning a classifier is the flexibility in thresholding the scores so as to obtain false/true positive rates that fit the particular operational constraints in which the decision is taken. A natural fairness requirement in this context is that a fair scoring function should lead to fair decisions *for all thresholds of interest*. To help fix ideas and grasp the methodological challenge, we describe below a concrete example to motivate our work.

Example 1 (Credit-risk screening)**.** *A bank grants a loan to a client with socio-economic features X if his/her score $s(X)$ is above a certain threshold t. As the degree of risk aversion of the bank may vary, the precise deployment threshold t is unknown when choosing the scoring function s, although the bank is generally interested in regimes where the probability of default is sufficiently small (low FPR). The bank would like to design a scoring function that ranks higher the clients that are more likely to repay the loan (*ranking performance*), while ensuring that any threshold in the regime of interest will lead to similar false negative rates across sensitive groups (*fairness constraint*).*

Contributions. In this work, we provide a thorough study of fairness in bipartite ranking. Our starting point is a number of fairness measures introduced independently in recent papers from different communities (Borkan et al., 2019; Beutel et al., 2019; Kallus and Zhou, 2019). We first show that these are special cases of a general family of fairness constraints based on the AUC, that we precisely characterize. We then argue that, because it is defined from scalar summaries of functional curves, AUC-based fairness is oblivious to potentially large disparities between groups at particular locations of the score distribution (see Fig. 1, bottom left). As a consequence, they fail to address use-cases where fairness is needed at specific thresholds (as in Example 1). To overcome these limitations, we introduce a novel functional view of fairness based on ROC curves. These richer *pointwise* ROC-*based constraints* can be instantiated to align group-wise score distributions at specific functional points (see Fig. 1, bottom right) and thereby ensure that classifiers obtained by thresholding the scoring function satisfy classification fairness for a certain range of thresholds, as desired in cases like Example 1.

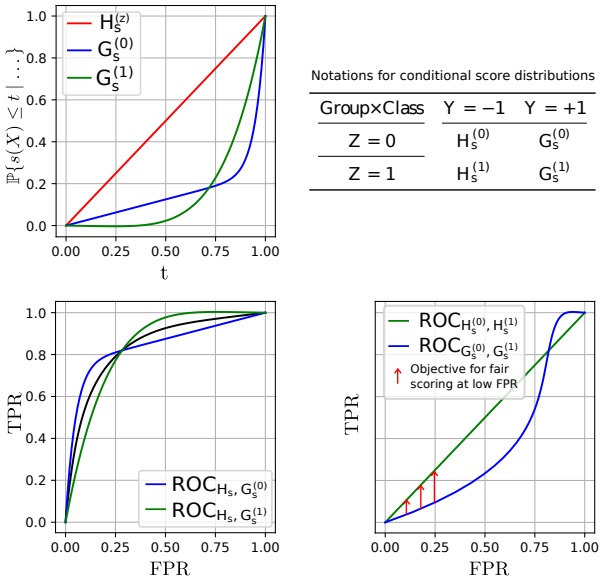

Figure 1: Illustrating the limitations of AUC-based fairness. Here, the group-wise positive/negative distributions (top) satisfy $\mathrm{AUC}_{H_s, G_s^{(0)}} = \mathrm{AUC}_{H_s, G_s^{(1)}}$ (bottom left), but yield very different TPR's at low FPR's (bottom left). Our new ROC-based constraints can align scores distributions where it matters, *e.g.* for low FPR's as in Example 1 (bottom right).

Based on the above, we then introduce empirical risk minimization formulations for learning fair scoring functions under both AUC and ROC-based fairness constraints and establish the first generalization bounds for fair bipartite ranking. Due to the complex nature of the ranking measures, our proof techniques largely differ from the classification results of Donini et al. (2018) as they require non standard tech-

nical tools (*e.g.* to control deviations of ratios of U-statistics). In addition to our conceptual contributions and theoretical analysis, we propose efficient training algorithms based on gradient descent and illustrate the practical relevance of our approach on synthetic and real datasets.

Outline. The paper is organized as follows. Section 2 reviews bipartite ranking as well as existing fairness notions for classification and ranking. Section 3 studies AUC-based fairness constraints and propose richer ROC-based constraints. In Section 4, we formulate the problem of fair scoring under both AUC and ROC-based fairness constraints and prove statistical learning guarantees. Section 5 presents numerical experiments, and we conclude in Section 6. Due to space limitations, some technical details and additional experiments are postponed to the supplementary.

2 BACKGROUND & RELATED WORK

In this section, we introduce the main concepts involved in the subsequent analysis and review related work. Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$ and the pseudo-inverse of any cumulative distribution function (c.d.f.) function $F : \mathbb{R} \to [0,1]$ by $F^{-1}(u) = \inf\{t \in \mathbb{R} : F(t) \geqslant u\}$.

2.1 Probabilistic Framework

Let X and Y be two random variables: Y denotes the binary output label (taking values in $\{-1,+1\}$) and X denotes the input features, taking values in a space $\mathcal{X} \subset \mathbb{R}^d$ with $d \geqslant 1$ and modeling some information hopefully useful to predict Y. For convenience, we introduce the proportion of positive instances $p := \mathbb{P}\{Y = +1\}$, as well as G and H, the conditional distributions of X given $Y = +1$ and $Y = -1$ respectively. The joint distribution of (X, Y) is fully determined by the triplet (p, G, H). Another way to specify the distribution of (X, Y) is through the pair (μ, η) where μ denotes the marginal distribution of X and η the function $\eta(x) := \mathbb{P}\{Y = +1 \mid X = x\}$. With these notations, one may write $\eta(x) = p(dG/dH)(x)/(1 - p + p(dG/dH)(x))$ and $\mu = pG + (1-p)H$.

In the context of fairness, we consider a third random variable Z which denotes the sensitive attribute taking values in $\{0,1\}$. The pair (X, Y) is said to belong to salient group 0 (resp. 1) when $Z = 0$ (resp. $Z = 1$). The distribution of the triplet (X, Y, Z) can be expressed as a mixture of the distributions of $X, Y|Z = z$. Following the conventions described above, we introduce the quantities $p_z, G^{(z)}, H^{(z)}$ as well as $\mu^{(z)}, \eta^{(z)}$. For instance, $p_0 = \mathbb{P}\{Y = +1|Z = 0\}$ and the distribution of $X|Y = +1, Z = 0$ is written $G^{(0)}$, *i.e.* for $A \subset \mathcal{X}$, $G^{(0)}(A) = \mathbb{P}\{X \in A|Y = +1, Z = 0\}$. We denote the probability of belonging to group z by $q_z := \mathbb{P}\{Z = z\}$, with $q_0 = 1 - q_1$.

2.2 Bipartite Ranking

The goal of bipartite ranking is to learn an order relationship on $\mathcal{X}$ for which positive instances are ranked higher than negative ones with high probability. This order is defined by transporting the natural order on the real line to the feature space through a scoring function $s : \mathcal{X} \to \mathbb{R}$. Given a distribution F over $\mathcal{X}$ and a scoring function s, we denote by F_s the cumulative distribution function of $s(X)$ when X follows F. Specifically:

$$G_s(t) := \mathbb{P}\{s(X) \leqslant t \mid Y = +1\} = G(s(X) \leqslant t),$$
$$H_s(t) := \mathbb{P}\{s(X) \leqslant t \mid Y = -1\} = H(s(X) \leqslant t).$$

ROC analysis. ROC curves are widely used to visualize the dissimilarity between two real-valued distributions in many applications, *e.g.* anomaly detection, medical diagnosis, information retrieval.

Definition 1 (ROC curve)**.** *Let g and h be two cumulative distribution functions on $\mathbb{R}$. The* ROC *curve related to g and h is the graph of the mapping:*

$$\mathrm{ROC}_{h,g} : \alpha \in [0,1] \mapsto 1 - g \circ h^{-1}(1 - \alpha).$$

When g and h are continuous, it can alternatively be defined as the parametric curve $t \in \mathbb{R} \mapsto (1 - h(t), 1 - g(t))$.

The classic area under the ROC curve (AUC) criterion is a scalar summary of the functional measure of dissimilarity ROC. Formally, we have:

$$\mathrm{AUC}_{h,g} := \int \mathrm{ROC}_{h,g}(\alpha) d\alpha = \mathbb{P}\{S > S'\} + \frac{1}{2}\mathbb{P}\{S = S'\},$$

where S and S' denote independent random variables, whose c.d.f.'s are h and g respectively.

In bipartite ranking, one focuses on the ability of the scoring function s to separate positive from negative data. This is reflected by ROC_{H_s,G_s}, which gives the false positive rate *vs.* true positive rate of binary classifiers $g_{s,t} : x \mapsto 2 \cdot \mathbb{I}\{s(x) > t\} - 1$ obtained by thresholding s at all possible thresholds $t \in \mathbb{R}$. The global summary AUC_{H_s,G_s} serves as a standard performance measure (Clémençon et al., 2008).

Empirical estimates. In practice, the scoring function s is learned based on a training set $\{(X_i, Y_i)\}_{i=1}^n$ of n i.i.d. copies of the random pair (X, Y). Let n_+ and n_- be the number of positive and negative data points

respectively. We introduce $\widehat{G}_s$ and $\widehat{H}_s$, the empirical counterparts of G_s and H_s:

$$\widehat{G}_s(t) := (1/n_+) \sum_{i=1}^n \mathbb{I}\{Y_i = +1, s(X_i) \leqslant t\},$$
$$\widehat{H}_s(t) := (1/n_-) \sum_{i=1}^n \mathbb{I}\{Y_i = -1, s(X_i) \leqslant t\}.$$

Note that the denominators n_+ and n_- are sums of i.i.d. random (indicator) variables. For any two distributions F, F' over $\mathbb{R}$, we denote the empirical counterparts of $\mathrm{AUC}_{F,F'}$ and $\mathrm{ROC}_{F,F'}$ by $\widehat{\mathrm{AUC}}_{F,F'} := \mathrm{AUC}_{\widehat{F},\widehat{F'}}$ and $\widehat{\mathrm{ROC}}_{F,F'}(\cdot) := \mathrm{ROC}_{\widehat{F},\widehat{F'}}(\cdot)$ respectively. In particular, we have:

$$\widehat{\mathrm{AUC}}_{H_s,G_s} := \frac{1}{n_+ n_-} \sum_{i<j} K((s(X_i), Y_i), (s(X_j), Y_j)),$$

where $K((t,y),(t',y')) = \mathbb{I}\{(y-y')(t-t') > 0\} + \mathbb{I}\{y \neq y', t = t'\}/2$ for any $t, t' \in \mathbb{R}^2, y, y' \in \{-1,+1\}^2$. Empirical risk minimization for bipartite ranking typically consists in maximizing $\widehat{\mathrm{AUC}}_{H_s,G_s}$ over a class of scoring functions (see *e.g.* Clémençon et al., 2008; Zhao et al., 2011).

2.3 Fairness in Binary Classification

In binary classification, the goal is to learn a mapping function $g : \mathcal{X} \mapsto \{-1, +1\}$ that predicts the output label Y from the input random variable X as accurately as possible (as measured by an appropriate loss function). Any classifier g can be defined by its unique acceptance set $A_g := \{x \in \mathcal{X} \mid g(x) = +1\} \subset \mathcal{X}$.

Existing notions of fairness for binary classification (see Zafar et al., 2019, for a detailed treatment) aim to ensure that g makes similar predictions (or errors) for the two groups. We mention here the common fairness definitions that depend on the ground truth label Y. *Parity in mistreatment* requires that the proportion of errors is the same for the two groups:

$$M^{(0)}(g) = M^{(1)}(g), \tag{1}$$

where $M^{(z)}(g) := \mathbb{P}\{g(X) \neq Y \mid Z = z\}$. While this requirement is natural, it considers that all errors are equal: in particular, one can have a high false positive rate (FPR) $H^{(1)}(A_g)$ for one group and a high false negative rate (FNR) $G^{(0)}(A_g)$ for the other. This can be considered unfair when acceptance is an advantage, *e.g.* being granted a loan in Example 1). A solution is to consider *parity in false positive rates* and/or *parity in false negative rates*, which respectively write:

$$H^{(0)}(A_g) = H^{(1)}(A_g) \text{ and } G^{(0)}(A_g) = G^{(1)}(A_g). \tag{2}$$

Remark 1 (Connection to bipartite ranking)**.** *A score function $s : \mathcal{X} \to \mathbb{R}$ induces an infinite collection of binary classifiers $g_{s,t} : x \mapsto 2 \cdot \mathbb{I}\{s(x) > t\} - 1$. While one could fix a threshold $t \in \mathbb{R}$ in advance and enforce fairness on $g_{s,t}$, we are interested here in notions of fairness for the score function itself (see Example 1).*

2.4 Fairness in Ranking

Fairness for rankings has been mostly considered in the informational retrieval and recommender systems communities. Given a set of items with *known relevance scores*, they aim to extract a (partial) ranking that balances utility and notions of fairness at the group or individual level, or through a notion of exposure over several queries (Zehlike et al., 2017; Celis et al., 2018; Biega et al., 2018; Singh and Joachims, 2018). Singh and Joachims (2019) and Beutel et al. (2019) extend the above work to the *learning to rank* framework, where the task is to learn relevance scores and ranking policies from a certain number of observed *queries* that consist of query-item features and item relevance scores. This is fundamentally different from the bipartite ranking setting considered here.

AUC constraints. In a setting closer to ours, Kallus and Zhou (2019) introduce measures to quantify the fairness of a known scoring function on binary labeled data (they do not address learning). Their approach is based on the AUC, which can be seen as a measure of homogeneity between distributions (Clémençon et al., 2009). Similar definitions of fairness are also considered in (Beutel et al., 2019; Borkan et al., 2019).

Introduce $G_s^{(z)}$ (resp. $H_s^{(z)}$) as the c.d.f. of the score on the positives (resp. negatives) of group $z \in \{0,1\}$, *i.e.* $G_s^{(z)}(t) = G^{(z)}(s(X) \leqslant t)$ and $H_s^{(z)}(t) = H^{(z)}(s(X) \leqslant t)$, for any $t \in \mathbb{R}$. Precise examples of AUC-based fairness constraints include: 1) the *intra-group pairwise* AUC *fairness* (Beutel et al., 2019),

$$\mathrm{AUC}_{H_s^{(0)},G_s^{(0)}} = \mathrm{AUC}_{H_s^{(1)},G_s^{(1)}}, \tag{3}$$

which requires the ranking performance to be equal *within* groups, 2) the *Background Negative Subgroup Positive (BNSP)* AUC *fairness* (Borkan et al., 2019),

$$\mathrm{AUC}_{H_s,G_s^{(0)}} = \mathrm{AUC}_{H_s,G_s^{(1)}}, \tag{4}$$

which enforces that positive instances from either group have the same probability of being ranked higher than a negative example, 3) the *inter-group pairwise* AUC *fairness* (Kallus and Zhou, 2019),

$$\mathrm{AUC}_{H_s^{(0)},G_s^{(1)}} = \mathrm{AUC}_{H_s^{(1)},G_s^{(0)}}, \tag{5}$$

which imposes that the positives of a group can be distinguished from the negatives of the other group as effectively for both groups. Many more AUC-based fairness constraints are possible: we give examples (some of them novel) in the supplementary material.

3 FROM AUC TO ROC-BASED FAIRNESS CONSTRAINTS

In this section, we first provide a new general framework to characterize all relevant AUC constraints. We then highlight some limitations of AUC fairness constraints, which serve as motivation to introduce our richer *pointwise ROC-based fairness constraints*.

3.1 A Family of AUC Fairness Constraints

All proposed AUC-based fairness constraints in the literature follow a common structure, which we precisely characterize.

Denote by (e_1, e_2, e_3, e_4) the canonical basis of $\mathbb{R}^4$, as well as by $\mathbf{1}$ the constant vector $\mathbf{1} = \sum_{k=1}^4 e_k$. AUC constraints are expressed in the form of equalities of the AUC's between mixtures of the c.d.f.'s $D(s)$, with: $D(s) := (H_s^{(0)}, H_s^{(1)}, G_s^{(0)}, G_s^{(1)})^\top$. Formally, introducing the probability vectors $\alpha, \beta, \alpha', \beta' \in \mathcal{P}$ where $\mathcal{P} = \{v \mid v \in \mathbb{R}_+^4, \mathbf{1}^\top v = 1\}$, they write as:

$$\mathrm{AUC}_{\alpha^\top D(s), \beta^\top D(s)} = \mathrm{AUC}_{\alpha'^\top D(s), \beta'^\top D(s)}. \quad (6)$$

However, observe that Eq. (6) is under-specified in the sense that it includes constraints that actually give an advantage to one of the groups.

We thus introduce a general framework to formulate all *relevant* AUC-based constraints (and only those) as a linear combination of 5 elementary constraints. Given a scoring function s, let the vector $C(s) = (C_1(s), \ldots, C_5(s))^\top$ where the $C_l(s)$'s are elementary fairness measurements. Specifically, the value of $|C_1(s)|$ (resp. $|C_2(s)|$) quantifies the resemblance of the distribution of the negatives (resp. positives) between the two sensitive attributes:

$$C_1(s) = \mathrm{AUC}_{H_s^{(0)}, H_s^{(1)}} - 1/2,$$
$$C_2(s) = 1/2 - \mathrm{AUC}_{G_s^{(0)}, G_s^{(1)}},$$

while $C_3(s)$, $C_4(s)$ and $C_5(s)$ measure the difference in ability of a score to discriminate between positives and negatives for any two pairs of sensitive attributes:

$$C_3(s) = \mathrm{AUC}_{H_s^{(0)}, G_s^{(0)}} - \mathrm{AUC}_{H_s^{(0)}, G_s^{(1)}},$$
$$C_4(s) = \mathrm{AUC}_{H_s^{(0)}, G_s^{(1)}} - \mathrm{AUC}_{H_s^{(1)}, G_s^{(0)}},$$
$$C_5(s) = \mathrm{AUC}_{H_s^{(1)}, G_s^{(0)}} - \mathrm{AUC}_{H_s^{(1)}, G_s^{(1)}}.$$

The family of fairness constraints we consider is then the set of linear combinations of the $C_l(s) = 0$:

$$\mathcal{C}_\Gamma(s): \quad \Gamma^\top C(s) = \textstyle\sum_{l=1}^5 \Gamma_l C_l(s) = 0, \quad (7)$$

where $\Gamma = (\Gamma_1, \ldots, \Gamma_5)^\top \in \mathbb{R}^5$.

Theorem 1. *The following statements are equivalent:*

1. *Eq.* (6) *is satisfied for any measurable scoring function s when $H^{(0)} = H^{(1)}$, $G^{(0)} = G^{(1)}$ and $\mu(\eta(X) = p) < 1$,*
2. *Eq.* (6) *is equivalent to $\mathcal{C}_\Gamma(s)$ for some $\Gamma \in \mathbb{R}^5$,*
3. *$(e_1 + e_2)^\top[(\alpha - \alpha') - (\beta - \beta')] = 0$.*

Theorem 1 shows that our general family defined by Eq. (7) compactly captures all relevant AUC-based fairness constraints (including those proposed by Beutel et al., 2019; Borkan et al., 2019; Kallus and Zhou, 2019) while ruling out the ones that are not satisfied when $H^{(0)} = H^{(1)}$ and $G^{(0)} = G^{(1)}$ (which are in fact *unfairness* constraints). Their parameters Γ are provided in Table 1. We refer to the supplementary for the proof of this result and examples of novel fairness constraints that can be expressed with Eq. (7).

As we show in Section 4.1, our unifying framework enables the design of general formulations and statistical guarantees for learning fair scoring functions, which can then be instantiated to the specific notion of AUC-based fairness that the practitioner is interested in.

3.2 Limitations of AUC-based Constraints

To illustrate the fundamental limitations of AUC-based fairness constraints, we will rely on the credit-risk screening use case described in Example 1. Imagine that the scoring function s gives the c.d.f.'s $H_s^{(z)}$ and $G_s^{(z)}$ shown in Fig. 1 (top). Looking at $G_s^{(1)}$, we can see that creditworthy ($Y = +1$) individuals of the sensitive group $Z = 1$ do not have scores smaller than 0.5 and have an almost constant positive density of scores between 0.6 and 1. On the other hand, the scores of creditworthy individuals of group $Z = 0$ are sometimes low but are mostly concentrated around 1 (greater than 0.80), as seen from $G_s^{(0)}$. The distribution of scores for individuals who do not repay their loan ($Y = -1$) is the same across groups.

Even though the c.d.f.'s $G_s^{(0)}$ and $G_s^{(1)}$ are very different, the scoring function s satisfies the AUC constraint in Eq. (4), as can be seen from Fig. 1 (bottom left). This means that creditworthy individuals from either group have the same probability of being ranked higher than a "bad borrower". However, using high thresholds (which correspond to low probabilities of default on the granted loans) will lead to unfair decisions for one group. For instance, using $t = 0.85$ gives a FNR of 30% for group 0 and of 60% for group 1, as can be seen from Fig. 1 (top). If the proportion of creditworthy people is the same in each group ($p_0 q_0 = p_1 q_1$), we would reject twice as much creditworthy people of

Table 1: Value of Γ in our formulation of Eq. (7) for AUC-based constraints introduced in previous work.

AUC-based fairness constraint	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5
Intra-group pairwise (Beutel et al., 2019), subgroup AUC (Borkan et al., 2019)	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
BNSP AUC (Borkan et al., 2019), pairwise accuracy (Beutel et al., 2019)	0	0	$\frac{q_0(1-p_0)}{1-p}$	0	$\frac{q_1(1-p_1)}{1-p}$
BPSN AUC (Borkan et al., 2019; Beutel et al., 2019; Kallus and Zhou, 2019)	0	0	$\frac{q_0 p_0}{2p}$	$\frac{1}{2}$	$\frac{q_1 p_1}{2p}$
Zero Average Equality Gap (Borkan et al., 2019)	0	1	0	0	0
Inter-group pairwise (Beutel et al., 2019), xAUC (Kallus and Zhou, 2019)	0	0	0	1	0

group 1 than of group 0! This is blatantly unfair in the sense of parity in FNR defined in Eq. (2).

In general, fairness constraints defined by the equality between two AUC's only quantify a stochastic order between distributions, not the equality between these distributions. In fact, for continuous ROCs, the equality between their two AUC's only implies that the two ROC's intersect at some unknown point. As a consequence, AUC-based fairness can only guarantee that there exists *some* threshold $t \in \mathbb{R}$ that induces a non-trivial classifier $g_{s,t} : x \mapsto 2 \cdot \mathbb{I}\{s(x) > t\} - 1$ satisfying a notion of fairness for classification (see the supplementary for details). Unfortunately, the value of t and the corresponding FPR of the ROC curves are not known in advance and are difficult to control. For the distributions of Fig. 1, we see that the classifier $g_{s,t}$ is fair in FNR only for $t = 0.72$ (20% FNR for each group) but has a rather high FPR (i.e., probability of default) of $\sim$25%, which may be not sustainable for the bank.

3.3 Learning with Pointwise ROC-based Fairness Constraints

To impose richer and more targeted fairness conditions, we propose to use *pointwise* ROC-*based fairness constraints* as an alternative to AUC-based constraints. We start from the "ideal fairness goal" of enforcing the equality of the score distributions of the positives (resp. negatives) between the two groups, *i.e.* $G_s^{(0)} = G_s^{(1)}$ (resp. $H_s^{(0)} = H_s^{(1)}$). This strong functional criterion can be expressed in terms of ROC curves. For $\alpha \in [0, 1]$, consider the deviations between the *positive* (resp. *negative*) *inter-group ROCs* and the identity function:

$$\Delta_{G,\alpha}(s) := \mathrm{ROC}_{G_s^{(0)}, G_s^{(1)}}(\alpha) - \alpha,$$
$$\big(\text{resp. } \Delta_{H,\alpha}(s) := \mathrm{ROC}_{H_s^{(0)}, H_s^{(1)}}(\alpha) - \alpha\big).$$

The aforementioned condition of equality between the distribution of the positives (resp. negatives) of the two groups are equivalent to satisfying $\Delta_{G,\alpha}(s) = 0$ (resp. $\Delta_{H,\alpha}(s) = 0$) for any $\alpha \in [0, 1]$. When both of those conditions are satisfied, all of the AUC-based fairness constraints covered by Eq. (7) are verified, as it is easy to see that $C_l(s) = 0$ for all $l \in \{1, \ldots, 5\}$. Furthermore, guarantees on the fairness of classifiers $g_{s,t}$ induced by s hold for all possible thresholds t. While this strong property is in principle desirable, it puts overly restrictive constraints on s that will often completely jeopardize its ranking performance.

We thus propose a general approach to implement the satisfaction of a *finite* number of fairness constraints on $\Delta_{H,\alpha}(s)$ and $\Delta_{G,\alpha}(s)$ for specific values of α that are relevant to the use case at hand. Our criterion is flexible enough to address the limitations of AUC-based constraints outlined above. Specifically, a practitioner can choose points for $\Delta_{H,\alpha}$ and $\Delta_{G,\alpha}$ to guarantee the fairness of classifiers obtained by thresholding the scoring function at the desired trade-offs between, say, FPR and FNR. Furthermore, we show in Proposition 1 below (proof in supplementary) that under some regularity assumption on the ROC curve (Assumption 1), if a small number of fairness constraints m_F are satisfied at discrete points $\alpha_F^{(1)}, \ldots, \alpha_F^{(m_F)}$ of an interval for $F \in \{H, G\}$, then one obtains guarantees in sup norm on $\alpha \mapsto \Delta_{F,\alpha}$ (and therefore fair classifiers) in the entire interval $[\alpha_F^{(1)}, \alpha_F^{(m_F)}]$. This result is crucial in applications where the threshold used at deployment can vary in a whole interval, such as biometric verification (Grother and Ngan, 2019) and credit-risk screening (see Example 1).

Assumption 1. *The class $\mathcal{S}$ of scoring functions take values in $(0, T)$ for some $T > 0$, and the family of cdfs $\mathcal{K} = \{G_s^{(z)}, H_s^{(z)} : s \in \mathcal{S}, z \in \{0, 1\}\}$ satisfies: (a) any $K \in \mathcal{K}$ is continuously differentiable, and (b) there exists $b, B > 0$ s.t. $\forall (K, t) \in \mathcal{K} \times (0, T)$, $b \leqslant |K'(t)| \leqslant B$. The latter condition is satisfied when scoring functions do not have flat or steep parts, see Clémençon and Vayatis (2007) (Remark 7) for a discussion.*

Proposition 1. *Under Assumption 1, if $\exists F \in \{H, G\}$ s.t. for every $k \in \{1, \ldots, m_F\}$, $|\Delta_{F,\alpha_F^{(k)}}(s)| \leqslant \epsilon$, then:*

$$\sup_{\alpha \in [0,1]} \big|\Delta_{F,\alpha}(s)\big| \leqslant \epsilon + \frac{B + b}{2b} \max_{k \in \{0, \ldots, m\}} \big|\alpha_F^{(k+1)} - \alpha_F^{(k)}\big|,$$

with the convention $\alpha_F^{(0)} = 0$ and $\alpha_F^{(m_F+1)} = 1$.

To illustrate how ROC-based fairness constraints can be designed in a practical case, we return to our credit

lending example. In Fig. 1 (bottom right), we have $\Delta_{H,\alpha}(s) = 0$ for any $\alpha \in [0,1]$ since $H_s^{(0)} = H_s^{(1)}$. However, $\Delta_{G,\alpha}(s)$ can be large: this is the case in particular for small α's (low FPR). If the goal is to obtain fair classifiers in FNR for high thresholds (i.e., low FPR), we should seek a scoring function s with $\Delta_{G,\alpha} \simeq 0$ for any $\alpha \leqslant \alpha_{\max}$, where $\alpha_{\max}$ is the maximum TPR the bank will operate at (see Fig. 1, bottom right). The value of $\alpha_{\max}$ can be chosen based on the performance of a score learned without fairness constraint if the bank seeks to limit FPR's or maximize its potential earnings. Learning with constraints for α's in an evenly spaced grid on $[0, \alpha_{\max}]$ will ensure that the resulting s yields fair classifiers $g_{s,t}$ for high thresholds t, as confirmed experimentally in Section 5.

4 LEARNING UNDER AUC AND ROC FAIRNESS CONSTRAINTS

In this section, we first introduce empirical risk minimization problems for learning under the AUC and ROC-based constraints introduced in Section 3. Then, we prove statistical learning guarantees in the form of generalization bounds, which fill a gap in the existing literature for AUC-based constraints and provide a theoretical justification for our novel ROC-based constraints. Finally, we briefly describe how to empirically minimize such criteria with gradient-based algorithms.

4.1 Learning with AUC-based Constraints

We first formulate the problem of bipartite ranking under AUC-based fairness constraints. Introducing fairness as a hard constraint is tempting, but may be costly in terms of ranking performance. In general, there is indeed a trade-off between the ranking performance and the level of fairness. For a family of scoring functions $\mathcal{S}$ and some instantiation Γ of our general fairness definition in Eq. (7), we thus define the learning problem as follows:

$$\max_{s \in \mathcal{S}} \quad \mathrm{AUC}_{H_s, G_s} - \lambda \left| \Gamma^\top C(s) \right|, \tag{8}$$

where $\lambda \geqslant 0$ is a hyperparameter balancing ranking performance and fairness.

For the sake of simplicity and concreteness, in the rest of this section we focus on a special case of Eq. (8), namely when $C(s)$ corresponds to the fairness definition in Eq. (3). One can easily extend our analysis to any other instance of our general definition in Eq. (7). We denote by s_λ^* the scoring function that maximizes the objective $L_\lambda(s)$ of Eq. (8), where:

$$L_\lambda(s) := \mathrm{AUC}_{H_s, G_s} - \lambda \left| \mathrm{AUC}_{H_s^{(0)}, G_s^{(0)}} - \mathrm{AUC}_{H_s^{(1)}, G_s^{(1)}} \right|.$$

Given a training set $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ of n i.i.d. copies of the random triplet (X, Y, Z), we denote by $n^{(z)}$ the number of points in group $z \in \{0,1\}$, and by $n_+^{(z)}$ (resp. $n_-^{(z)}$) the number of positive (resp. negative) points in z. The empirical counterparts of $H_s^{(z)}$ and $G_s^{(z)}$ are:

$$\widehat{H}_s^{(z)}(t) = (1/n_-^{(z)}) \sum_{i=1}^n \mathbb{I}\{Z_i = z, Y_i = -1, s(X_i) \leqslant t\},$$
$$\widehat{G}_s^{(z)}(t) = (1/n_+^{(z)}) \sum_{i=1}^n \mathbb{I}\{Z_i = z, Y_i = +1, s(X_i) \leqslant t\}.$$

Recalling the notation $\widehat{\mathrm{AUC}}_{F,F'} := \mathrm{AUC}_{\widehat{F},\widehat{F}'}$ from Section 2.2, the empirical problem writes:

$$\widehat{L}_\lambda(s) := \widehat{\mathrm{AUC}}_{H_s, G_s} - \lambda \left| \widehat{\mathrm{AUC}}_{H_s^{(0)}, G_s^{(0)}} - \widehat{\mathrm{AUC}}_{H_s^{(1)}, G_s^{(1)}} \right|.$$

We denote its maximizer by $\widehat{s}_\lambda$. We can now state our statistical learning guarantees for fair ranking.

Theorem 2. *Assume the class of functions $\mathcal{S}$ is VC-major with finite VC-dimension $V < +\infty$ and that there exists $\epsilon > 0$ s.t. $\min_{z \in \{0,1\}, y \in \{-1,1\}} \mathbb{P}\{Y = y, Z = z\} \geqslant \epsilon$. Then, for any $\delta > 0$, for all $n > 1$, we have w.p. at least $1 - \delta$:*

$$\epsilon^2 \cdot [L_\lambda(s_\lambda^*) - L_\lambda(\widehat{s}_\lambda)] \leqslant C\sqrt{V/n} \cdot (4\lambda + 1/2) + \sqrt{\frac{\log(13/\delta)}{n-1}} \cdot \Big(4\lambda + (4\lambda + 2)\epsilon\Big) + O(n^{-1}).$$

Theorem 2 establishes a learning rate of $O(1/\sqrt{n})$ for our problem of ranking under AUC-based fairness constraints, which holds for any distribution of (X, Y, Z) as long as the probability of observing each combination of label and group is bounded away from zero. As the natural estimate of the AUC involves sums of dependent random variables, the proof of Theorem 2 does not follow from usual concentration inequalities on standard averages. Indeed, it requires controlling the uniform deviation of ratios of U-processes indexed by a class of functions of controlled complexity.

4.2 Learning with ROC-based Constraints

We now turn to the problem of bipartite ranking under ROC-based fairness constraints. Recall from Section 3.3 that we aim to satisfy some constraints on $\Delta_{H,\alpha}(s)$ and $\Delta_{G,\alpha}(s)$ for specific values of α. Denote by $m_H, m_G \in \mathbb{N}$ be the number of constraints for the negatives and the positives respectively, as well as $\alpha_H = [\alpha_H^{(1)}, \ldots, \alpha_H^{(m_H)}] \in [0,1]^{m_H}$ and $\alpha_G = [\alpha_G^{(1)}, \ldots, \alpha_G^{(m_G)}] \in [0,1]^{m_G}$ the points at which they apply (sorted in strictly increasing order).

With the notation $\Lambda := (\alpha, \lambda_H, \lambda_G)$, we can introduce the learning objective $L_\Lambda(s)$ defined as:

$$\mathrm{AUC}_{H_s, G_s} - \sum_{k=1}^{m_H} \lambda_H^{(k)} \left| \Delta_{H, \alpha_H^{(k)}}(s) \right| - \sum_{k=1}^{m_G} \lambda_G^{(k)} \left| \Delta_{G, \alpha_G^{(k)}}(s) \right|,$$

where $\lambda_H = [\lambda_H^{(1)}, \ldots, \lambda_H^{(m_H)}] \in \mathbb{R}_+^{m_H}$ and $\lambda_G = [\lambda_G^{(1)}, \ldots, \lambda_G^{(m_G)}] \in \mathbb{R}_+^{m_G}$ are hyperparameters.

The empirical counterpart $\widehat{L}_\Lambda(s)$ of L_Λ is defined as:

$$\widehat{\mathrm{AUC}}_{H_s,G_s} - \sum_{k=1}^{m_H} \lambda_H^{(k)} \left|\widehat{\Delta}_{H,\alpha_H^{(k)}}(s)\right| - \sum_{k=1}^{m_G} \lambda_G^{(k)} \left|\widehat{\Delta}_{G,\alpha_G^{(k)}}(s)\right|,$$

where $\widehat{\Delta}_{H,\alpha}(s) = \widehat{\mathrm{ROC}}_{G_s^{(0)},G_s^{(1)}}(\alpha) - \alpha$ and $\widehat{\Delta}_{G,\alpha}(s) = \widehat{\mathrm{ROC}}_{H_s^{(0)},H_s^{(1)}}(\alpha) - \alpha$ for any $\alpha \in [0,1]$.

We now prove statistical guarantees for the maximization of $\widehat{L}_\Lambda(s)$. We denote by s_Λ^* the maximizer of L_Λ over $\mathcal{S}$, and by $\hat{s}_\Lambda$ the maximizer of $\widehat{L}_\Lambda$ over $\mathcal{S}$. Our analysis relies on the regularity assumption on the ROC curve provided in Section 3.3 (Assumption 1).

Theorem 3. *Under Assumption 1 and those of Theorem 2, for any $\delta > 0$, $n > 1$, w.p. $\geqslant 1 - \delta$:*

$$\epsilon^2 \cdot [L_\Lambda(s_\Lambda^*) - L_\Lambda(\hat{s}_\Lambda)] \leqslant C\,(1/2 + 2\epsilon C_{\Lambda,\mathcal{K}})\sqrt{V/n} + 2\epsilon\,(1 + 3C_{\Lambda,\mathcal{K}})\sqrt{\frac{\log(19/\delta)}{n-1}} + O(n^{-1}),$$

where $C_{\Lambda,\mathcal{K}} = (1 + B/b)(\bar{\lambda}_H + \bar{\lambda}_G)$, with $\bar{\lambda}_H = \sum_{k=1}^{m_H} \lambda_H^{(k)}$ and $\bar{\lambda}_G = \sum_{k=1}^{m_G} \lambda_G^{(k)}$.

Theorem 3 generalizes the learning rate of $O(1/\sqrt{n})$ of Theorem 2 to ranking under ROC-based constraints. Its proof also relies on results for U-processes, but further requires a study of the deviations of the empirical ROC curve seen as ratios of empirical processes indexed by $\mathcal{S} \times [0,1]$. In that regard, our analysis builds upon the decomposition proposed in Hsieh and Turnbull (1996), which enables the derivation of uniform bounds over $\mathcal{S} \times [0,1]$ from results on standard empirical processes (van der Vaart and Wellner, 1996).

4.3 Algorithmic Details

In practice, maximizing $\widehat{L}_\lambda$ or $\widehat{L}_\Lambda$ directly by gradient ascent is not feasible since the criteria are not continuous. We use classic smooth surrogate relaxations of the AUCs or ROCs based on the logistic function $\sigma : x \mapsto 1/(1+e^{-x})$. We also remove the absolute values in $\widehat{L}_\lambda$ and $\widehat{L}_\Lambda$, and instead rely on parameters that are modified adaptively during the training process. We solve the problem using a stochastic gradient ascent algorithm, and modify the introduced parameters every fixed number of iterations based on fairness statistics evaluated on a small validation set. We refer to the supplementary material for more details on the algorithms we use in our experiments.

The hyperparameter λ should be tuned to achieve the desired trade-off between ranking performance and fairness. For learning under a ROC-based constraint,

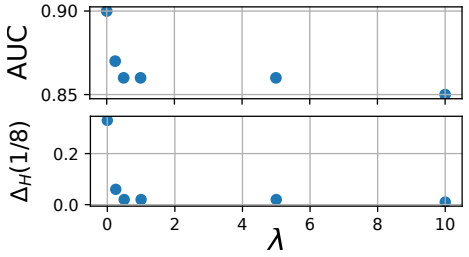

Figure 2: Ranking accuracy (AUC) and a ROC-based constraint at $\Delta_H(1/8)$ as a function of the hyperparameter λ, on the Adult dataset.

Fig. 2 provides examples of trade-offs for different λ's on the dataset *Adult* presented in Section 5.

5 EXPERIMENTS

In this section, we present a subset of our experimental results, which we think nicely illustrates the differences between AUC and ROC-based fairness. It also shows how these constraints can be used to achieve a trade-off between ranking performance and the desired notion of fairness in practical use cases. Due to space limitations, we refer to the supplementary material for the presentation of all details on the experimental setup, as well as additional results.

Results are summarized in Fig. 3, which shows ROC curves for 2-layer neural scoring functions learned with and without fairness constraints on 2 real datasets: *Compas* and *Adult* (used *e.g.* in Donini et al., 2018).

Compas is a recidivism prediction dataset. We define the sensitive variable to be $Z = 1$ if the individual is categorized as African-American and 0 otherwise. In contrast to credit-risk screening, here being labeled positive (i.e., recidivist) is a disadvantage, so we consider the *Background Positive Subgroup Negative (BPSN)* AUC *fairness constraint* defined as $\mathrm{AUC}_{H_s^{(0)},G_s} = \mathrm{AUC}_{H_s^{(1)},G_s}$, which is equivalent to Eq. (4) with positive and negative labels swapped. BPSN forces the probabilities that a negative from a given group is mistakenly ranked higher than a positive to be the same across groups. While the scoring function learned without fairness constraint systematically makes more ranking errors for non-recidivist African-Americans (Fig. 3-a), we can see that learning with the AUC-constraint achieves its goal as it makes the area under $\mathrm{ROC}_{H_s^{(1)},G_s}$ and $\mathrm{ROC}_{H_s^{(0)},G_s}$ very similar (Fig. 3-c). However, slightly more of such errors are still made in the top 25% of the scores, which is the region where the prediction threshold could be set in practice for taking decisions such as denying bail. We thus configure our ROC-based fairness constraints to align the distributions of positives and negatives

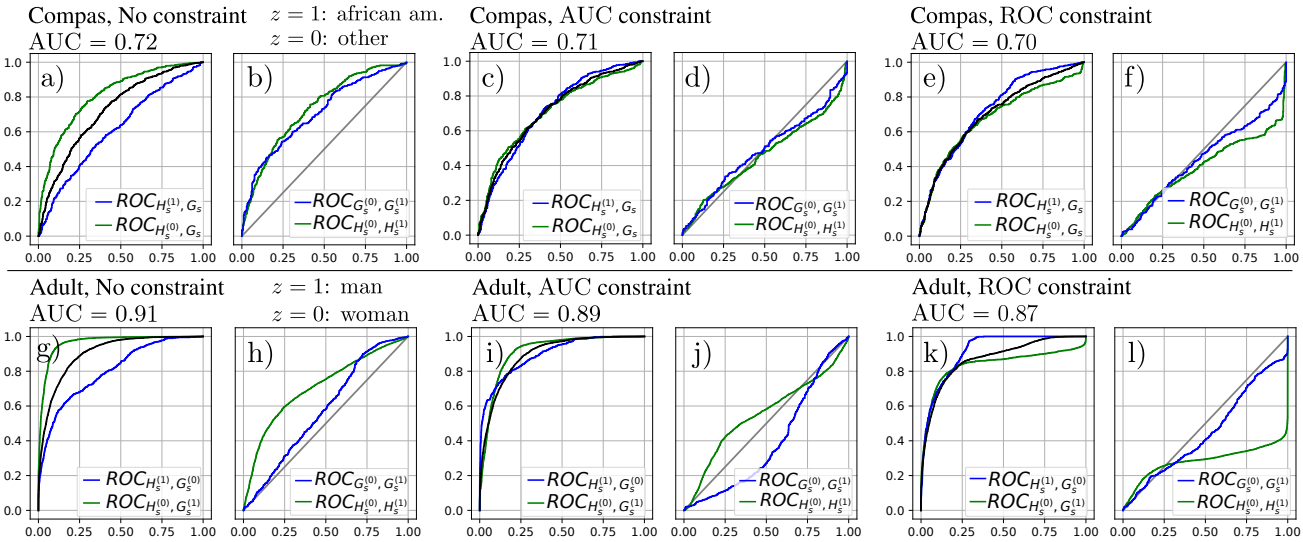

Figure 3: ROC curves on the test set of Adult and Compas for a score learned without and with fairness constraints. Black curves represent ROC_{H_s,G_s}. We also report the corresponding ranking performance AUC_{H_s,G_s}.

across both groups by penalizing solutions with high $|\Delta_{G,1/8}(s)|$, $|\Delta_{G,1/4}(s)|$, $|\Delta_{H,1/8}(s)|$ and $|\Delta_{H,1/4}(s)|$. In line with our theoretical analysis (see the discussion in Section 3.3), we can see from $\mathrm{ROC}_{G_s^{(0)},G_s^{(1)}}$ and $\mathrm{ROC}_{H_s^{(0)},H_s^{(1)}}$ that this suffices to learn a scoring function that achieves equality of the positive and negative distributions in the entire interval $[0, 1/4]$ of interest (Fig. 3-f). In turn, $\mathrm{ROC}_{H_s^{(1)},G_s}$ and $\mathrm{ROC}_{H_s^{(0)},G_s}$ become essentially equal in this region as desired (Fig. 3-e). Note that on this dataset, both the AUC and ROC constraints are achieved with minor impact on the ranking performance, as seen from the AUC scores.

We now turn to the *Adult* dataset, where we set Z to denote the gender (0 for female) and $Y = 1$ indicates that the person makes over \$50K/year. For this dataset, we plot $\mathrm{ROC}_{H_s^{(1)},G_s^{(0)}}$ and $\mathrm{ROC}_{H_s^{(0)},G_s^{(1)}}$ and observe that without fairness constraint, men who make less than \$50K are much more likely to be mistakenly ranked above a woman who actually makes more, than the other way around (Fig. 3-g). The learned score thus reproduces a common gender bias. To fix this, the appropriate notion of AUC-based fairness is Eq. (5). We see that learning under this constraint successfully equates the area under $\mathrm{ROC}_{H_s^{(1)},G_s^{(0)}}$ and $\mathrm{ROC}_{H_s^{(0)},G_s^{(1)}}$ (Fig. 3-i). However, this comes at the cost of introducing a small bias against men in the top scores. As seen from $\mathrm{ROC}_{H_s^{(0)},H_s^{(1)}}$ and $\mathrm{ROC}_{G_s^{(0)},G_s^{(1)}}$, positive women now have higher scores overall than positive men, while negative men have higher scores than negative women (Fig. 3-j). These observations illustrate the limitations of AUC fairness (see Section 3.2). To address them, we use the same ROC constraints as for *Compas* so as to align the positive and negative distributions of each group in $[0, 1/4]$. This is again achieved almost perfectly in the entire interval (Fig. 3-l). While the degradation in ranking performance is more noticeable on this dataset, a clear advantage from ROC-based fairness is that the scoring function can be thresholded to obtain fair classifiers at a wide range of thresholds.

6 DISCUSSION

In this work, we studied the problem of fairness for scoring functions learned from binary labeled data. We proposed a general framework for designing AUC-based fairness constraints, introduced novel ROC-based constraints, and derived statistical guarantees for learning scoring functions under such constraints. Although we focused on ROC curves, our framework can be adapted to *precision-recall curves* (as they are a function of the FPR and TPR (Clémençon and Vayatis, 2009)). It can also be extended to *similarity ranking*, a variant of bipartite ranking covering applications like biometric authentication (Vogel et al., 2018).

Recent work derived analytical expressions of optimal fair models for learning problems other than bipartite ranking (Menon and Williamson, 2018; Chzhen et al., 2020). A promising direction for future work is to derive a similar result for scoring functions. This would enable us to propose a compelling theoretical study of the trade-offs between performance and fairness in bipartite ranking, and lay the foundations for provably fair extensions of ROC curve optimization algorithms based on recursive partitioning (Clémençon et al., 2011; Clémençon and Vayatis, 2010).

References

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 2018.

S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

Y. Bechavod and K. Ligett. Learning fair classifiers: A regularization-inspired approach. *CoRR*, abs/1707.00044, 2017.

A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019*, pages 2212–2220. ACM, 2019.

A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, pages 405–414. ACM, 2018.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion of The 2019 World Wide Web Conference (WWW)*, 2019.

L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018*, volume 107 of *LIPIcs*, pages 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

J. Chen. Fair lending needs explainable models for responsible recommendation. *CoRR*, abs/1809.04684, 2018.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees. HAL, archives ouvertes, Mar. 2020.

S. Clémençon and N. Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8: 2671–2699, 2007.

S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

S. Clémençon and N. Vayatis. The RankOver algorithm: overlaid classification rules for optimal ranking. *Constructive Approximation*, 32:619–648, 2010.

S. Clémençon, M. Depecker, and N. Vayatis. Adaptive partitioning schemes for bipartite ranking. *Machine Learning*, 83(1):31–69, 2011.

S. Clémençon and N. Vayatis. Nonparametric estimation of the precision-recall curve. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 185–192. ACM, 2009.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and Empirical Minimization of U-Statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

S. Clémençon, M. Depecker, and N. Vayatis. AUC optimization and the two-sample problem. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009*, pages 360–368. Curran Associates, Inc., 2009.

R. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 2796–2806, 2018.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012*, pages 214–226. ACM, 2012.

P. Grother and M. Ngan. Face Recognition Vendor Test (FRVT) — Performance of Automated Gender Classification Algorithms. Technical Report NISTIR 8052, National Institute of Standards and Technology (NIST), 2019.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 3315–3323, 2016.

F. Hsieh and B. W. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24(1): 25–40, 1996.

N. Kallus and A. Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the XAUC metric. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 3433–3443. 2019.

J. M. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

A. K. Menon and R. C. Williamson. Bipartite ranking: a risk-theoretic perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016.

A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR, 2018.

G. Pleiss, M. Raghavan, F. Wu, J. M. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5680–5689, 2017.

C. Rudin. Ranking with a p-norm push. In *Learning Theory, 19th Annual Conference on Learning Theory, COLT 2006*, volume 4005 of *Lecture Notes in Computer Science*, pages 589–604. Springer, 2006.

C. Rudin, C. Wang, and B. Coker. The age of secrecy and unfairness in recidivism prediction. *CoRR*, abs/1811.00731, 2018.

A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, pages 2219–2228. ACM, 2018.

A. Singh and T. Joachims. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 5427–5437, 2019.

A. W. van der Vaart and J. a. Wellner. *Weak convergence and empirical processes*. 1996.

R. Vogel, A. Bellet, and S. Clémençon. A probabilistic theory of supervised similarity learning for pointwise ROC curve optimization. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5062–5071. PMLR, 2018.

B. E. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953. PMLR, 2017.

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*, pages 1171–1180. ACM, 2017a.

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, AISTATS 2017*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 2017b.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017*, pages 1569–1578. ACM, 2017.

P. Zhao, S. C. H. Hoi, R. Jin, and T. Yang. Online AUC maximization. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 233–240. Omnipress, 2011.