Learning Fair Scoring Functions: Fairness Definitions, Algorithms and Generalization Bounds for Bipartite Ranking

Robin Vogel [1 2] **Aurélien Bellet** [3] **Stéphan Clémençon** [2]

Abstract

Many applications of artificial intelligence, ranging from credit lending to the design of medical diagnosis support tools through recidivism prediction, involve *scoring* individuals using a learned function of their attributes. These predictive risk scores are used to rank a set of people, and/or take individual decisions about them based on whether the score exceeds a certain threshold that may depend on the context in which the decision is taken. The level of delegation granted to such systems will heavily depend on how questions of *fairness* can be answered. While this concern has received a lot of attention in the classification setup, the design of relevant fairness constraints for the problem of learning scoring functions has not been much investigated. In this paper, we propose a flexible approach to group fairness for the scoring problem with binary labeled data, a standard learning task referred to as *bipartite ranking*. We argue that the functional nature of the ROC curve, the gold standard measuring ranking performance in this context, leads to several possible ways of formulating fairness constraints. We introduce general classes of fairness conditions in bipartite ranking and establish generalization bounds for scoring rules learned under such constraints. Beyond the theoretical formulation and results, we design practical learning algorithms and illustrate our approach with numerical experiments.

1. Introduction

With the availability of data at ever finer granularity through the Internet-of-Things and the development of technological bricks to efficiently store and process this data, the infatuation with machine learning and artificial intelligence is spreading to nearly all fields (science, transportation, energy, medicine, security, banking, insurance, commerce, etc.). Expectations are high. AI is supposed to allow for the development of personalized medicine that will adapt a treatment to the patient's genetic traits. Autonomous vehicles will be safer and be in service for longer. 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, machine learning systems that make significant decisions for humans, regarding for instance credit lending in the banking sector, diagnosis in medicine or recidivism prediction in criminal justice (see Chen, 2018; Deo, 2015; Rudin et al., 2018), should guarantee that they do not penalize certain groups of individuals.

Hence, stimulated by the societal demand, notions of *fairness* in machine learning and guarantees that they can be fulfilled by decision-making 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) or (Kleinberg & Raghavan, 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 machine learning focuses on binary classification, the flagship problem in statistical learning theory. In this context, fairness constraints force the classifiers to have the same true positive rate (or false positive rate) across the sensitive groups. For instance, Hardt & Srebro (2016) and Pleiss et al. (2017) propose to modify a pre-trained classifier in order to fulfill such constraints without deteriorating classification performance. Other work incorporates fairness constraints in the learning stage (see *e.g.* Agarwal et al., 2017; Woodworth et al., 2017; Zafar et al., 2017a;b; 2019; Menon & Williamson, 2018; Bechavod & Ligett, 2018, among others). Statistical guarantees (in the form of generalization bounds) for classifiers obtained through empirical risk minimization under fairness constraints are established in (Donini et al., 2018).

The present paper is also devoted to algorithmic fairness, but for a different problem: namely, learning scoring functions from binary labeled data. This statistical learning problem, usually referred to as *bipartite ranking*, is of consider-

[1]IDEMIA, France [2]LTCI, Télécom Paris, Institut Polytechnique de Paris, France [3]INRIA, France. Correspondence to: Robin Vogel <robin.vogel@telecom-paris.fr>.

able importance in the applications. It covers in particular tasks such as credit scoring in banking, pathology scoring in medicine or recidivism scoring in criminal justice, for which fairness requirements are a major concern (Kallus & 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 *vs* the true positive rate (the higher the curve, the more accurate the ranking induced by s). Sup-norm optimization of the ROC curve has been investigated in (Clémençon & Vayatis, 2009; 2010), while most of the literature focuses on the maximization of scalar summaries of the ROC curve such as the AUC criterion (see *e.g.* Agarwal et al., 2005; Clémençon et al., 2008; Zhao et al., 2011) or alternative measures (Rudin, 2006; Clémençon & Vayatis, 2007; Menon & Williamson, 2016).

We propose a thorough study of fairness in bipartite ranking, where the goal is to guarantee that sensitive variables have little impact on the rankings induced by a scoring function. Similarly to ranking performance, there are various possible options to measure the fairness of a scoring function. As a first go, we introduce a general family of AUC-based fairness constraints which encompasses recently proposed notions (Borkan et al., 2019; Beutel et al., 2019; Kallus & Zhou, 2019) in a unified framework. However, we argue that AUC criteria may not always be appropriate insofar as two ROC curves with very different shapes may have exactly the same AUC. This motivates our design of richer definitions of fairness for scoring functions related to the ROC curves themselves. Crucially, these definitions have strong implications for fair classification: classifiers obtained by thresholding such fair scoring functions approximately satisfy definitions of classification fairness for a wide range of thresholds. We establish the first generalization bounds for scoring functions learned under AUC and ROC-based fairness constraints, following in the footsteps of Donini et al. (2018) for fair empirical risk minimizers in classification. Beyond our theoretical analysis, we propose training algorithms based on gradient descent and illustrate the practical relevance of our approach on synthetic and real datasets.

The rest of the paper is organized as follows. In Section 2, we briefly recall the key concepts of bipartite ranking and review the related work in fairness for classification and ranking. In Section 3, we introduce our general family of AUC-based fairness constraints and use it to formulate optimization problems and statistical guarantees for learning fair scoring functions. The limitations of AUC-based fairness constraints are discussed in Section 4, leading to the design of richer ROC-based (functional) fairness definitions for which we also provide generalization bounds. Finally, Section 5 presents illustrative numerical experiments on synthetic and real data, and we conclude in Section 6. Due to space limitations, technical details and additional experiments can be found in the supplementary material.

2. Background and Related Work

In this section, we introduce the main concepts involved in the subsequent analysis. We start by introducing the probabilistic framework we consider. We then recall the problem of bipartite ranking and its connections to ROC analysis, and briefly discuss the formalization of fairness in the context of classification. Finally, we review the related work in fairness for ranking, focusing on relevant AUC-based fairness constraints introduced in prior 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 feature 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 regression function $\eta(x) := \mathbb{P}\{Y = +1 \mid X = x\}$. Equipped 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. This order is defined by transporting the natural order on the real line to the feature space through a scoring rule (or scorer) $s : \mathcal{X} \to \mathbb{R}$. Given a distribution F over $\mathcal{X}$ and a scorer 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 one-dimensional distributions in a large variety of applications such as anomaly detection, medical diagnosis, information retrieval, etc.

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

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

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

The L_1 distance of $\mathrm{ROC}_{g,h}$ to the diagonal conveniently quantifies the deviation from the homogeneous case, leading to the classic area under the ROC curve (AUC) criterion:

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

where S and S' denote independent random variables, drawn from $h(dt)$ and $g(dt)$ respectively.

In the context of bipartite ranking, one is interested in the ability of the scorer s to separate positive from negative data, which is reflected by the curve ROC_{H_s,G_s}. The global summary AUC_{H_s,G_s} serves as the standard performance measure (Clémençon et al., 2008).

Empirical estimates. In practice, the scorer 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:

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

Note that the denominators n_+ and n_- are sums of i.i.d. random 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:

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

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 rules (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.* for job applications. 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 scorer $s : \mathcal{X} \to \mathbb{R}$ induces an infinite collection of binary classifiers $\{g_{s,t}(x) := \mathrm{sign}(s(x) - t)\}_{t \in \mathbb{R}}$. While one could fix a threshold $t \in \mathbb{R}$ and try to enforce fairness on $g_{s,t}$, we are interested in notions of fairness for the scorer itself, independently of a particular choice of threshold.*

2.4. Fairness in Ranking

Fairness for rankings has only recently become a research topic of interest, and most of the work originates from 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 & Joachims, 2018). Singh & 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 (which

are typically not binary). This framework is fundamentally different from the bipartite ranking setting considered here.

AUC constraints. In a setting closer to ours, Kallus & Zhou (2019) introduce fairness constraints to better quantify the fairness of a known scoring functions on binary labeled data (they do not address learning). Similar definitions of fairness are also considered by Beutel et al. (2019), and by Borkan et al. (2019) in a classification context. Below, we present these definitions in the unified form of *equalities between two* AUC*s*. In general, the AUC can be seen as a measure of homogeneity between two distributions. Its empirical counterpart (called the Mann-Whitney statistic in hypothesis testing) is often used to test equality between distributions, see (Vayatis et al., 2009) for details on this interpretation of AUCs.

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.*

$$G_s^{(z)}(t) = G^{(z)}\left(s(X) \leqslant t\right), H_s^{(z)}(t) = H^{(z)}\left(s(X) \leqslant t\right),$$

for any $t \in \mathbb{R}$. Both Beutel et al. (2019) and Borkan et al. (2019) proposed the following fairness constraints:

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

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

Eq. (3) is referred to as *intra-group pairwise* or *subgroup AUC* fairness and Eq. (4) as *pairwise accuracy* or *Background Positive Subgroup Negative (BNSP) AUC* fairness, depending on the authors. Eq. (3) requires the ranking performance (as measured by the AUC) to be equal *within* groups, which is relevant for instance in situations where groups are ranked separately (*e.g.*, candidates for two types of jobs). Eq. (4) enforces that positive instances from either group have the same probability of being ranked higher than a negative example, and can be seen as the ranking counterpart of parity in FNR for classification, see Eq. (2). (Borkan et al., 2019; Kallus & Zhou, 2019) also consider the following notions:

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

$$\mathrm{AUC}_{G_s,G_s^{(0)}} = \mathrm{AUC}_{G_s,G_s^{(1)}}. \tag{6}$$

Borkan et al. (2019) refers to Eq. (5) as *Backgroup Positive Subgroup Negative (BPSN)* AUC and can be seen as the counterpart of parity in FPR for classification, see Eq. (2). The notion of *Average Equality Gap (AEG)* introduced by Borkan et al. (2019) can be written $\mathrm{AUC}(G_s, G_s^{(z)}) - 1/2$ for $z \in \{0,1\}$. Eq. (6) thus corresponds to an AEG of zero, which means that the scores of the positives of any group are not stochastically larger than those of the other. Beutel et al. (2019) and Kallus & Zhou (2019) also define respectively the *inter-group pairwise fairness* or xAUC *disparity*:

$$\mathrm{AUC}_{H_s^{(0)},G_s^{(1)}} = \mathrm{AUC}_{H_s^{(1)},G_s^{(0)}}, \tag{7}$$

which imposes that the positives of a group can be distinguished from the negatives of the other group as effectively for both groups. Next, we generalize these AUC-based definitions and derive generalization bounds and algorithms for learning scoring functions under such fairness constraints.

3. Fair Scoring via AUC Constraints

In this section, we give a thorough treatment of the problem of statistical learning of scoring functions under AUC-based fairness constraints. First, we introduce a general family of AUC-based fairness definitions which encompasses those presented in Section 2.4 as well as many others. We then derive generalization bounds for the bipartite ranking problem under AUC-based fairness constraints. Finally, we propose a practical algorithm to learn such fair scoring functions.

3.1. A Family of AUC-based Fairness Definitions

Many sensible fairness definitions can be expressed in terms of the AUC between two distributions. We now introduce a framework to formulate AUC-based fairness constraints as a linear combination of a set of 5 elementary fairness constraints, and prove its generality. Given a scorer s, we introduce the vector $C(s) = (C_1(s), \ldots, C_5(s))^\top$, where the $C_l(s)$'s, $l \in \{1, \ldots, 5\}$, are elementary fairness measurements. More precisely, 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 positive and negative 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 elementary fairness constraints are the equations $C_l(s) = 0$ for any $l \in \{1, \ldots, 5\}$.

The family of fairness constraints we consider is the set of linear combinations of the elementary fairness constraints:

$$\mathcal{C}_\Gamma(s): \quad \Gamma^\top C(s) = \textstyle\sum_{l=1}^5 \Gamma_l C_l(s) = 0, \tag{8}$$

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

The following theorem shows that the family $(\mathcal{C}_\Gamma(s))_{\Gamma \in \mathbb{R}^5}$ covers a wide array of possible fairness constraints in the form of equalities of the AUC's between mixtures of the distributions $D(s) := (H_s^{(0)}, H_s^{(1)}, G_s^{(0)}, G_s^{(1)})^\top$. Denote

by (e_1, e_2, e_3, e_4) the canonical basis of $\mathbb{R}^4$, as well as the constant vector $\mathbf{1} = \sum_{k=1}^4 e_k$. Introducing the probability vectors $\alpha, \beta, \alpha', \beta' \in \mathcal{P}$ where $\mathcal{P} = \{v \mid v \in \mathbb{R}_+^4, \mathbf{1}^\top v = 1\}$, we define the following constraint:

$$\mathrm{AUC}_{\alpha^\top D(s), \beta^\top D(s)} = \mathrm{AUC}_{\alpha'^\top D(s), \beta'^\top D(s)}. \quad (9)$$

Theorem 1. *The following propositions are equivalent:*

1. *Eq. (9) is satisfied for any measurable scorer s when $H^{(0)} = H^{(1)}$, $G^{(0)} = G^{(1)}$ and $\mu(\eta(X) = p) < 1$,*
2. *Eq. (9) 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. (8) compactly captures all relevant AUC-based fairness constraints while ruling out the ones that are not satisfied when $H^{(0)} = H^{(1)}$ and $G^{(0)} = G^{(1)}$. Such undesirable fairness constraints are those which actually give an advantage to one of the groups, such as $\mathrm{AUC}_{G_s^{(0)} G_s^{(1)}} = 2\mathrm{AUC}_{H_s, G_s} - 1$ which is a special case of Eq. (9) that requires the scores of the positives of group 1 to be higher than those of group 0.

All AUC-based fairness constraints proposed in previous work (see Section 2.4) can be written as instances of our general definition for a specific choice of Γ, see Table 1. Note that Γ might depend on the quantities q_0, p_0, q_1, p_1. Interestingly, new fairness constraints can be expressed using our general formulation. Denoting $F_s^{(0)} = (1 - p_0)H_s^{(0)} + p_0 G_s^{(0)}$, consider for instance the following constraint:

$$\mathrm{AUC}_{F_s^{(0)}, G_s^{(0)}} = \mathrm{AUC}_{F_s^{(0)}, G_s^{(1)}}. \quad (10)$$

It equalizes the expected position of the positives of each group with respect to a *reference group* (here group 0). Another fairness constraint of interest is based on the rate of misranked pairs when one element is in a specific group:

$$E(s, z) := (1/2) \cdot \mathbb{P}\{s(X) = s(X) \mid Y \neq Y', Z = z\} + \mathbb{P}\{(s(X) - s(X))(Y - Y') > 0 \mid Y \neq Y', Z = z\}.$$

The equality $E(s, 0) = E(s, 1)$ can be seen as the analogue of *parity in mistreatment* for the task of ordering pairs, see Eq. (1). It is easy to see that this constraint can be written in the form of Eq. (9) and that point 1 of Theorem 1 holds, hence it is equivalent to $\mathcal{C}_\Gamma(s)$ for some $\Gamma \in \mathbb{R}^5$.

3.2. Statistical Learning Guarantees

We now formulate the problem of fair ranking based on the fairness definitions introduced above. While it is tempting to introduce fairness as a hard constraint, this may come at a large cost in terms of the ability of such scorers to separate positive from negative data points. In general, there is a trade-off between the ranking performance and the level of fairness, as illustrated by the following example.

Table 1. Value of $\Gamma = (\Gamma_l)_{l=1}^5$ for all of the AUC-based fairness constraints in the paper for the general formulation of Eq. (8).

Eq.	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5
(3)	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
(4)	0	0	$\frac{q_0(1-p_0)}{1-p}$	0	$\frac{q_1(1-p_1)}{1-p}$
(5)	0	0	$\frac{q_0 p_0}{2p}$	$\frac{1}{2}$	$\frac{q_1 p_1}{2p}$
(6)	0	1	0	0	0
(7)	0	0	0	1	0
(10)	0	p_0	$1 - p_0$	0	0

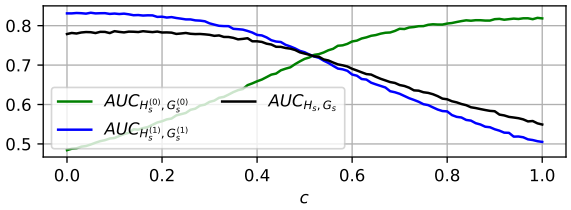


Figure 1. Plotting Example 1 for $q_1 = 17/20$. Under the fairness definition Eq. (3), a fair solution exists for $c = 1/2$, but the ranking performance for $c < 1/2$ is significantly higher.

Example 1. *Let $\mathcal{X} = [0, 1]^2$. For any $x = (x_1\ x_2)^\top \in \mathcal{X}$, let $\mu^{(0)}(x) = \mu^{(1)}(x) = 1$, as well as $\eta^{(0)}(x) = x_1$ and $\eta^{(1)}(x) = x_2$. We have $\mu(x) = 1$ and $\eta(x) = q_0 x_1 + q_1 x_2$. Consider linear scorers of the form $s_c(x) = cx_1 + (1-c)x_2$ parameterized by $c \in [0, 1]$. Fig. 1 plots AUC_{H_s, G_s} and $AUC_{H_s^{(z)}, G_s^{(z)}}$ for $z \in \{0, 1\}$ as a function of c, illustrating the trade-off between fairness and ranking performance.*

For a family of scoring functions $\mathcal{S}$ and some instantiation Γ of our general fairness definition in Eq. (8), we thus define the learning problem as follows:

$$\max_{s \in \mathcal{S}} \quad \mathrm{AUC}_{H_s, G_s} - \lambda \left| \Gamma^\top C(s) \right|, \quad (11)$$

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 the special case of the fairness definition in Eq. (3) — one can easily extend our analysis to any other instance of our general definition in Eq. (8). The objective in Eq. (11) then writes:

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

and we denote its maximizer by s_λ^*.

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 this group. The

empirical counterparts of $H_s^{(z)}$ and $G_s^{(z)}$ are then given by:

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

Recalling the notation $\widehat{\mathrm{AUC}}_{F,F'} := \mathrm{AUC}_{\widehat{F},\widehat{F}'}$ introduced in Section 2.2, the empirical problem can thus be written:

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

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 $1 - \delta$:*

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

Theorem 2 establishes a learning rate of $O(1/\sqrt{n})$ for our problem of ranking under AUC-based fairness constraint, 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.

3.3. Training Algorithm

In practice, maximizing $\widehat{L}_\lambda$ directly by gradient ascent is not feasible since the criterion is not continuous. As standard in the literature, we can use smooth surrogate relaxations of the AUCs based on the logit function $\sigma : x \mapsto 1/(1 + e^{-x})$. Again, we illustrate our approach for the fairness definition in Eq. (8). The surrogate relaxation of $\widehat{\mathrm{AUC}}_{H_s,G_s}$ writes:

$$\widetilde{\mathrm{AUC}}_{H_s,G_s} = \frac{1}{n_+ n_-} \sum_{i<j} \sigma\left[(s(x_i) - s(x_j))(y_i - y_j)\right].$$

Similarly, for $z \in \{0, 1\}$, we obtain $\widetilde{\mathrm{AUC}}_{H_s^{(z)},G_s^{(z)}}$ by averaging over pairs of positive/negative points in group z. The overall relaxed objective we aim to maximize is then:

$$\widetilde{\mathrm{AUC}}_{H_s,G_s} - \lambda \cdot c\big(\widetilde{\mathrm{AUC}}_{H_s^{(1)},G_s^{(1)}} - \widetilde{\mathrm{AUC}}_{H_s^{(0)},G_s^{(0)}}\big).$$

We solve the problem using a stochastic gradient descent algorithm in which the constant $c \in [-1, 1]$ is set adaptively during the training process based on a small validation set. Specifically, if more errors are done on group 0 than group 1, *i.e.* $\widehat{\mathrm{AUC}}_{H_s^{(1)},G_s^{(1)}} > \widehat{\mathrm{AUC}}_{H_s^{(0)},G_s^{(0)}}$ on the validation set, we set c to $\min(c + \Delta c, 1)$ (where Δc is a small positive constant) so as to increase the weight of those errors. In the other case, we set c to $\max(c - \Delta c, -1)$. We update the value of c every fixed number n_{adapt} of iterations. Details on the implementation are given in the supplementary material.

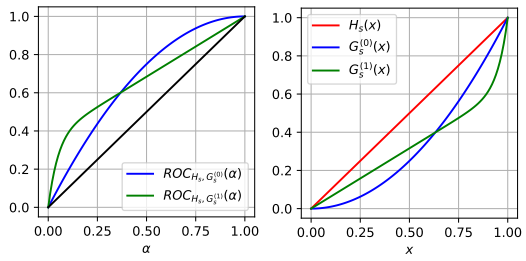


Figure 2. In this example, Eq. (4) is verified, but $G_s^{(0)}$ and $G_s^{(1)}$ are very different. Indeed, $\sup_{t\in[0,1]} |G_s^{(0)}(t) - G_s^{(1)}(t)| \approx 0.10$.

4. Beyond AUC-based Fairness Constraints

In this section, we highlight some limitations of the AUC-based fairness constraints studied in Section 3. These serve as a motivation to introduce new pointwise ROC-based fairness constraints. We then present generalization bounds and gradient descent algorithms for learning scoring functions under such constraints.

4.1. Limitations of AUC-based Constraints

As mentioned in Section 2.4, the equality of two AUCs can be used to measure homogeneity between two distributions. However it only quantifies a stochastic order between the two distributions, and not the equality: see Fig. 2 for an example where two very different distributions are indistinguishable in terms of AUC. For continuous ROCs, the equality between their two AUC's only implies that the two ROCs intersect at some unknown point, as shown by Proposition 1 (a simple consequence of the mean value theorem). Borkan et al. (2019, Theorem 3.3 therein) corresponds to the special case of Proposition 1 when $h = g, h' \neq g'$.

Proposition 1. *Let h, g, h', g' be cdfs on $\mathbb{R}$ s.t. $\mathrm{ROC}_{h,g}$ and $\mathrm{ROC}_{h',g'}$ are continuous. If $\mathrm{AUC}_{h,g} = \mathrm{AUC}_{h',g'}$, then there exists $\alpha \in (0, 1)$, s.t. $g \circ h^{-1}(\alpha) = g' \circ h'^{-1}(\alpha)$.*

Proposition 1 implies that when a scorer s satisfies some AUC-based fairness constraint, there exists a threshold $t \in \mathbb{R}$ inducing a non-trivial classifier $g_{s,t} := \mathrm{sign}(s(x) - t)$ that satisfies a notion of fairness for classification.

Corollary 1. *Under appropriate conditions on the scorer s (i.e, $s \in \mathcal{S}$ where $\mathcal{S}$ satisfies Assumption 1), we have that:*

- *If $p_0 = p_1$ and s satisfies Eq. (3), then there exists $(t_0, t_1) \in (0, T)^2$, s.t. $M^{(0)}(g_{s,t_0}) = M^{(1)}(g_{s,t_1})$, which resembles parity in mistreatment (Eq. (1)).*
- *If s satisfies Eq. (4) or (6) or (10), then $g_{s,t}$ satisfies fairness in FNR (Eq. (2)) for some threshold $t \in (0, T)$.*
- *If s satisfies Eq. (5), then $g_{s,t}$ satisfies parity in FPR (Eq. (2)) for some threshold $t \in (0, T)$.*

Unfortunately, AUC-based fairness constraints guarantee classification fairness for a *single* threshold t, corresponding to a specific point α of the ROC curve (over which one has no control). In contrast, in many applications, one is interested in learning a scoring function s which induces classifiers $g_{s,t}$ that are fair in *particular regions* of the ROC curve. One striking example is in biometric verification, where one is interested in low false positive rates (*i.e.*, large thresholds t), see Grother & Ngan (2019). In many practical scenarios, learning with AUC-based fairness constraints thus leads to inadequate scorers.

4.2. Pointwise ROC-based Fairness Constraints

To impose more restrictive fairness conditions, the ideal goal is to enforce 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 stronger 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) between the two groups are equivalent to satisfying, for any $\alpha \in [0,1]$:

$$\Delta_{G,\alpha}(s) = 0 \quad \big(\text{resp. } \Delta_{H,\alpha}(s) = 0\big). \tag{12}$$

When both conditions in Eq. (12) are satisfied, all of the fairness constraints covered by Theorem 1 are verified, since $C_l(s) = 0$ for any $l \in \{1,\dots,5\}$. Furthermore, guarantees on the fairness of classifiers induced by s (see Corollary 1) hold for all thresholds. While this strong property is desirable, it is challenging to enforce in practice due to its functional nature, and in many cases it may only be achievable by completely jeopardizing the ranking performance.

We thus propose to implement a trade-off between the ranking performance and the satisfaction of a finite number of fairness constraints on the value of $\Delta_{H,\alpha}(s)$ and $\Delta_{G,\alpha}(s)$ for specific values of α. Let $m_H, m_G \in \mathbb{N}$ be the number of constraints for the negatives and the positives respectively, as well as $\alpha_H = [\alpha_H^{(1)},\dots,\alpha_H^{(m_H)}] \in [0,1]^{m_H}$ and $\alpha_G = [\alpha_G^{(1)},\dots,\alpha_G^{(m_G)}] \in [0,1]^{m_G}$ the points at which they apply (sorted in strictly increasing order). With the notation $\Lambda := (\alpha, \lambda_H, \lambda_G)$, we can introduce the criterion $L_\Lambda(s)$, defined as:

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

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

This criterion is flexible enough to address the scenarios outlined in Section 4.1. In particular, under some regularity assumption on the ROC curve, Proposition 2 shows, that if a small number of constraints m_F are satisfied for $F \in \{H,G\}$, one obtains guarantees in sup norm on $\alpha \mapsto \Delta_{F,\alpha}$.

Assumption 1. *The candidate scoring functions $\mathcal{S}$ take their 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 as soon as the score functions do not present neither flat nor steep parts, see Clémençon & Vayatis (2007, Remark 7 therein) for a discussion.*

Proposition 2. *Under Assumption 1, if $\exists F \in \{H,G\}$ s.t. for every $k \in \{1,\dots,m_F\}$, $|\Delta_{F,\alpha_F^{(k)}}(s)| \leqslant \epsilon$, then:*

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

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

4.3. Statistical Guarantees

We now proceed to prove statistical guarantees for the maximization of L_Λ. Its empirical counterpart $\widehat{L}_\Lambda(s)$ writes:

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

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 study the generalization properties of the scorers that maximize $\hat{L}_\Lambda$. We denote by s_Λ^* the maximizer of L_Λ over $\mathcal{S}$, and $\hat{s}_\Lambda$ the maximizer of $\widehat{L}_\Lambda$ over $\mathcal{S}$.

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

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

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 $\bar{\lambda}_G = \sum_{k=1}^{m_G} \lambda_G^{(k)}$.

4.4. Training Algorithm

To maximize $\hat{L}_\Lambda$, we can use a similar stochastic gradient descent procedure as the one introduced in Section 3.3. We refer to the supplementary material for more details.

5. Experiments

In this section, we illustrate our approaches on synthetic and real data. We learn linear scorers for synthetic data and neu-

Table 2. Results on test set. For synthetic data, results are averaged over 100 runs (std. dev. are all smaller than 0.02). The strength of fairness constraints and regularization is chosen by cross-validation to obtain interesting trade-offs, see supplementary for more detail.

Method	AUC-based fairness						ROC-based fairness			
Value of λ	$\lambda = 0$		$\lambda > 0$				$\lambda_H^{(k)} = \lambda_H > 0$			
Toy data	AUC	ΔAUC	AUC	ΔAUC	$\lvert\Delta_H(\frac{3}{4})\rvert$	–	AUC	ΔAUC	$\lvert\Delta_H(\frac{3}{4})\rvert$	–
Example 1	**0.79**	0.28	0.73	**0.00**	–	–	–	–	–	–
Example 2	**0.80**	**0.02**	**0.80**	**0.02**	0.38	–	0.75	0.06	**0.00**	–
Real data	AUC	ΔAUC	AUC	ΔAUC	$\lvert\Delta_H(\frac{1}{8})\rvert$	$\lvert\Delta_H(\frac{1}{4})\rvert$	AUC	ΔAUC	$\lvert\Delta_H(\frac{1}{8})\rvert$	$\lvert\Delta_H(\frac{1}{4})\rvert$
German	**0.77**	0.05	0.76	0.04	**0.03**	0.01	0.73	**0.02**	**0.03**	**0.00**
Adult	**0.90**	0.04	0.89	**0.03**	0.27	0.28	0.85	0.05	**0.07**	**0.11**
Compas	**0.71**	0.03	0.70	**0.01**	0.12	0.10	0.66	0.02	**0.07**	**0.03**
Bank	**0.93**	0.15	0.77	**0.03**	0.10	0.18	0.81	0.29	**0.03**	**0.02**

ral network-based scorers for real data, using L_2 regularization on the model parameters. Results are summarized in Table 2, where AUC denotes the ranking accuracy AUC_{H_s,G_s}, and ΔAUC denotes the absolute difference of the terms in the AUC-based fairness constraint. We highlight in bold the best ranking accuracy, and the fairest algorithm for the relevant constraint. We refer to the supplementary for more details on the setup and further illustrations of the results.

Synthetic data. First, we illustrate learning with the AUC constraint in Eq. (3) on the simple problem in Example 1. Precisely, learning scorers with different λ's allow different trade-offs between ranking performance and fairness (larger λ leads to more fairness). Example 2 allows to compare AUC-based and ROC-based approaches. The former uses Eq. (3) as constraint and the latter penalizes $\Delta_{H,3/4}(s) \neq 0$. The different constraints lead to scorers with specific trade-offs between fairness and performance. Results with AUC-based fairness are the same for $\lambda = 0$ and $\lambda = 1$ because the optimal scorer for ranking satisfies Eq. (3). In the supplementary, we show that our algorithm recovers optimal scorers (as measured by the loss) for both examples.

Example 2. *Set* $\mathcal{X} = [0,1]^2$. *For any* $x \in \mathcal{X}$ *with* $x = (x_1\ x_2)^\top$, *set* $\mu^{(0)}(x) = (16/\pi) \cdot \mathbb{I}\{x^2 + y^2 \leqslant 1/2\}$, $\mu^{(1)}(x) = (16/3\pi) \cdot \mathbb{I}\{1/2 \leqslant x^2 + y^2 \leqslant 1\}$, *and* $\eta^{(0)}(x) = \eta^{(1)}(x) = (2/\pi) \cdot \arctan(x_2/x_1)$.

Real data. We now experiment with four datasets that are popular in the fairness literature: *German Credit*, *Adult Income Dataset* and *Compas Dataset* are used in (Donini et al., 2018) among others, and *Bank Marketing* is featured in (Zafar et al., 2019). In the case of *Compas* (recidivism prediction), being labeled positive is a disadvantage so the approach with AUC-based fairness uses the constraint in Eq. (5) to balance FPRs. Conversely, for *German* (credit scoring) a positive label is an advantage, so we choose Eq. (4) to balance FNRs. For the other datasets, the problem has no clear connotation so we select Eq. (3) to force the same ranking accuracy within each group. Inspired by the operational considerations presented in Section 4.1, the ROC-based approach is configured to align the distribution of FPR for low FPRs between both groups by penalizing solutions with high $|\Delta_{H,1/8}(s)|$ and $|\Delta_{H,1/4}(s)|$.

Results demonstrate that our approaches consistently balance ranking performance and the chosen notion of fairness. Naturally, the ability to reach strong fairness at a reasonable ranking performance cost strongly depends on the problem. In the supplementary, we give the full ROC curves for the learned scorers. In line with Proposition 2, scorers learned with ROC-based constraints achieve fairness not only for the two discrete points considered at training but for a whole interval, implying fairness for the classifiers induced by these scorers at a wide range of thresholds.

6. Conclusion

In this paper, we considered the issue of fairness for scoring functions learned from binary classification data. We proposed a general notion of fairness based on the AUC criterion, encompassing those previously introduced, and provided statistical guarantees for scorers learned via empirical AUC maximization under such fairness constraints. The analysis was also extended to a richer fairness notion based on ROC curves rather than their scalar summaries. From a practical perspective, we showed how to implement gradient descent algorithms to solve these learning problems and illustrated our concepts and methods via numerical experiments. We point out that our framework can be extended to *similarity ranking*, a variant of bipartite ranking covering key applications such as scoring for biometric identification, see *e.g.* Vogel et al. (2018). In future work, we plan to investigate how the unrelaxed versions of our fairness constraints can be incorporated to ROC curve optimization algorithms based on recursive partitioning, such as those developed in Clémençon et al. (2010) or Clémençon & Vayatis (2010).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

Agarwal, A., Beygelzimer, A., Dudik, M., and Langford, J. A reductions approach to fair classification. In *FAT*, 2017.

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

Bechavod, Y. and Ligett, K. Penalizing unfairness in binary classification. *arXiv:1707.00044v3*, 2018.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., and Goodrow, C. Fairness in recommendation ranking through pairwise comparisons. In *KDD*, 2019.

Biega, A. J., Gummadi, K. P., and Weikum, G. Equity of attention: Amortizing individual fairness in rankings. In *SIGIR*, 2018.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. *arXiv:1903.04561*, 2019.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification : a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pp. 169–207. 2004.

Celis, L. E., Straszak, D., and Vishnoi, N. K. Ranking with fairness constraints. In *ICALP*, 2018.

Chen, J. Fair lending needs explainable models for responsible recommendation. *arXiv:1809.04684*, 2018.

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

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

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

Clémençon, S., Depecker, M., and Vayatis, N. Adaptive partitioning schemes for bipartite ranking. *Machine Learning*, 2010.

Clémençon, S., Colin, I., and Bellet, A. Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics. *Journal of Machine Learning Research*, 17 (76):1–36, 2016.

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

Deo, R. Machine learning in medicine. *Circulation*, 132 (20):19201930, 2015.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical risk minimization under fairness constraints. In *NeurIPS*, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *ITCS*, 2012.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

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

Gyrfi, L. *Principles of Nonparametric Learning*. Springer, 2002.

Hardt, M., P. E. and Srebro, N. Equality of opportunity in supervised learning. In *NIPS*, 2016.

Kallus, N. and Zhou, A. The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the xAUC Metric. In *NeurIPS*. 2019.

Kleinberg, J., M. S. and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017.

Lee, A. J. *U-statistics: Theory and practice*. Marcel Dekker, Inc., New York, 1990.

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

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *FAT*, 2018.

Papa, G., Clémençon, S., and Bellet, A. SGD Algorithms based on Incomplete U-statistics: Large-Scale Minimization of Empirical Risk. In *NIPS*, 2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. In *NIPS*, 2017.

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv:1609.04747*, 2016.

Rudin, C. Ranking with a P-Norm Push. In *COLT*, 2006.

Rudin, C., Wang, C., and Coker, B. The age of secrecy and unfairness in recidivism prediction. *arXiv:1811.00731*, 2018.

Shorack, G. and Wellner, J. A. *Empirical Processes with applications to Statistics*. Classics in Applied Mathematics. SIAM, 1989.

Singh, A. and Joachims, T. Fairness of exposure in rankings. In *KDD*, 2018.

Singh, A. and Joachims, T. Policy learning for fairness in ranking. In *NeurIPS*, 2019.

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

van der Vaart, A. W. Asymptotic Statistics. *Cambridge Series in Statistical and Probabilistic Mathematics*, 2000.

Vayatis, N., Depecker, M., and Clémençcon, S. J. Auc optimization and the two-sample problem. In *NIPS*. 2009.

Vogel, R., Bellet, A., and Clémençon, S. A probabilistic theory of supervised similarity learning for pointwise ROC curve optimization. In *ICML*, 2018.

Woodworth, B., Gunasekar, S., Ohannessian, M., and Srebro, N. Learning non-discriminatory predictors. In *COLT*, 2017.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 2017a.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017b.

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

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. FA*IR: A Fair Top-k Ranking Algorithm. In *CIKM*, 2017.

Zhao, P., Hoi, S., Jin, R., and Yang, T. AUC Maximization. In *ICML*, 2011.

SUPPLEMENTARY MATERIAL

A. Technical Proofs

A.1. Definitions

We recall a few useful definitions.

Definition 2 (VC-major class of functions – van der Vaart & Wellner, 1996)**.** *A class of functions $\mathcal{F}$ such that $\forall f \in \mathcal{F}$, $f : \mathcal{X} \to \mathbb{R}$ is called VC-major if the major sets of the elements in $\mathcal{F}$ form a VC-class of sets in $\mathcal{X}$. Formally, $\mathcal{F}$ is a VC-major class if and only if:*

$$\{\{x \in \mathcal{X} \mid f(x) > t\} \mid f \in \mathcal{F}, t \in \mathbb{R}\} \text{ is a VC-class of sets.}$$

Definition 3 (U-statistic of degree 2 – Lee, 1990)**.** *Let $\mathcal{X}$ be some measurable space and $V_1, \ldots, V_n$ i.i.d. random variables valued in $\mathcal{X}$ and $K : \mathcal{X}^2 \to \mathbb{R}$ a measurable symmetric mapping s.t. $h(V_1, V_2)$ is square integrable. The functional $U_n(h) = (1/n(n-1)) \sum_{i \neq j} h(V_i, V_j)$ is referred to as a symmetric U-statistic of degree two with kernel h. It classically follows from Lehmann-Scheffé's lemma that it is the unbiased estimator of the parameter $\mathbb{E}[h(V_1, V_2)]$ with minimum variance.*

A.2. Proof of Theorem 1

Denote $D(s) = (D_1(s), D_2(s), D_3(s), D_4(s))^\top := (H_s^{(0)}, H_s^{(1)}, G_s^{(0)}, G_s^{(1)})^\top$. For any $(i, j) \in \{1, \ldots, 4\}^2$, we introduce the notation:

$$\mathrm{AUC}_{D_i, D_j} : s \mapsto \mathrm{AUC}_{D_i(s), D_j(s)}.$$

Introduce a function M such that $M(s) \in \mathbb{R}^{4 \times 4}$ for any $s : X \to \mathbb{R}$, and for any $(i, j) \in \{1, \ldots, 4\}$, the (i, j) coordinate of M writes:

$$M_{i,j} = \mathrm{AUC}_{D_i, D_j} - \frac{1}{2}.$$

First note that, for any s, $M(s)$ is antisymmetric *i.e.* $M_{j,i}(s) = -M_{i,j}(s)$ for any $(i, j) \in \{1, \ldots, 4\}^2$. Then, with $(\alpha, \beta) \in \mathcal{P}^2$, we have that:

$$\mathrm{AUC}_{\alpha^\top D, \beta^\top D} = \alpha^\top M \beta - \frac{1}{2} = \langle M, \alpha\beta^\top \rangle - \frac{1}{2},$$

where $\langle M, M' \rangle = \mathrm{tr}(M^\top M')$ is the standard dot product between matrices. Eq. (9) can be written as:

$$\langle M, \alpha\beta^\top - \alpha'\beta'^\top \rangle = 0. \tag{13}$$

Case of $\alpha = \alpha'$ and $\beta - \beta' = \delta(e_i - e_j)$.

Consider the specific case where $\alpha = \alpha'$ and $\beta - \beta' = \delta(e_i - e_j)$ with $i \neq j$ and $\delta \neq 0$, then

$$\langle M, \alpha(\beta - \beta')^\top \rangle = \delta K_{i,j}^{(\alpha)},$$

where:

$$\begin{aligned} K_{i,j}^{(\alpha)} &= \langle M, \alpha(e_i - e_j)^\top \rangle = \sum_{k=1}^{4} \alpha_k \left[\mathrm{AUC}_{D_k, D_i} - \mathrm{AUC}_{D_k, D_j} \right], \\ &= (\alpha_i + \alpha_j) \left[\frac{1}{2} - \mathrm{AUC}_{D_i, D_j} \right] + \sum_{k \notin \{i,j\}} \alpha_k \left[\mathrm{AUC}_{D_k, D_i} - \mathrm{AUC}_{D_k, D_j} \right], \end{aligned}$$

The preceding definition implies that $K_{i,j}^{(\alpha)} = -K_{j,i}^{(\alpha)}$. Using $\sum_{k=1}^{K} \alpha_k = 0$, we can express every $K_{i,j}^{(\alpha)}$ as a linear combinations of the C_l's plus a remainder, precisely:

$$\begin{aligned}
K_{1,2}^{(\alpha)} &= -\left(\alpha_1 + \alpha_2\right) C_1 - \alpha_3(C_3 + C_4) - \alpha_4(C_4 + C_5),\\
K_{1,3}^{(\alpha)} &= \left(\frac{1}{2} - \mathrm{AUC}_{D_1,D_3}\right) + \alpha_2(-C_1 + C_3 + C_4) + \alpha_4(-C_2 + C_3),\\
K_{1,4}^{(\alpha)} &= \left(\frac{1}{2} - \mathrm{AUC}_{D_1,D_4}\right) + \alpha_2(-C_1 + C_4 + C_5) + \alpha_3(C_2 - C_3 - C_4),\\
K_{2,3}^{(\alpha)} &= \left(\frac{1}{2} - \mathrm{AUC}_{D_2,D_3}\right) + \alpha_1(C_1 - C_3 - C_4) + \alpha_4(-C_2 + C_5),\\
K_{2,4}^{(\alpha)} &= \left(\frac{1}{2} - \mathrm{AUC}_{D_2,D_4}\right) + \alpha_1(C_1 - C_4 - C_5) + \alpha_3(C_2 - C_5),\\
K_{3,4}^{(\alpha)} &= \left(\alpha_3 + \alpha_4\right) C_2 + \alpha_1 C_3 + \alpha_2 C_5.
\end{aligned}$$

Hence, it suffices that $\{i,j\} = \{1,2\}$ or $\{i,j\} = \{3,4\}$ for Eq. (13) to be equivalent to $\mathcal{C}_\Gamma$ for some $\Gamma \in \mathbb{R}^5$.

Case of $\alpha = \alpha'$.

Any of the $\beta - \beta'$ can be written as a positive linear combination of $e_i - e_j$ with $i \neq j$, since:

$$\beta - \beta' = \frac{1}{4} \sum_{i \neq j} \left(\beta_i + \beta'_j\right) \left(e_i - e_j\right),$$

which means that, since $K_{i,j}^{(\alpha)} = -K_{j,i}^{(\alpha)}$:

$$\langle M, \alpha(\beta - \beta')^\top \rangle = \frac{1}{4} \sum_{i \neq j} \left(\beta_i + \beta'_j\right) K_{i,j}^{(\alpha)} = \frac{1}{4} \sum_{i<j} \left(\left[\beta_i - \beta_j\right] - \left[\beta'_i - \beta'_j\right]\right) K_{i,j}^{(\alpha)}. \tag{14}$$

Note that any linear combination of the $K_{1,3}^{(\alpha)}$, $K_{1,4}^{(\alpha)}$, $K_{2,3}^{(\alpha)}$ and $K_{2,4}^{(\alpha)}$:

$$\gamma_1 \cdot K_{1,3}^{(\alpha)} + \gamma_2 \cdot K_{1,4}^{(\alpha)} + \gamma_3 \cdot K_{2,3}^{(\alpha)} + \gamma_4 \cdot K_{2,4}^{(\alpha)},$$

where $\gamma \in R^4$ with $\mathbf{1}^\top \gamma = 0$ can be written as a weighted sum of the C_l for $l \in \{1, \ldots, 5\}$.

Hence, it suffices that $\beta_1 + \beta_2 = \beta'_1 + \beta'_2$ for Eq. (14) to be equivalent to some $\mathcal{C}_\Gamma$ for some $\Gamma \in \mathbb{R}^5$.

General case.

Note that, using the antisymmetry of M and Eq. (14):

$$\begin{aligned}
\langle M, \alpha\beta^\top - \alpha'\beta'^\top \rangle &= \langle M, \alpha(\beta - \beta')^\top \rangle + \langle M, (\alpha - \alpha')\beta'^\top \rangle,\\
&= \langle M, \alpha(\beta - \beta')^\top \rangle - \langle M, \beta'(\alpha - \alpha')^\top \rangle,\\
&= \frac{1}{4} \sum_{i<j} \left[\left(\left[\beta_i - \beta_j\right] - \left[\beta'_i - \beta'_j\right]\right) K_{i,j}^{(\alpha)} - \left(\left[\alpha_i - \alpha_j\right] - \left[\alpha'_i - \alpha'_j\right]\right) K_{i,j}^{(\beta')} \right],
\end{aligned}$$

Hence, it suffices that $(e_1 + e_2)^\top [(\alpha - \alpha') - (\beta - \beta')] = 0$ for Eq. (13) to be equivalent to some $\mathcal{C}_\Gamma$ for some $\Gamma \in \mathbb{R}^5$.

Conclusion.

We denote the three propositions of Theorem 1 as P_1, P_2 and P_3.

Assume that $H^{(0)} = H^{(1)}$, $G^{(0)} = G^{(1)}$ and $\mu(\eta(X) = 1/2) < 1$, then $C_l = 0$ for any $l \in \{1, \ldots, 5\}$, which gives:

$$\begin{aligned}
\langle M(s), \alpha\beta^\top - \alpha'\beta'^\top \rangle &= \frac{1}{4} \left(\frac{1}{2} - \mathrm{AUC}_{H_s,G_s}\right) \left(\sum_{i \in \{1,2\}} \sum_{j \in \{3,4\}} \left[\left(\left[\beta_i - \beta_j\right] - \left[\beta'_i - \beta'_j\right]\right) - \left(\left[\alpha_i - \alpha_j\right] - \left[\alpha'_i - \alpha'_j\right]\right) \right] \right),\\
&= \left(\frac{1}{2} - \mathrm{AUC}_{H_s,G_s}\right) (e_1 + e_2)^\top [(\alpha - \alpha') - (\beta - \beta')],
\end{aligned}$$

It is known that:

$$\mathrm{AUC}_{H_\eta,G_\eta} = \frac{1}{2} + \frac{1}{4p(1-p)} \iint \big|\eta(x) - \eta(x')\big| d\mu(x) d\mu(x'),$$

which means that $\mathrm{AUC}_{H_\eta,G_\eta} = 1/2$ implies that $\eta(X) = p$ almost surely (a.s.), and the converse is true.

Assume P_1 is true, then $\mathrm{AUC}_{H_\eta,G_\eta} > 1/2$, hence $(e_1+e_2)^\top[(\alpha-\alpha') - (\beta-\beta')] = 0$ because Eq. (13) is verified for η, and we have shown $P_1 \implies P_3$.

Assume P_3 is true, then $\langle M, \alpha\beta^\top - \alpha'\beta'^\top \rangle$ writes as a linear combination of the C_l's, $l \in \{1,\ldots,5\}$, and we have shown that $P_3 \implies P_2$.

Assume P_2 is true, then observe that if $H^{(0)} = H^{(1)}$ and $G^{(0)} = G^{(1)}$, then any $\mathcal{C}_\Gamma$ is satisfied for any $\Gamma \in \mathbb{R}^5$, and we have shown that $P_2 \implies P_1$, which concludes the proof.

A.3. Proof of Theorem 2

Usual arguments imply that: $L_\lambda(\widehat{s}_\lambda) - L_\lambda(s^*_\lambda) \leqslant 2 \cdot \sup_{s\in\mathcal{S}} \left|\widehat{L}_\lambda(s) - L_\lambda(s)\right|$. Introduce the quantities:

$$\widehat{\Delta} = \sup_{s\in\mathcal{S}} \left|\widehat{\mathrm{AUC}}_{H_s,G_s} - \mathrm{AUC}_{H_s,G_s}\right|, \qquad \widehat{\Delta}_0 = \sup_{s\in\mathcal{S}} \left|\widehat{\mathrm{AUC}}_{H_s^{(0)},G_s^{(0)}} - \mathrm{AUC}_{H_s^{(0)},G_s^{(0)}}\right|,$$
$$\text{and} \quad \widehat{\Delta}_1 = \sup_{s\in\mathcal{S}} \left|\widehat{\mathrm{AUC}}_{H_s^{(1)},G_s^{(1)}} - \mathrm{AUC}_{H_s^{(1)},G_s^{(1)}}\right|.$$

The triangular inequality implies that: $\sup_{s\in\mathcal{S}} \left|\widehat{L}_\lambda(s) - L_\lambda(s)\right| \leqslant \widehat{\Delta} + \lambda\widehat{\Delta}_0 + \lambda\widehat{\Delta}_1$.

Case of $\widehat{\Delta}$: Note that:

$$\widehat{\mathrm{AUC}}_{H_s,G_s} = (n(n-1)/2n_+n_-) \cdot \widehat{U}_K(s),$$
$$\text{where} \quad \widehat{U}_K(s) = \frac{2}{n(n-1)} \sum_{1\leqslant i<j\leqslant n} K((s(X_i),Y_i,Z_i),(s(X_j),Y_j,Z_j)),$$

and $K((t,y,z),(t',y',z')) = \mathbb{I}\{(y-y')(t-t') > 0\} + (1/2)\cdot\mathbb{I}\{y \neq y', t = t'\}$. The quantity $\widehat{U}_K(s)$ is a known type of statistic and is called a U-statistic, see Definition 3 for the definition and (Lee, 1990) for an overview. We write $U_K(s) := \mathbb{E}[\widehat{U}_K(s)] = 2p(1-p)\mathrm{AUC}_{H_s,G_s}$.

Following Clémençon et al. (2008), we have the following lemma.

Lemma 1. *(Clémençon et al., 2008, Corollary 3) Assume that $\mathcal{S}$ is a VC-major class of functions (see Definition 2) with finite* VC *dimension $V < +\infty$. We have w.p. $\geqslant 1-\delta$: $\forall n > 1$,*

$$\sup_{S\in\mathcal{S}} \left|\widehat{U}_K(s) - U_K(s)\right| \leqslant 2C\sqrt{\frac{V}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n-1}}, \tag{15}$$

where C is a universal constant, explicited in Bousquet et al. (2004, page 198 therein).

Introducing $\widehat{m} := n_+n_-/n^2 - p(1-p)$, we have that, since $\sup_{s\in\mathcal{S}} |\widehat{U}_K(s)| \leqslant 2n_+n_-/(n(n-1))$:

$$\widehat{\Delta} \leqslant \left|\frac{n(n-1)}{2n_+n_-} - \frac{1}{2p(1-p)}\right| \cdot \sup_{s\in\mathcal{S}} \left|\widehat{U}_K(s)\right| + \frac{1}{2p(1-p)} \cdot \sup_{s\in\mathcal{S}} \left|\widehat{U}_K(s) - U_K(s)\right|,$$
$$\leqslant \frac{1}{p(1-p)} \left|\widehat{m} + \frac{n_+n_-}{n^2(n-1)}\right| + \frac{1}{2p(1-p)} \cdot \sup_{s\in\mathcal{S}} \left|\widehat{U}_K(s) - U_K(s)\right|.$$

The properties of the shatter coefficient described in Gyrfi (2002, Theorem 1.12 therein) and the fact that $\mathcal{S}$ is VC major, imply that the class of sets: $\{((x,y),(x',y')) \mid (s(x)-s(x'))(y-y') > 0\}_{s\in\mathcal{S}}$ is VC with dimension V.

The right-hand side term above is covered by Lemma 1, and we deal now with the left-hand side term.

Hoeffding's inequality implies, that w.p. $\geqslant 1-\delta$, we have that, for all $n \geqslant 1$,

$$\left|\frac{n_+}{n} - p\right| \leqslant \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}. \tag{16}$$

Since $n_- = n - n_+$, we have that:

$$\widehat{m} = (1-2p)\left(\frac{n_+}{n} - p\right) - \left(\frac{n_+}{n} - p\right)^2.$$

It follows that:

$$\left|\widehat{m} + \frac{n_+ n_-}{n^2(n-1)}\right| \leqslant |\widehat{m}| + \frac{1}{4(n-1)} \leqslant (1-2p)\sqrt{\frac{\log(2/\delta)}{2n}} + A_n(\delta),$$

where $A_n(\delta) = \frac{\log(2/\delta)}{2n} + \frac{1}{4(n-1)} = O\left(n^{-1}\right)$.

Finally, a union bound between Eq. (15) and Eq. (16) gives that, using the majoration $1/(2n) \leqslant 1/(n-1)$: w.p. $\geqslant 1-\delta$, for any $n > 1$:

$$p(1-p)\cdot\widehat{\Delta} \leqslant C\sqrt{\frac{V}{n}} + 2(1-p)\sqrt{\frac{\log(3/\delta)}{n-1}} + A_n(2\delta/3). \tag{17}$$

Case of $\widehat{\Delta}_0$: Note that:

$$\widehat{\mathrm{AUC}}_{H_s^{(0)},G_s^{(0)}} = \left(n(n-1)/2n_+^{(0)}n_-^{(0)}\right)\cdot \widehat{U}_{K^{(0)}}(s),$$

where $K^{(0)}((t,y,z),(t',y',z')) = \mathbb{I}\{z=0, z'=0\}\cdot K((t,y,z),(t',y',z'))$. We denote:

$$U_{K^{(0)}}(s) := \mathbb{E}[\widehat{U}_{K^{(0)}}(s) = 2q_0^2 p_0(1-p_0)\cdot \mathrm{AUC}_{H_s^{(0)},G_s^{(0)}}.$$

Following the proof of the bound for $\widehat{\Delta}$, introducing $\widehat{m}_0 := n_+^{(0)}n_-^{(0)}/n^2 - q_0^2 p_0(1-p_0)$,

$$\widehat{\Delta}_0 \leqslant \frac{1}{q_0^2 p_0(1-p_0)}\left|\widehat{m}_0 + \frac{n_+^{(0)}n_-^{(0)}}{n^2(n-1)}\right| + \frac{1}{2q_0^2 p_0(1-p_0)}\cdot \sup_{s\in\mathcal{S}}\left|\widehat{U}_{K^{(0)}}(s) - U_{K^{(0)}}(s)\right|.$$

The right-hand side term above is once again covered by Lemma 1. We deal now with the left-hand side term, note that:

$$\begin{aligned}
\widehat{m}_0 &= \frac{n_+^{(0)}n^{(0)}}{n^2} - q_0^2 p_0 - \left(\left[\frac{n_+^{(0)}}{n}\right]^2 - q_0^2 p_0^2\right), \\
&= q_0 p_0\left(\frac{n^{(0)}}{n} - q_0\right) + q_0(1-2p_0)\left(\frac{n_+^{(0)}}{n} - q_0 p_0\right) + \left(\frac{n_+^{(0)}}{n} - q_0 p_0\right)\left(\frac{n^{(0)}}{n} - q_0\right) - \left(\frac{n_+^{(0)}}{n} - q_0 p_0\right)^2.
\end{aligned}$$

A union bound of two Hoeffding inequalities gives that w.p. $\geqslant 1-\delta$, for any $n > 1$, we have simultaneously:

$$\left|\frac{n^{(0)}}{n} - q_0\right| \leqslant \sqrt{\frac{\log(4/\delta)}{2n}} \quad \text{and} \quad \left|\frac{n_+^{(0)}}{n} - q_0 p_0\right| \leqslant \sqrt{\frac{\log(4/\delta)}{2n}}. \tag{18}$$

It follows that:

$$\left|\widehat{m}_0 + \frac{n_+^{(0)}n_-^{(0)}}{n^2(n-1)}\right| \leqslant |\widehat{m}_0| + \left|\frac{\left(n^{(0)}\right)^2}{4n^2(n-1)}\right| \leqslant q_0(1-p_0)\sqrt{\frac{\log(4/\delta)}{2n}} + B_n(\delta),$$

where $B_n(\delta) = \frac{1}{4(n-1)} + \frac{\log(4/\delta)}{n}$.

Finally, a union bound between Eq. (15) and Eq. (18) gives, using the majoration $1/(2n) \leqslant 1/(n-1)$: w.p. $\geqslant 1-\delta$, for any $n > 1$:

$$q_0^2 p_0(1-p_0) \cdot \widehat{\Delta}_0 \leqslant C\sqrt{\frac{V}{n}} + (1 + q_0(1-p_0))\sqrt{\frac{\log(5/\delta)}{n}} + B_n(4\delta/5). \tag{19}$$

Case of $\widehat{\Delta}_1$:

One can prove a similar result as Eq. (19) for $\widehat{\Delta}_1$: w.p. $\geqslant 1-\delta$, for any $n > 1$:

$$q_1^2 p_1(1-p_1) \cdot \widehat{\Delta}_1 \leqslant C\sqrt{\frac{V}{n}} + (1 + q_1(1-p_1))\sqrt{\frac{\log(5/\delta)}{n}} + B_n(4\delta/5). \tag{20}$$

Conclusion:

Under the assumption $\min_{z\in\{0,1\}} \min_{y\in\{-1,1\}} \mathbb{P}\{Y = y, Z = z\} \geqslant \epsilon$, note that $\min(p, 1-p) \geqslant 2\epsilon$. A union bound between Eq. (17), Eq. (19), and Eq. (20). gives that: for any $\delta > 0$ and for all $n > 1$, w.p. $\geqslant 1-\delta$,

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

which concludes the proof.

A.4. Proof of Proposition 1

Consider $f : [0,1] \mapsto [-1,1]$: $f(\alpha) = \mathrm{ROC}_{h,g}(\alpha) - \mathrm{ROC}_{h',g'}(\alpha)$, it is continuous, hence integrable, and with:

$$F(t) = \int_0^t f(\alpha)dt,$$

Note that $F(1) = \mathrm{AUC}_{h,g} - \mathrm{AUC}_{h',g'} = 0 = F(0)$. The mean value theorem implies that there exists $\alpha \in (0,1)$ such that:

$$\mathrm{ROC}_{h,g}(\alpha) = \mathrm{ROC}_{h',g'}(\alpha).$$

A.5. Proof of Corollary 1

Eq. (3)

Assume that s satisfies Eq. (3), Proposition 1 implies that there exists an $\alpha \in (0,1)$, such that:

$$G_s^{(0)} \circ \left(H_s^{(0)}\right)^{-1}(\alpha) = G_s^{(1)} \circ \left(H_s^{(1)}\right)^{-1}(\alpha),$$

Introduce $t_z = (H_s^{(z)})^{-1}(\alpha)$ then $G_s^{(z)}(t_z) = H_s^{(z)}(t_z) = \alpha$ for any $z \in \{0,1\}$, since $H_s^{(z)}$ is increasing. Also,

$$M^{(z)}(g_{s,t_z}) = \mathbb{P}\{g_{s,t_z}(X) \neq Y \mid Z = z\} = p_z G_s^{(z)}(t_z) + (1-p_z)(1 - H_s^{(z)}(t_z)) = (2\alpha - 1)p_z + (1-\alpha),$$

which implies the result.

Eq. (4)

Assume that s satisfies $Eq.$ (4), Proposition 1 implies that there exists an $\alpha \in (0,1)$, such that:

$$G_s^{(0)} \circ H_s^{-1}(\alpha) = G_s^{(1)} \circ H_s^{-1}(\alpha),$$

which translates to:

$$G^{(0)}\left(s(X) \leqslant H_s^{-1}(\alpha)\right) = G^{(1)}\left(s(X) \leqslant H_s^{-1}(\alpha)\right),$$

hence $g_{s,t}$ satisfies fairness in FNR (Eq. (2)) for the threshold $t = H_s^{-1}(\alpha)$.

Eq. (5)

Assume that s satisfies $Eq.$ (5), Proposition 1 implies that there exists an $\alpha \in (0,1)$, such that:

$$G_s \circ (H_s^{(0)})^{-1}(\alpha) = G_s \circ (H_s^{(1)})^{-1}(\alpha),$$

which implies, since G_s, $H_s^{(0)}$ and $H_s^{(1)}$ are increasing:

$$H_s^{(0)} \circ (H_s^{(0)})^{-1}(\alpha) = H_s^{(1)} \circ (H_s^{(0)})^{-1}(\alpha),$$

and:

$$H^{(0)}\left(s(X) > (H_s^{(0)})^{-1}(\alpha)\right) = H^{(1)}\left(s(X) > (H_s^{(0)})^{-1}(\alpha)\right),$$

hence $g_{s,t}$ satisfies fairness in FPR (Eq. (2)) for the threshold $t = (H_s^{(0)})^{-1}(\alpha)$.

Eq. (6)

Assume that s satisfies $Eq.$ (6), Proposition 1 implies that there exists an $\alpha \in (0,1)$, such that:

$$G_s^{(0)} \circ G_s^{-1}(\alpha) = G_s^{(1)} \circ G_s^{-1}(\alpha),$$

which translates to:

$$G^{(0)}\left(s(X) > G_s^{-1}(\alpha)\right) = G^{(1)}\left(s(X) > G_s^{-1}(\alpha)\right),$$

hence $g_{s,t}$ satisfies fairness in FNR (Eq. (2)) for the threshold $t = G_s^{-1}(\alpha)$.

Eq. (10)

Assume that s satisfies $Eq.$ (10), Proposition 1 implies that there exists an $\alpha \in (0,1)$, such that:

$$G_s^{(0)} \circ (F_s^{(0)})^{-1}(\alpha) = G_s^{(1)} \circ (F_s^{(0)})^{-1}(\alpha),$$

which translates to:

$$G^{(0)}\left(s(X) > (F_s^{(0)})^{-1}(\alpha)\right) = G^{(1)}\left(s(X) > (F_s^{(0)})^{-1}(\alpha)\right),$$

hence $g_{s,t}$ satisfies fairness in FNR (Eq. (2)) for the threshold $t = (F_s^{(0)})^{-1}(\alpha)$.

A.6. Proof of Proposition 2

For any $F \in \{H, G\}$, note that:

$$\sup_{\alpha\in[0,1]} |\Delta_{F,\alpha}(s)| \leqslant \max_{k\in\{0,\dots,m\}} \sup_{x\in[\alpha_F^{(k)},\alpha_F^{(k+1)}]} |\Delta_{F,\alpha}(s)|.$$

$\mathrm{ROC}_{F_s^{(0)},F_s^{(1)}}$ is differentiable, and its derivative is bounded by B/b. Indeed, for any $K_1, K_2 \in \mathcal{K}$, since K_1 is continuous and increasing, the inverse function theorem implies that $(K_1)^{-1}$ is differentiable. It follows that $K_2 \circ K_1^{-1}$ is differentiable and that its derivative satisfies:

$$\left(K_2 \circ K_1^{-1}\right)' = \frac{K_2' \circ K_1^{-1}}{K_1' \circ K_1^{-1}} \leqslant \frac{B}{b}.$$

Let $k \in \{0,\dots,m\}$, and $\alpha \in [\alpha_F^{(k)}, \alpha_F^{(k+1)}]$. Since $\alpha \mapsto \Delta_{F,\alpha}(s)$ is continuously differentiable, then α simultaneously satisfies, with the assumption that $|\Delta_{F,\alpha_F^{(k)}}(s)| \leqslant \epsilon$ for any $k \in \{1,\dots,K\}$:

$$|\Delta_{F,\alpha}(s)| \leqslant \epsilon + \left(1 + \frac{B}{b}\right)\left|\alpha_F^{(k)} - \alpha\right| \quad \text{and} \quad |\Delta_{F,\alpha}(s)| \leqslant \epsilon + \left(1 + \frac{B}{b}\right)\left|\alpha - \alpha_F^{(k+1)}\right|,$$

which implies that $|\Delta_{F,\alpha}(s)| \leqslant \epsilon + (1 + B/b)\left|\alpha_F^{(k+1)} - \alpha_F^{(k)}\right|/2$.

Finally, we have shown that:

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

A.7. Proof of Theorem 3

Usual arguments imply that: $L_\Lambda(\widehat{s}_\Lambda) - L_\Lambda(s_\Lambda^*) \leqslant 2 \cdot \sup_{s\in\mathcal{S}} \left|\widehat{L}_\Lambda(s) - L_\Lambda(s)\right|$. As in Appendix A.3, the triangle inequality implies that:

$$\begin{aligned}
\left|\widehat{L}_\Lambda(s) - L_\Lambda(s)\right| &\leqslant \left|\widehat{\mathrm{AUC}}_{H_s,G_s} - \mathrm{AUC}_{H_s,G_s}\right| + \sum_{k=1}^{m_H} \lambda_H^{(k)} \left|\left|\widehat{\Delta}_{H,\alpha_k}(s)\right| - |\Delta_{H,\alpha_k}(s)|\right| + \sum_{k=1}^{m_G} \lambda_G^{(k)} \left|\left|\widehat{\Delta}_{G,\alpha_k}(s)\right| - |\Delta_{G,\alpha_k}(s)|\right|, \\
&\leqslant \left|\widehat{\mathrm{AUC}}_{H_s,G_s} - \mathrm{AUC}_{H_s,G_s}\right| + \sum_{k=1}^{m_H} \lambda_H^{(k)} \left|\widehat{\Delta}_{H,\alpha_k}(s) - \Delta_{H,\alpha_k}(s)\right| + \sum_{k=1}^{m_G} \lambda_G^{(k)} \left|\widehat{\Delta}_{G,\alpha_k}(s) - \Delta_{G,\alpha_k}(s)\right|.
\end{aligned}$$

It follows that:

$$\begin{aligned}
\sup_{s\in\mathcal{S}} \left|\widehat{L}_\Lambda(s) - L_\Lambda(s)\right| \leqslant \sup_{s\in\mathcal{S}} &\left|\widehat{\mathrm{AUC}}_{H_s,G_s} - \mathrm{AUC}_{H_s,G_s}\right| + \bar{\lambda}_H \cdot \sup_{s,\alpha\in\mathcal{S}\times[0,1]} \left|\widehat{\Delta}_{H,\alpha}(s) - \Delta_{H,\alpha}(s)\right| \\
&+ \bar{\lambda}_G \cdot \sup_{s,\alpha\in\mathcal{S}\times[0,1]} \left|\widehat{\Delta}_{G,\alpha}(s) - \Delta_{G,\alpha}(s)\right|,
\end{aligned}$$

and each of the terms is studied independently. The first term is already dealt with in Appendix A.3, and the second and third terms have the same nature, hence we choose to focus on $\widehat{\Delta}_{G,\alpha}(s) - \Delta_{G,\alpha}(s)$.

Note that:

$$\begin{aligned}
\widehat{\Delta}_{G,\alpha}(s) - \Delta_{G,\alpha}(s) &= \widehat{\mathrm{ROC}}_{G_s^{(0)},G_s^{(1)}}(\alpha) - \mathrm{ROC}_{G_s^{(0)},G_s^{(1)}}(\alpha), \\
&= \left[G_s^{(1)} \circ \left(G_s^{(0)}\right)^{-1} - \widehat{G}_s^{(1)} \circ \left(\widehat{G}_s^{(0)}\right)^{-1}\right](1-\alpha), \\
&= \underbrace{\left[G_s^{(1)} \circ \left(G_s^{(0)}\right)^{-1} - G_s^{(1)} \circ \left(\widehat{G}_s^{(0)}\right)^{-1}\right](1-\alpha)}_{T_1(s,\alpha)} + \underbrace{\left[G_s^{(1)} \circ \left(\widehat{G}_s^{(0)}\right)^{-1} - \widehat{G}_s^{(1)} \left(\widehat{G}_s^{(0)}\right)^{-1}\right](1-\alpha)}_{T_2(s,\alpha)}.
\end{aligned}$$

Hence:

$$\sup_{s,\alpha\in\mathcal{S}\times[0,1]} \left|\widehat{\Delta}_{G,\alpha}(s) - \Delta_{G,\alpha}(s)\right| \leqslant \sup_{s,\alpha\in\mathcal{S}\times[0,1]} |T_1(s,\alpha)| + \sup_{s,\alpha\in\mathcal{S}\times[0,1]} |T_2(s,\alpha)|,$$

and we study each of these two terms independently.

Dealing with $\sup_{s,\alpha\in\mathcal{S}\times[0,1]}|T_1(s,\alpha)|$.

Introduce the following functions, for any $z \in \{0,1\}$:

$$\widehat{U}_{n,s}^{(z)}(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i = +1, Z_i = z, s(X_i) \leqslant t\} \quad \text{and} \quad U_{n,s}^{(z)}(t) := \mathbb{E}\left[\widehat{U}_{n,s}^{(z)}(t)\right],$$

then $\widehat{G}_s^{(z)}(t) = (n/n_+^{(z)}) \cdot \widehat{U}_{n,s}^{(z)}(t)$ and $G_s^{(z)}(t) = (1/q_z p_z) \cdot U_{n,s}^{(z)}(t)$ for any $t \in (0,T)$.

The properties of the generalized inverse of a composition of functions, see van der Vaart (2000, see Lemma 21.1, page 304 therein), give, for any $u \in [0,1]$:

$$\left(\widehat{G}_s^{(0)}\right)^{-1}(u) = \left(\widehat{U}_{n,s}^{(0)}\right)^{-1}\left(\frac{n_+^{(0)} u}{n}\right). \tag{21}$$

The assumption on $\mathcal{K}$ implies that $G_s^{(0)}$ is increasing. Define $k_s^{(0)} = G_s^{(0)} \circ s$, for any $t \in (0, T)$, we have:

$$\widehat{U}_{n,s}^{(0)}(t) = \widehat{U}_{n,k_s^{(0)}}\left(G_s^{(0)}(t)\right). \tag{22}$$

Combining Eq. (21) and Eq. (22), we have, for any $u \in [0, 1]$:

$$\left(\widehat{G}_s^{(0)}\right)^{-1}(u) = \left(G_s^{(0)}\right)^{-1} \circ \left(\widehat{U}_{n,k_s^{(0)}}^{(0)}\right)^{-1}\left(\frac{n_+^{(0)} u}{n}\right).$$

Since $G_s^{(0)}$ is continuous and increasing, the inverse function theorem implies that $(G_s^{(0)})^{-1}$ is differentiable. It follows that:

$$\frac{d}{du}\left(G_s^{(1)} \circ (G_s^{(0)})^{-1}(u)\right) = \frac{\left(G_s^{(1)}\right)'\left((G_s^{(0)})^{-1}(u)\right)}{\left(G_s^{(0)}\right)'\left((G_s^{(0)})^{-1}(u)\right)} \leqslant \frac{B}{b},$$

and the mean value inequality implies:

$$\sup_{s,\alpha \in \mathcal{S}\times[0,1]} |T_2(s,\alpha)| \leqslant (B/b) \cdot \sup_{s,\alpha\in\mathcal{S}\times[0,1]} \left| \left(\widehat{U}_{n,k_s^{(0)}}^{(0)}\right)^{-1}\left(\frac{n_+^{(0)}\alpha}{n}\right) - \alpha \right|.$$

Conditioned upon the Z_i's and Y_i's, the quantity

$$\sqrt{n}\left(\left(\frac{n}{n_+^{(0)}}\right)\hat{U}_{n,k_s^{(0)}}^{(0)}(\alpha) - \alpha\right),$$

is a standard empirical process, and it follows from Shorack & Wellner (1989, page 86 therein), that:

$$\sup_{\alpha\in[0,1]} \left| \left(\widehat{U}_{n,k_s^{(0)}}^{(0)}\right)^{-1}\left(\frac{n_+^{(0)}\alpha}{n}\right) - \alpha \right| = \sup_{\alpha\in[0,1]} \left| \frac{n}{n_+^{(0)}} \widehat{U}_{n,k_s^{(0)}}^{(0)}(\alpha) - \alpha \right|.$$

Similar arguments as those seen in Appendix A.3 imply:

$$\begin{aligned}
\sup_{s,\alpha\in\mathcal{S}\times[0,1]} |T_2(s,\alpha)| &\leqslant (B/b) \cdot \sup_{s,\alpha\in\mathcal{S}\times[0,1]} \left| \frac{n}{n_+^{(0)}} \widehat{U}_{n,k_s^{(0)}}^{(0)}(\alpha) - \alpha \right|, \\
&\leqslant \frac{B}{bq_0p_0} \cdot \left| \frac{n_+^{(0)}}{n} - q_0p_0 \right| + \frac{B}{bq_0p_0} \cdot \sup_{s,\alpha\in\mathcal{S}\times[0,1]} \left| \widehat{U}_{n,k_s^{(0)}}^{(0)}(\alpha) - q_0p_0\alpha \right|,
\end{aligned}$$

A standard learning bound, see (Boucheron et al., 2005, Theorem 3.2 and 3.4 therein, page 326-328), implies that: for any $\delta > 0, n > 0$, w.p. $\geqslant 1 - \delta$,

$$\sup_{s,\alpha\in\mathcal{S}\times[0,1]} \left| \widehat{U}_{n,k_s^{(0)}}^{(0)}(\alpha) - U_{n,k_s^{(0)}}^{(0)}(\alpha) \right| \leqslant C\sqrt{\frac{V}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}, \tag{23}$$

where C is a universal constant.

A union bound between Eq. (23) and a standard Hoeffding inequality for $n_+^{(0)}$ gives: for any $\delta > 0, n > 1$, w.p. $\geqslant 1 - \delta$,

$$\sup_{s\in\mathcal{S}} |T_2(s,\alpha)| \leqslant \frac{BC}{bq_0p_0}\sqrt{\frac{V}{n}} + \frac{3B}{bq_0p_0}\sqrt{\frac{\log(4/\delta)}{2n}}. \tag{24}$$

Dealing with $\sup_{s,\alpha\in\mathcal{S}\times[0,1]} |T_2(s,\alpha)|$.

We recall that $\widehat{G}_s^{(z)}(t) = (n/n_+^{(z)}) \cdot \widehat{U}_{n,s}^{(z)}(t)$ and $G_s^{(z)}(t) = (1/q_z p_z) \cdot U_{n,s}^{(z)}(t)$ for any $t \in (0,T)$.

First note that, using the same type of arguments as in Appendix A.3:

$$\begin{aligned}\sup_{s,\alpha \in \mathcal{S}\times[0,1]} |T_1(s,\alpha)| &\leqslant \sup_{s,t \in \mathcal{S}\times(0,T)} \left|\widehat{G}_s^{(1)}(t) - G_s^{(1)}(t)\right|, \\ &\leqslant \frac{1}{q_1 p_1}\left|\frac{n_+^{(1)}}{n} - q_1 p_1\right| + \frac{1}{q_1 p_1} \cdot \sup_{s,t \in \mathcal{S}\times(0,T)} \left|\widehat{U}_{n,s}^{(1)}(t) - U_{n,s}^{(1)}(t)\right|.\end{aligned}$$

The same arguments as for Eq. (23) apply, which means that: for any $\delta > 0, n > 0$, w.p. $\geqslant 1-\delta$,

$$\sup_{s,t \in \mathcal{S}\times(0,T)} \left|\widehat{U}_{n,s}^{(1)}(t) - U_{n,s}^{(1)}(t)\right| \leqslant C\sqrt{\frac{V}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}, \tag{25}$$

where C is a universal constant.

A union bound of Eq. (25) and a standard Hoeffding inequality for $n_+^{(1)}$ finally imply that: for any $\delta > 0, n > 1$,

$$\sup_{s \in \mathcal{S}} |T_1(s,\alpha)| \leqslant \frac{C}{q_1 p_1}\sqrt{\frac{V}{n}} + \frac{3}{q_1 p_1}\sqrt{\frac{\log(4/\delta)}{2n}}. \tag{26}$$

Conclusion.

Combining Eq. (24) and Eq. (26), one obtains that: for any $\delta > 0, n > 1$, w.p. $\geqslant 1-\delta$,

$$\sup_{s,\alpha \in \mathcal{S}\times[0,1]} \left|\widehat{\Delta}_{G,\alpha}(s) - \Delta_{G,\alpha}(s)\right| \leqslant C\left(\frac{1}{q_1 p_1} + \frac{B}{b q_0 p_0}\right)\sqrt{\frac{V}{n}} + \left(\frac{3}{q_1 p_1} + \frac{3B}{b q_0 p_0}\right)\sqrt{\frac{\log(8/\delta)}{2n}}. \tag{27}$$

and a result with similar form can be shown for $\sup_{s,\alpha \in \mathcal{S}\times[0,1]} \left|\widehat{\Delta}_{H,\alpha}(s) - \Delta_{H,\alpha}(s)\right|$ by following the same steps.

Under the assumption $\min_{z\in\{0,1\}} \min_{y\in\{-1,1\}} \mathbb{P}\{Y = y, Z = z\} \geqslant \epsilon$, a union bound between Eq. (27), its equivalent for $\widehat{\Delta}_{H,\alpha}$ and Eq. (17) gives, with the majoration $1/(2n) \leqslant 1/(n-1)$: for any $\delta > 0, n > 1$, w.p. $\geqslant 1-\delta$,

$$\begin{aligned}\epsilon^2 \cdot (L_\Lambda(\widehat{s}_\Lambda) - L_\Lambda(s_\Lambda^*)) \leqslant 2\epsilon &\left(1 + 3(\bar{\lambda}_H + \bar{\lambda}_G)\left[1 + \frac{B}{b}\right]\right)\sqrt{\frac{\log(19/\delta)}{n-1}} \\ &+ C\left(\frac{1}{2} + 2\epsilon(\bar{\lambda}_H + \bar{\lambda}_G)\left[1 + \frac{B}{b}\right]\right)\sqrt{\frac{V}{n}} + O(n^{-1}),\end{aligned}$$

which concludes the proof.

B. Additional Experimental Results and Details

B.1. Details on the Training Algorithms

General principles.

Maximizing directly $\widehat{L}_\lambda$ by gradient ascent (GA) is not feasible, since the criterion is not continuous, hence not differentiable. Hence, we decided to approximate any indicator function $x \mapsto \mathbb{I}\{x > 0\}$ by a logit function $\sigma : x \mapsto 1/(1+e^{-x})$.

We learn with stochastic gradient descent using batches $\mathcal{B}_N$ of N elements sampled with replacement in the training set $\mathcal{D}_n = \{(X_i, Y_i, Z_i)\}_{i=1}^n$, with $\mathcal{B}_N = \{(x_i, y_i, z_i)\}_{i=1}^N$. We assume the existence of a small validation dataset $\mathcal{V}_m$, with $\mathcal{V}_m = \{(x_i^{(v)}, y_i^{(v)}, z_i^{(v)})\}_{i=1}^m$. In practice, one splits a total number of instances $n+m$ between the train and validation dataset.

The approximation of $\widehat{\mathrm{AUC}}_{H_s,G_s}$ on the batch writes:

$$\widetilde{\mathrm{AUC}}_{H_s,G_s} = \frac{1}{N_+ N_-}\sum_{i<j} \sigma\left[(s(x_i) - s(x_j))(y_i - y_j)\right],$$

where $N_+ := \sum_{i=1}^N \mathbb{I}\{y_i = +1\} =: N - N_-$. Similarly, we define $N_+^{(z)} := N^{(z)} - N_-^{(z)}$, with

$$N^{(z)} := \sum_{i=1}^N \mathbb{I}\{z_i = z\} \quad \text{and} \quad N_+^{(z)} := \sum_{i=1}^N \mathbb{I}\{z_i = z, y_i = +1\}.$$

Due to the high number of term involved involved in the summation, the computation of $\widetilde{\mathrm{AUC}}_{H_s,G_s}$ can be very expensive, and we rely on approximations called *incomplete U-statistics*, which simply average a random sample of B nonzero terms of the summation, see (Lee, 1990). We refer to (Clémençon et al., 2016; Papa et al., 2015) for details on their statistical efficiency and use in the context of SGD algorithms. Formally, we define the incomplete approximation with $B \in \mathbb{N}$ pairs of $\widetilde{\mathrm{AUC}}_{H_s,G_s}$ as:

$$\widetilde{\mathrm{AUC}}^{(B)}_{H_s,G_s} := \frac{1}{B} \sum_{(i,j)\in\mathcal{D}_B} \sigma\left[(s(x_i) - s(x_j))(y_i - y_j)\right],$$

where $\mathcal{D}_B$ is a random set of B pairs in the set of all possible pairs $\{(i,j) \mid 1 \leqslant i < j \leqslant N\}$.

For AUC-based constraints (Section 3).

Here, we give more details on our algorithm for the case of the AUC-based constraint Eq. (3). The generalization to other AUC-based fairness constraints is straightforward. For any $z \in \{0,1\}$ the relaxation of $\widetilde{\mathrm{AUC}}_{H^{(z)},G^{(z)}}$ on the batch writes:

$$\widetilde{\mathrm{AUC}}_{H_s^{(z)},G_s^{(z)}} = \frac{1}{N_+^{(z)} N_-^{(z)}} \sum_{\substack{i<j \\ z_i=z_j=z}} \sigma\left[(s(x_i) - s(x_j))(y_i - y_j)\right].$$

Similarly as $\widetilde{\mathrm{AUC}}_{H_s,G_s}$, we introduce the sampling-based approximations $\widetilde{\mathrm{AUC}}^{(B)}_{H_s^{(z)},G_s^{(z)}}$ for any $z \in \{0,1\}$.

To minimize the absolute value in Eq. (11), we introduce a parameter $c \in [-1,+1]$, which is modified slightly every n_{adapt} iterations so that it has the same sign as the evaluation of $\Gamma^\top C(s)$ on $\mathcal{V}_m$. This allows us to write a cost in the form of a weighted sum of AUC's, with weights that vary during the optimization process. Precisely, it is defined as:

$$\widetilde{L}_{\lambda,c}(s) := \left(1 - \widetilde{\mathrm{AUC}}_{H_s,G_s}\right) + \lambda \cdot c \left(\widetilde{\mathrm{AUC}}_{H_s^{(1)},G_s^{(1)}} - \widetilde{\mathrm{AUC}}_{H_s^{(0)},G_s^{(0)}}\right) + \frac{\lambda_{\text{reg}}}{2} \cdot \|W\|_2^2,$$

where λ_{reg} is a regularization parameter and $\|W\|_2^2$ is the sum of the squared L_2 norms of all of the weights of the model. The sampling-based approximation of $\widetilde{L}_{\lambda,c}$ writes:

$$\widetilde{L}^{(B)}_{\lambda,c}(s) := \left(1 - \widetilde{\mathrm{AUC}}^{(B)}_{H_s,G_s}\right) + \lambda \cdot c \left(\widetilde{\mathrm{AUC}}^{(B)}_{H_s^{(1)},G_s^{(1)}} - \widetilde{\mathrm{AUC}}^{(B)}_{H_s^{(0)},G_s^{(0)}}\right) + \frac{\lambda_{\text{reg}}}{2} \cdot \|W\|_2^2.$$

The algorithm is detailed in Algorithm 1, where sng the sign function, *i.e.* $\mathrm{sgn}(x) = 2\mathbb{I}\{x > 0\} - 1$ for any $x \in \mathbb{R}$.

For ROC-based constraints (Section 4).

We define an approximation of the quantities $\widehat{H}_s^{(z)}, \widehat{G}_s^{(z)}$ on $\mathcal{B}_N$, for any $z \in \{0,1\}$, as:

$$\widetilde{H}_s^{(z)}(t) = \frac{1}{N_-^{(z)}} \sum_{i=1}^N \mathbb{I}\{y_i = -1, z_i = z\} \cdot \sigma(t - s(x_i)),$$

$$\widetilde{G}_s^{(z)}(t) = \frac{1}{N_+^{(z)}} \sum_{i=1}^N \mathbb{I}\{y_i = +1, z_i = z\} \cdot \sigma(t - s(x_i)).$$

which can be respectively seen as relaxations of the false positive rate (i.e. $\bar{H}_s^{(z)}(t) = 1 - H_s^{(z)}(t)$) and true positive rate (i.e. $\bar{G}_s^{(z)}(t) = 1 - G_s^{(z)}(t)$) at threshold t and conditioned upon $Z = z$.

For any $F \in \{H, G\}, k \in \{1, \ldots, m_F\}$, we introduce a loss ℓ_F^k which gradients are meant to enforce the constraint $|\widehat{\Delta}_{F,\alpha_F^{(k)}}(s)| = 0$. This constraint can be seen as one that imposes equality between the true positive rates and false positive

Algorithm 1 Practical algorithm for learning with the AUC-based constraint Eq. (3).

Input: training set $\mathcal{D}_n$, validation set $\mathcal{V}_m$
$c \leftarrow 0$
for $i = 1$ **to** n_{iter} **do**
  $\mathcal{B}_N \leftarrow N$ observations sampled with replacement from $\mathcal{D}_n$
  $s \leftarrow$ updated score function using a gradient-based algorithm (*e.g.* ADAM), using the derivative of $\widetilde{L}^{(B)}_{\lambda,c}(s)$ on $\mathcal{B}_N$
  if $(n_{\text{iter}} \bmod n_{\text{adapt}}) = 0$ **then**
    $\Delta\text{AUC} \leftarrow \widehat{\text{AUC}}^{(B_v)}_{H_s^{(1)},G_s^{(1)}} - \widehat{\text{AUC}}^{(B_v)}_{H_s^{(0)},G_s^{(0)}}$ computed on $\mathcal{V}_m$
    $c \leftarrow c + \text{sgn}(\Delta\text{AUC}) \cdot \Delta c$
    $c \leftarrow \min(1, \max(-1, c))$
  end if
end for
Output: score function s

rates for the problem of discriminating between the negatives (resp. positives) of sensitive group 1 against those of sensitive group 0 when $F = H$ (resp. $F = G$). An approximation of this problem's false positive rate (resp. true positive rate) at threshold t is $\widetilde{F}_s^{(0)}(t)$ (resp. $\widetilde{F}_s^{(1)}(t)$). Introduce $c_F^{(k)}$ as a constant in $[-1, +1]$ and $t_F^{(k)}$ as a threshold in $\mathbb{R}$, the following loss $\ell_F^{(k)}$ seeks to equalize these two quantities at threshold $t_F^{(k)}$:

$$\ell_F^{(k)}(s) = c_F^{(k)} \cdot \left(\widetilde{F}_s^{(0)}\left(t_F^{(k)}\right) - \widetilde{F}_s^{(1)}\left(t_F^{(k)}\right) \right).$$

If the gap between $\widehat{F}_s^{(0)}(t_F^{(k)})$ and $\widehat{F}_s^{(1)}(t_F^{(k)})$ — evaluated on the validation set $\mathcal{V}_m$ — is not too large, the threshold $t_F^{(k)}$ is modified slightly every few iterations so that $\widehat{F}_s^{(0)}(t_F^{(k)})$ and $\widehat{F}_s^{(1)}(t_F^{(k)})$ both approach the target value $\alpha_F^{(k)}$. Otherwise, the parameter $c_F^{(k)}$ is slightly modified. The precise strategy to modify $c_F^{(k)}$ and $t_F^{(k)}$ is detailed in Algorithm 2, and we introduce a step Δt to modify the thresholds $t_F^{(k)}$.

The final loss writes:

$$\widetilde{L}_{\Lambda,c,t}(s) := \left(1 - \widetilde{\text{AUC}}_{H_s,G_s}\right) + \frac{1}{m_H} \sum_{k=1}^{m_H} \lambda_H^{(k)} \cdot \ell_H^{(k)}(s) + \frac{1}{m_G} \sum_{k=1}^{m_G} \lambda_G^{(k)} \cdot \ell_G^{(k)}(s) + \frac{\lambda_{\text{reg}}}{2} \cdot \|W\|_2^2,$$

and one can define $\widetilde{L}^{(B)}_{\Lambda,c,t}$ by approximating $\widetilde{\text{AUC}}_{H_s,G_s}$ above by $\widetilde{\text{AUC}}^{(B)}_{H_s,G_s}$. The full algorithm is given in Algorithm 2.

Score function and optimization.

We used a simple neural network of various depth D ($D = 0$ corresponds to linear scorer, while $D = 2$ corresponds to a network of 2 hidden layers) where each layer has the same width d (the dimension of the input space), except for the output layer which outputs a real score. We used ReLU's as activation functions. To center and scale the output score we used *batch normalization* (BN) (see Goodfellow et al., 2016, Section 8.7.1 therein) with fixed values $\gamma = 1, \beta = 0$ for the output value of the network. Algorithm 3 gives a formal description of the network architecture. The intuition for normalizing the output score is that the ranking losses only depend on the relative value of the score between instances, and the more *classification-oriented* losses of ROC-based constraints only depend on a threshold on the score. Empirically, we observed the necessity of renormalization for the algorithm with ROC-based constraints, as the loss $\ell_F^{(k)}$ is zero when $\widehat{F}_s^{(0)}(t_F^{(k)}) = \widehat{F}_s^{(1)}(t_F^{(k)}) \in \{0, 1\}$, which leads to scores that drift away from zero during the learning process, as it seeks to satisfy the constraint imposed by $\ell_F^{(k)}$. All of the network weigths were initialized using a simple centered normal random variable with standard deviation 0.01.

For both AUC-based and ROC-based constraints, optimization was done with the ADAM algorithm. It has an adaptative step size, so we did not modify its default parameters. Refer to (Ruder, 2016) for more details on gradient descent optimization algorithms.

Implementation details.

Algorithm 2 Practical algorithm for learning with ROC-based constraints.

Input: training set $\mathcal{D}_n$, validation set $\mathcal{V}_m$

$c_F^{(k)} \leftarrow 0$ for any $F \in \{H, G\}$, $k \in \{1, \ldots, m_F\}$

$t_F^{(k)} \leftarrow 0$ for any $F \in \{H, G\}$, $k \in \{1, \ldots, m_F\}$

for $i = 1$ **to** n_{iter} **do**

$\mathcal{B}_N \leftarrow N$ observations sampled with replacement from $\mathcal{D}_n$

$s \leftarrow$ updated score function using a gradient-based algorithm (*e.g.* ADAM), using the derivative of $\widetilde{L}^{(B)}_{\Lambda,c,t}(s)$ on $\mathcal{B}_N$

if $(n_{\text{iter}} \bmod n_{\text{adapt}}) = 0$ **then**

for any $F \in \{H, G\}$, $k \in \{1, \ldots, m_F\}$ **do**

$\Delta_F^{(k)} \leftarrow \widehat{F}_s^{(0)}\left(t_F^{(k)}\right) - \widehat{F}_s^{(1)}\left(t_F^{(k)}\right)$ computed on $\mathcal{V}_m$

$\Sigma_F^{(k)} \leftarrow \widehat{F}_s^{(0)}\left(t_F^{(k)}\right) + \widehat{F}_s^{(1)}\left(t_F^{(k)}\right) - 2\alpha_F^{(k)}$ computed on $\mathcal{V}_m$

if $\left|\Sigma_F^{(k)}\right| > \left|\Delta_F^{(k)}\right|$ **then**

$t_F^{(k)} \leftarrow t_F^{(k)} + \text{sgn}\left(\Sigma_F^{(k)}\right) \cdot \Delta t$

else

$c_F^{(k)} \leftarrow c_F^{(k)} + \text{sgn}\left(\Delta_F^{(k)}\right) \cdot \Delta c$

$c_F^{(k)} \leftarrow \min\left(1, \max\left(-1, c_F^{(k)}\right)\right)$

end if

end for

end if

end for

Output: score function s

For all experiments, we set aside 40% of the data for validation, *i.e.* $m = \lfloor 0.40(m+n) \rfloor$ with $\lfloor \cdot \rfloor$ the floor function, the batch size to $N = 100$ and the parameters of the loss changed every $n_{\text{adapt}} = 50$ iterations. For any sampling-based approximation computed on a batch $\mathcal{B}_N$, we set $B = 100$, and $B_{\text{v}} = 10^5$ for those on a validation set $\mathcal{V}_m$. The value Δc was always fixed to 0.01 and Δt to 0.001. We used linear scorers, *i.e.* $D = 0$, for the synthetic data experiments, and networks with $D = 2$ for real data.

The experiments were implemented in Python, and relied extensively on the libraries `numpy`, `TensorFlow` (Abadi et al., 2015), `scikit-learn` (Pedregosa et al., 2011) and `matplotlib` for plots.

B.2. Synthetic Data Experiments

For all of the synthetic data experiments, our objective is to show that the learning procedure recovers the optimal scorer when the dataset is large enough. Each of the 100 runs uses independently generated train, validation and test datasets. The variation that we report on 100 runs hence includes that of the data generation process. For each run, we chose a total of $n + m = 10,000$ points for the train and validation sets and a test dataset of size $n_{\text{test}} = 20,000$. Both algorithms ran for $n_{\text{iter}} = 10,000$ iterations, and with the same regularization strength $\lambda_{\text{reg}} = 0.01$.

Example 1.

The goal of this experiment is to show that we can effectively find trade-offs between ranking accuracy and satisfying Eq. (3). using the procedure described in Algorithm 1.

Algorithm 3 Network architecture.

Input: observation $x = h''_0 \in \mathbb{R}^d$,

for $k = 1$ **to** D **do**

Linear layer: $h_k = W_k^\top h''_{k-1} + b_k$ with $W_k \in \mathbb{R}^{d,d}, b_k \in \mathbb{R}^{d,1}$ learned by GD,

ReLu layer: $h''_k = \max(0, h'_k)$ where $\max$ is an element-wise maximum,

end for

Linear layer: $h_{D+1} = w_{D+1}^\top h''_D + b_{D+1}$ with $w_{D+1} \in \mathbb{R}^{d,1}, b_{D+1} \in \mathbb{R}$ learned by GD,

BN layer: $h'_{D+1} = (h_{D+1} - \mu_{D+1})/\sigma_{D+1}$, with $\mu_{D+1} \in \mathbb{R}, \sigma_{D+1} \in \mathbb{R}$ running averages,

Output: score $s(x)$ of x, with $s(x) = h'_{D+1} \in \mathbb{R}$.

The final solutions of Algorithm 1 with two different values of λ, parameterized by c, are shown in Fig. 3. A representation of the value of the corresponding scorers on $[0,1] \times [0,1]$ is provided in Fig. 4. The median ROC curves for two values of λ over 100 independent runs are shown in Fig. 5, with pointwise 95% confidence intervals.

Example 2.

The goal of this experiment is to show that Algorithm 2 can effectively learn a scorer s for which the α such that the classifier g_{s,t_α} is fair in FPR is specified in advance, and that that solution can be significantly different from the output of runs of Algorithm 1.

For that matter, we compare the solutions of optimizing the AUC without constraint, *i.e.* Algorithm 1 with $\lambda = 0$ with those of Algorithm 1 with $\lambda = 1$ and Algorithm 2 where we impose $\Delta_H(3/4) = 0$ with strength $\lambda_H = 1$. To summarize the results, we introduce the following family of scorers $s_c(x) = -c \cdot x_1 + (1-c) \cdot x_2$, parameterized by $c \in [0,1]$.

Fig. 6 shows that the AUC-based constraint has no effect on the solution, while the ROC-based constraint does and is respected with Algorithm 2. Fig. 7 gives two possible scorers with Algorithm 2. The median ROC curves for two values of Fig. 8 over 100 independent runs are shown in Fig. 5, with pointwise 95% confidence intervals.

B.3. Real data experiments

Databases.

We evaluate our algorithms on four common datasets in the fairness in machine learning literature. Those are:

- The *German Credit Dataset* (German), which was featured in (Zafar et al., 2019; Zehlike et al., 2017; Singh & Joachims, 2019; Donini et al., 2018), and consists in classifying people described by a set of attributes as good or bad credit risks. The sensitive variable is the gender of the individual, *i.e.* male or female. It contains 1,000 instances and we retain 30% of those for testing, and the rest for training/validation.

- The *Adult Income Dataset* (Adult), which was featured in (Zafar et al., 2019; Donini et al., 2018), is based on US census data and consists in predicting whether income exceeds \$50K a year. The sensitive variable is the gender of the individual, *i.e.* male or female. It contains 32.5K observations for training and validation, as well as 16.3K observations for testing. For simplicity, we removed the weights associated to each instance of the dataset.

- The *Compas Dataset* (Compas), which was featured in (Zehlike et al., 2017; Donini et al., 2018), consists in predicting recidivism of convicts in the US. The sensitive variable is the race of the individual, precisely $Z = 1$ if the individual is categorized as African-American and $Z = 0$ otherwise. It contains 9.4K observations, and we retain 20% of those for testing, and the rest for training/validation.

- The *Bank Marketing Dataset* (Bank), which was featured in (Zafar et al., 2019), consists in predicting whether a client will subscribe to a term deposit. The sensitive variable is the age of the individual, i.e $Z = 1$ when the age is between 25 and 60. It contains 45K observations, of which we retain 20% for testing, and the rest for training/validation.

For all of the datasets, we used one-hot encoding for any categorical variables. The number of training instances $n + m$, testing instances n_{test} and covariates per dataset d is summarized in Table 3.

Table 3. Number of observations and of covariates d per dataset.

Dataset	German	Adult	Compas	Bank
$n + m$	700	32.5K	7.5K	36K
n_{test}	300	16.3K	1.9K	9K
d	61	107	16	59

Parameters.

For every run of Algorithm 2, we set:

$$m_H = 2, \quad m_G = 0, \quad \alpha_H^{(1)} = \frac{1}{8}, \quad \alpha_H^{(2)} = \frac{1}{4}, \quad \text{and} \quad \lambda_H^{(1)} = \lambda_H^{(2)} = \lambda_H.$$

For all algorithms, we chose the parameter λ (resp. λ_H) from the candidate set $\in \{0, 0.25, 0.5, 1, 5, 10\}$. Denote by $\widetilde{s}$ the output of Algorithm 1 or Algorithm 2, we selected the parameter λ_{reg} that maximizes the criterion $L_\lambda(\widetilde{s})$ (resp. $L_\Lambda(\widetilde{s})$) on the validation dataset over the following candidate regularization strength set:

$$\lambda_{\text{reg}} \in \{1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}, 5 \times 10^{-1}, 1\}.$$

The validation dataset was only exploited to modify of a few parameters, and we only test 7 candidates for the regularization strength, so selecting λ_{reg} using this data is sensible. All of the numerical evaluations reported below are evaluations on the formerly unused test dataset.

Results of AUC-based constraints - Algorithm 1.

Results are summarized in Table 2 and show that for different values of λ, one obtains different trade-offs between accuracy and fairness. The values of $(\lambda, \lambda_{\text{reg}})$ selected for Table 2 were $\{(0, 0.5), (5, 0.5)\}$ for *German*, $\{(0, 0.05), (0.25, 0.05)\}$ for *Adult*, $\{(0, 0.05), (0.5, 0.05)\}$ for *Compas* and $\{(0, 0.05), (5, 0.5)\}$ for *Bank*.

Results of ROC-based constraints - Algorithm 2.

Results are summarized in Fig. 11. One can see that $\mathrm{ROC}_{H_s^{(0)}, H_s^{(1)}}$ stays close to the diagonal for the values of α around $1/8$ and $1/4$ (in most case, this is true for the whole interval $\alpha \in [0, 0.25]$). This implies that we achieve fairness in false positive rate for the induced classifiers $g_{s,t}$ for a wide range of thresholds t that correspond to the regime of low FPRs. The values of $(\lambda_H, \lambda_{\text{reg}})$ selected for Table 2 and Algorithm 2 in Fig. 11 were $(10, 0.5)$ for *German*, $(0.25, 0.05)$ for *Adult*, $(10, 0.005)$ for *Compas* and $(1, 0.1)$ for *Bank*. The values of $(\lambda, \lambda_{\text{reg}})$ selected for Algorithm 1 in Fig. 11 were the same as those selected for Table 2.

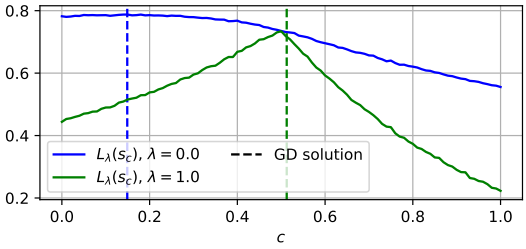
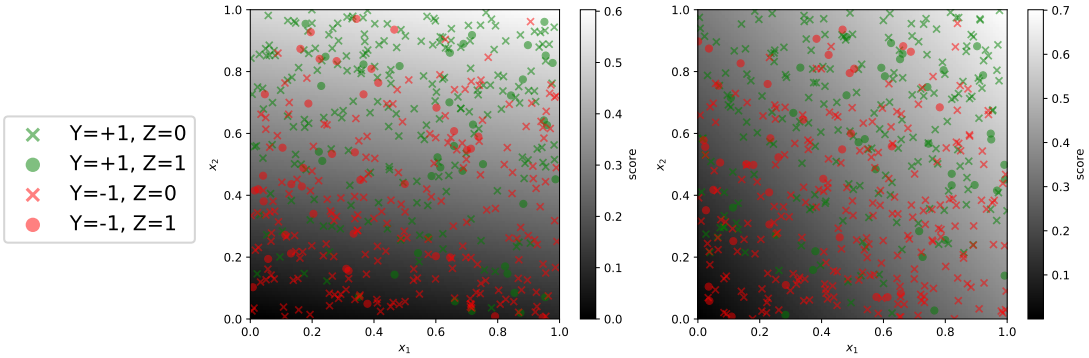
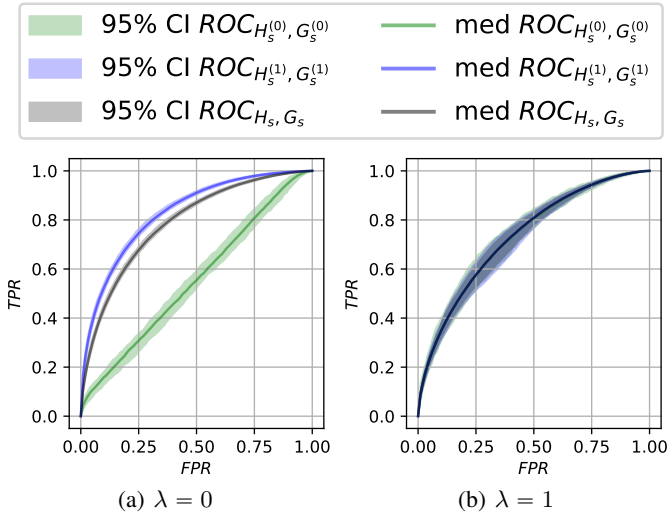


Figure 3. For Example 1, $L_\lambda(s_c)$ as a function of $c \in [0, 1]$ for $\lambda \in \{0, 1\}$, with the parametrization $s_c(x) = cx_1 + (1 - c)x_2$, and the values c for the scores obtained by gradient descent with Algorithm 1.



(a) $\lambda = 0$ (b) $\lambda = 1$

Figure 4. Values of the output score functions on $[0, 1]^2$ for Algorithm 1 ran on Example 1.



(a) $\lambda = 0$ (b) $\lambda = 1$

Figure 5. Result of Example 1 with Algorithm 1.

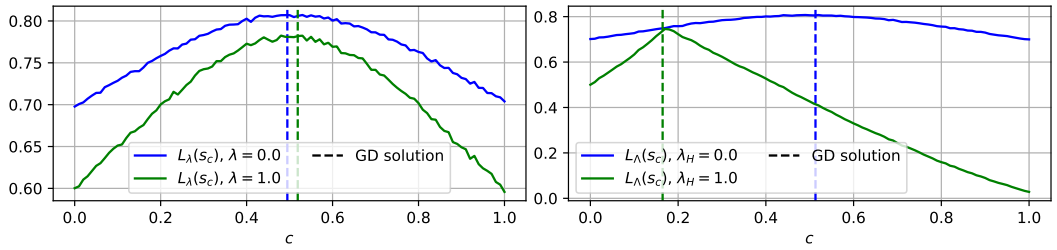
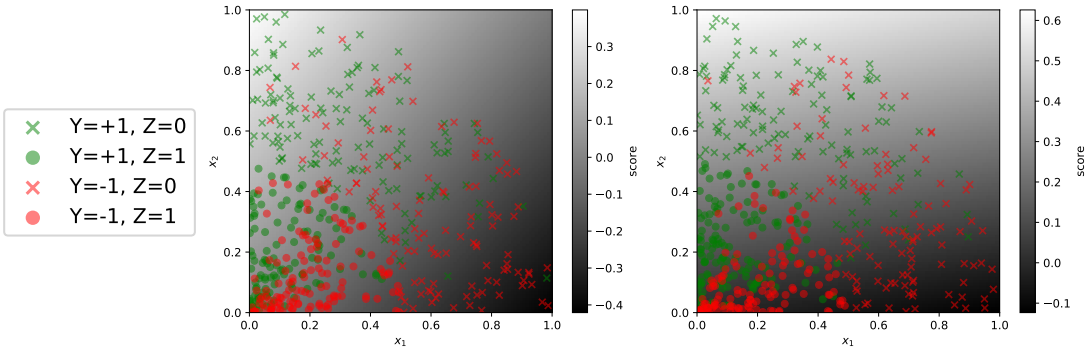
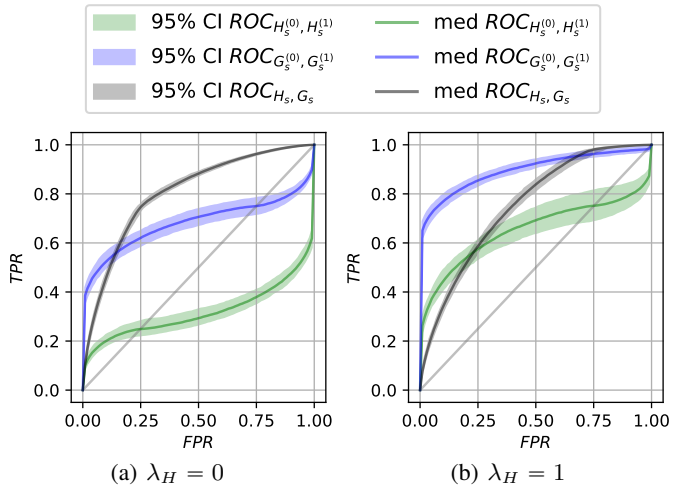


Figure 6. On the left (resp. right), for Example 2, $L_\lambda(s_c)$ (resp. $L_\Lambda(s_c)$) as a function of $c \in [0, 1]$ for $\lambda \in \{0, 1\}$ (resp. $\lambda_H \in \{0, 1\}$), with the parametrization $s_c(x) = -cx_1 + (1 - c)x_2$, and the values c for the scores obtained by gradient descent with Algorithm 1 (resp. Algorithm 2).



(a) $\lambda_H = 0$ (b) $\lambda_H = 1$

Figure 7. Values of the output score functions on $[0, 1]^2$ for Algorithm 2 ran on Example 2.



(a) $\lambda_H = 0$ (b) $\lambda_H = 1$

Figure 8. Result of Example 2 with Algorithm 2.

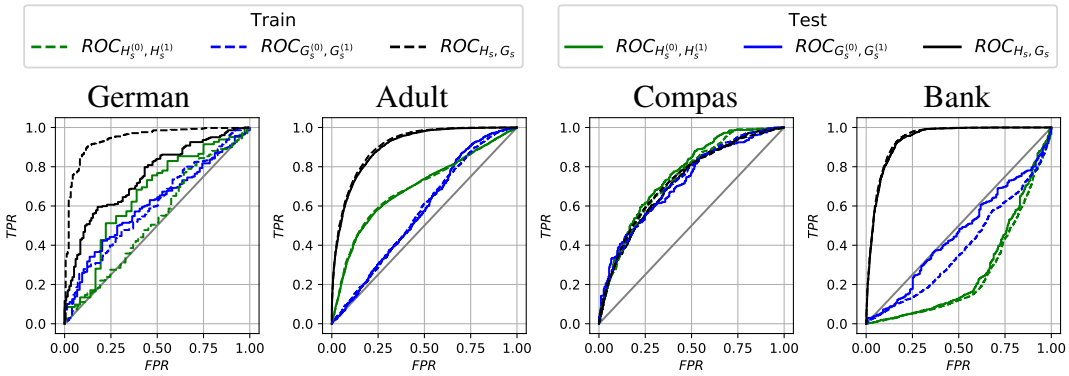


Figure 9. ROC curves for Algorithm 1 with $\lambda = 0$.

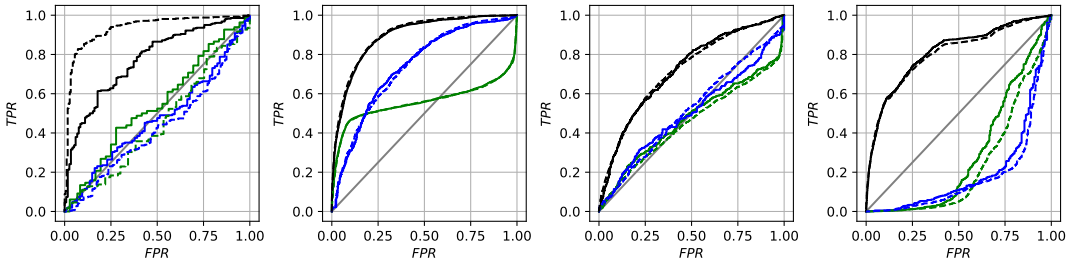


Figure 10. ROC curves for Algorithm 1 with $\lambda > 0$.

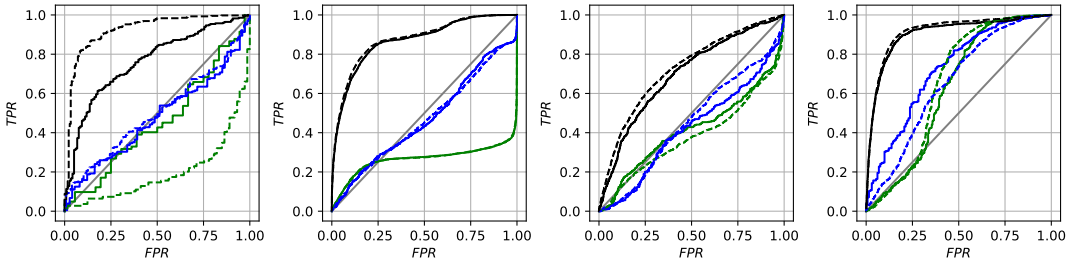


Figure 11. ROC curves for Algorithm 2 with $\lambda_H > 0$.