Similarity Ranking for Biometrics: Theory and Practice

Robin Vogel, PhD defense
LTCI, Télécom Paris & IDEMIA

	d'Alché-Buc Florence	President
	Sebban Marc	Reviewer
	Williamson Robert C.	Reviewer
Jury:	Maillard Oldaric-Ambrym	Examiner
	Valera Isabel	Examiner
	Clémençon Stephan	Supervisor
	Bellet Aurélien	Co-Supervisor
	Gentric Stéphane	Guest
	Despiegel Vincent	Guest

TELECOM Paris

IDEMIA

Outline

Introduction

I - Similarity Ranking

II - Fair Scoring Functions

III - Label Ranking

Perspectives

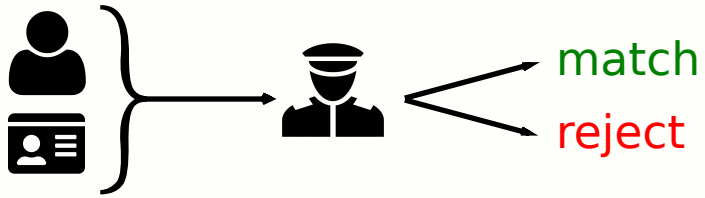
Biometrics in the Real World

Goal of machine learning (ML): to automate complex decisions.

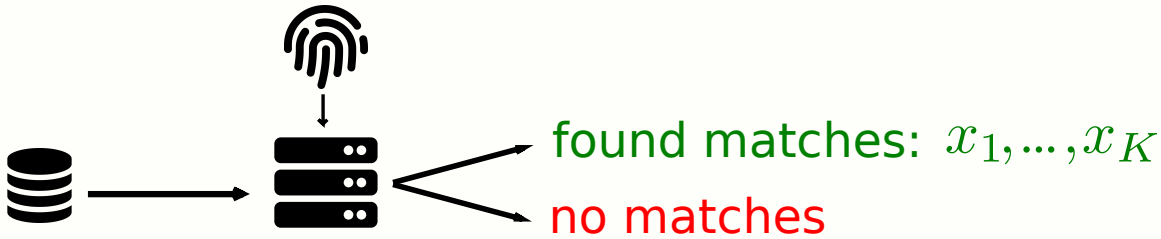
In particular, **Biometrics** automates **personal identification**.

The technology has important social benefits but is controversial.
→ The Aadhaar project vs China's Skynet [Jain et al., 2011].

Ex. 1: **Border control** (1:1 authentification)

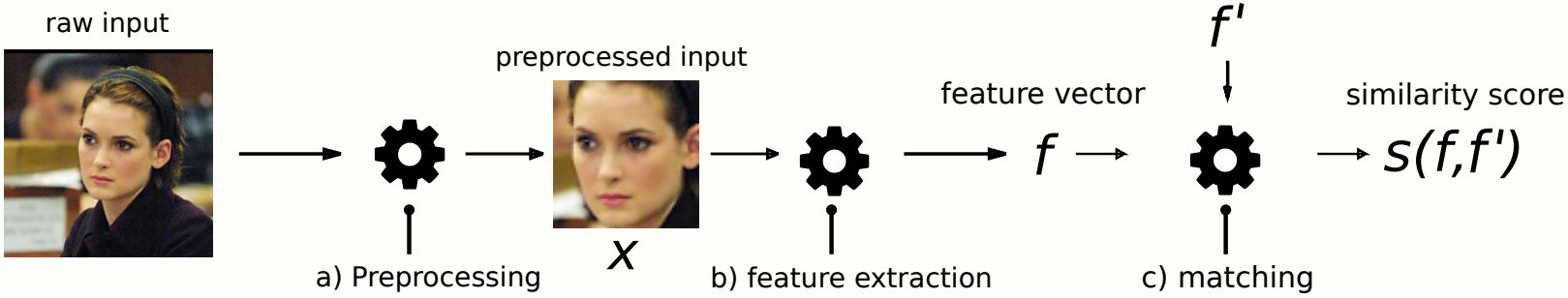


Ex. 2: **Suspect identification** (1:N identification)



Biometric Systems in the Lab

Biometric systems assign a score to a pair of observations (x, x'), with the following process:



Examples in facial recognition (FR):

1995's: a) Face pattern matching — b) EigenFaces — c) distance on (f, f')
2000's: a) Viola-Jones — b) LBP then LDA — c) some distance
2015's: a) MT—CNN — b) ResNet—XX — c) cosine distance

Biometric algorithms involve:

→ similarity learning, → pairwise learning,
→ low-dimensional representations.

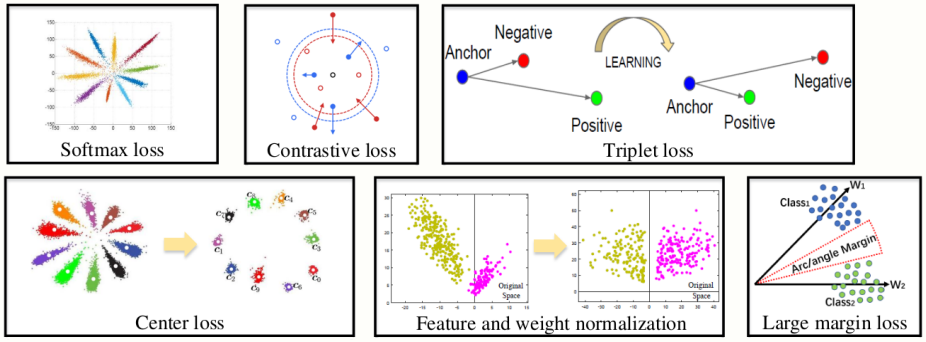
Training Biometric Systems

They are trained with large databases (DB). → large-scale learning.

The nature of the DBs can be very different.
→ cross-domain learning, → sampling bias.

Usual databases (few image per identity)
contain way more neg than pos pairs.
→ subsampling.

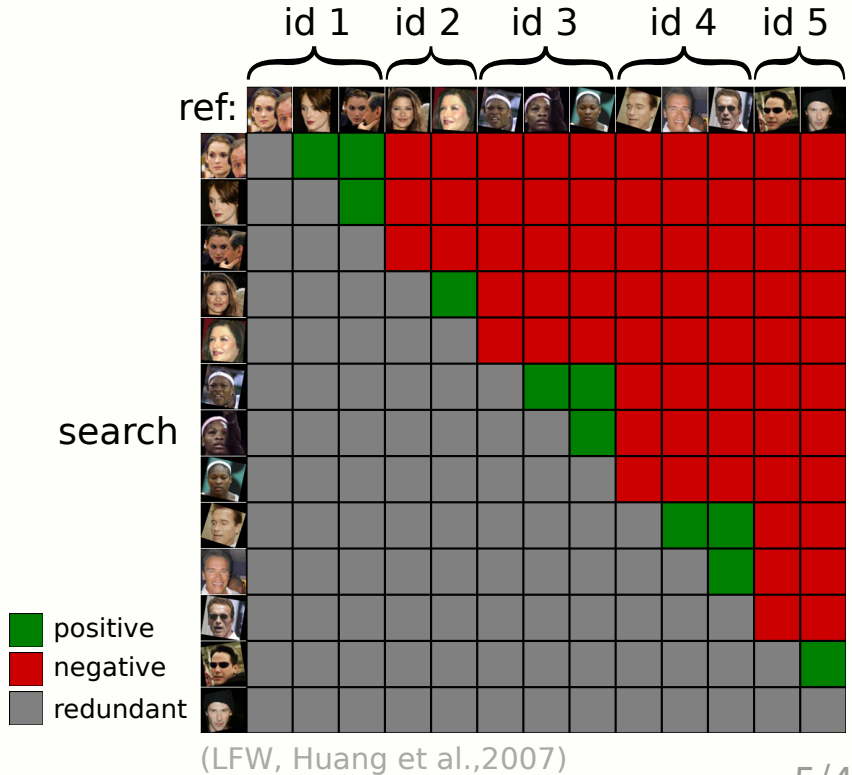
The system seeks to minimize some loss function over the DB.



(Wang et al. 2019)

How to choose the loss ?

Example face images

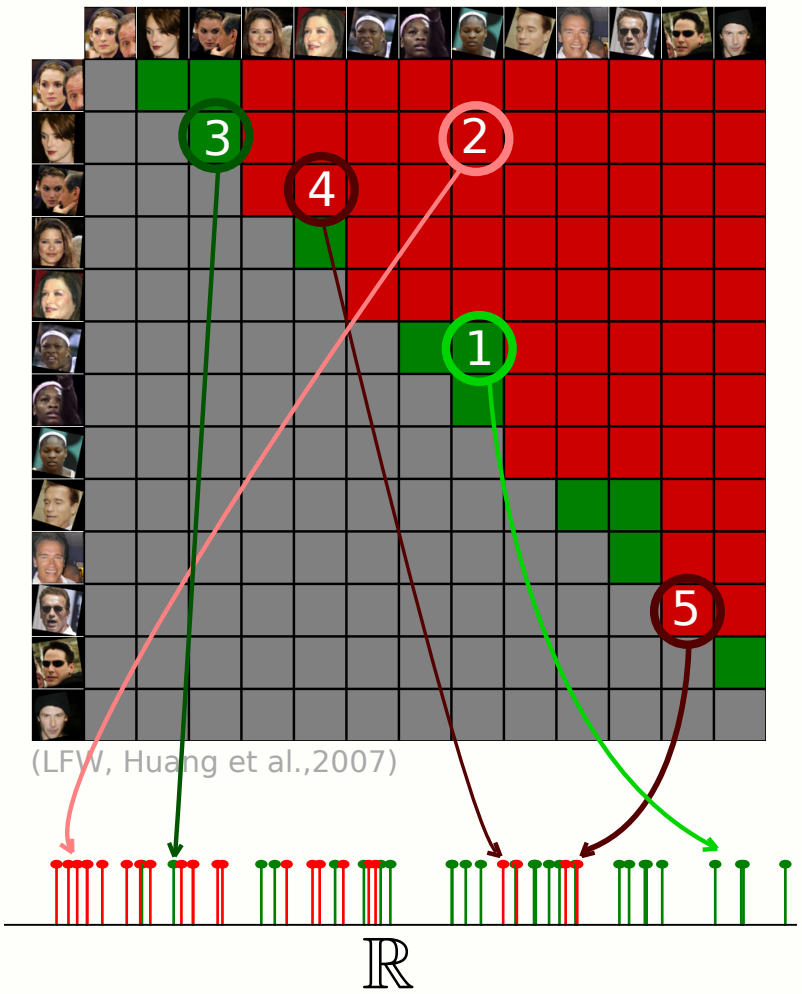


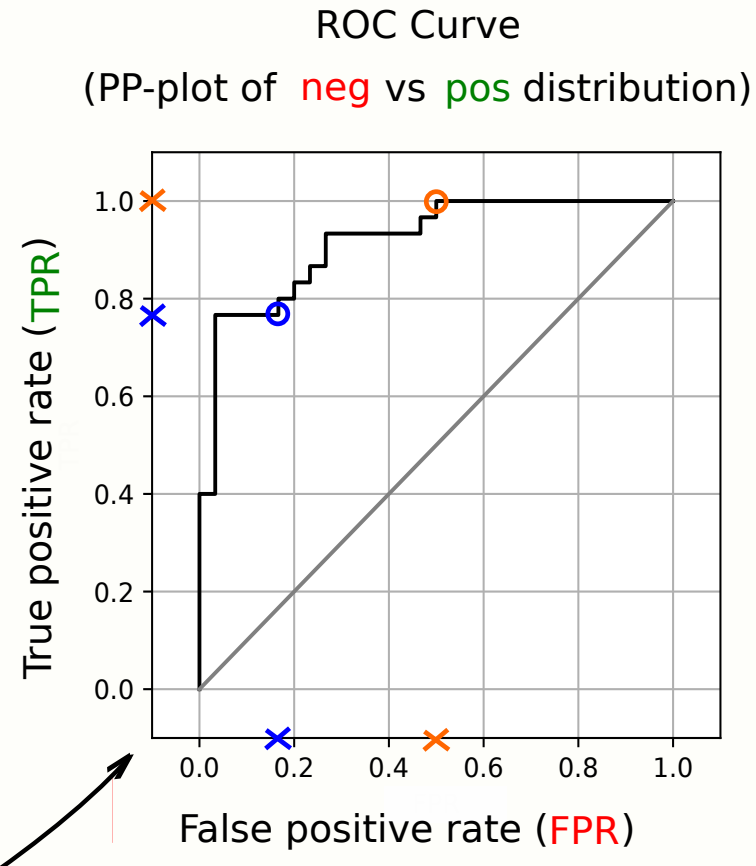
(LFW, Huang et al.,2007)

Evaluating Biometric Systems

The score/similarity function projects pairs on the real line.
The ROC evaluates *s w.r.t.* its ability to split pos and neg by thresholding.

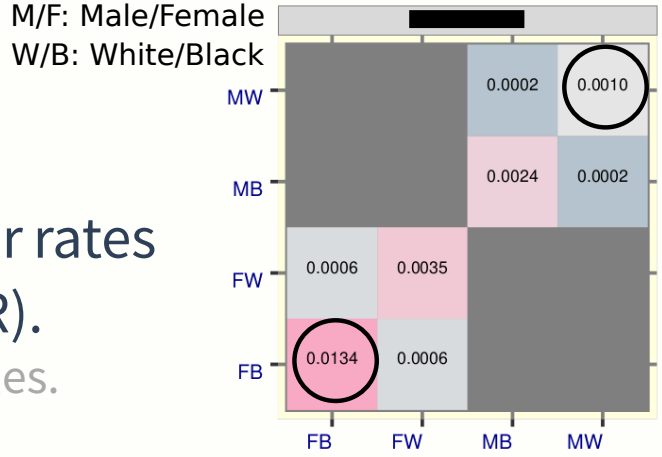
→ bipartite ranking, → theoretical guarantees, → performance criteria.





Fairness in Biometric Systems

FPR for t s.t. $FPR_{MW} = 10^{-3}$



(Grother and Ngan, 2019)

Reports of the NIST show discrepancies in error rates between social groups for face recognition (FR).

At fixed t, 13× more FP for black females than white males.

An explanation is racial bias in face DBs.

Large face DBs:

⇒ are celebrity databases.

⇒ underrepresent some social groups.

⇒ minorities have higher proportion of pos.

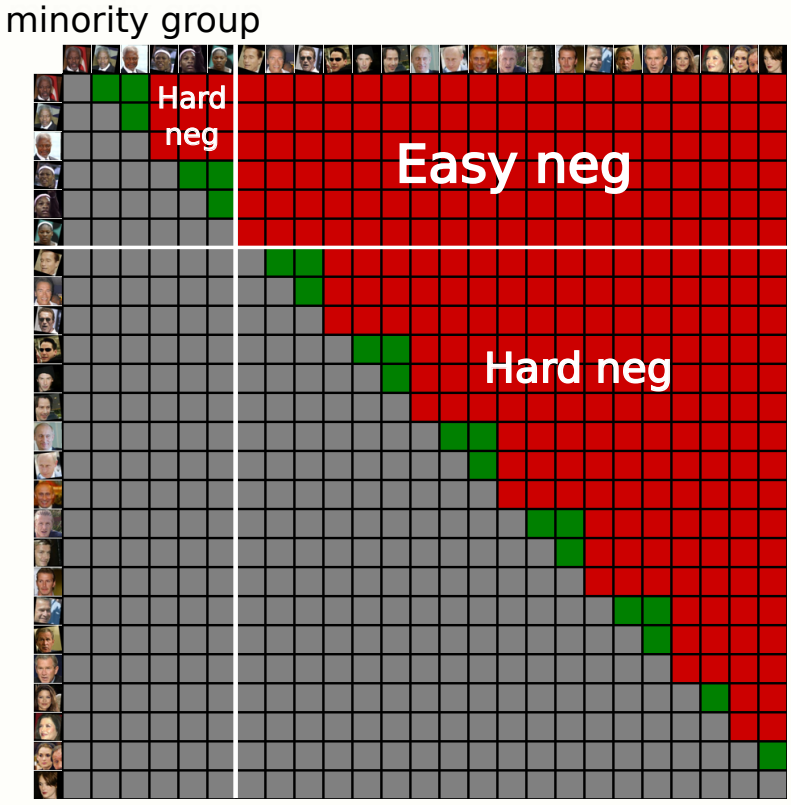
By optimizing for FR performance:

⇒ minorities have higher scores.

⇒ minorities are more likely to falsely match.

⇒ Watchlist surveillance brings racial profiling.

→ fairness.



(LFW, Huang et al.,2007)

Today's Agenda

Part I: We consider **metric learning as bipartite ranking on pairs**. We address pointwise ROC optimization, a natural criterion in biometrics. [Vogel et al., 2018, Clémençon and Vogel, 2019]

Part II: We address **fairness in bipartite ranking**, as a gateway to learning fair similarity functions. [Vogel et al., 2020b]

Part III: We tackle **label ranking**, *i.e.* ordering classes by probability using classification data. [Vogel and Clémençon, 2020]

Outline

Introduction

I - Similarity Ranking
- A Few Facts about Bipartite Ranking
- Pointwise ROC Optimization (pROC) in Similarity Ranking
- TreeRank for Similarity Ranking

II - Fair Scoring Functions

III - Label Ranking

Perspectives

A Probabilistic View of Bipartite Ranking

Objective: Rank observations $X_1, X_2, \ldots$ by order of relevance.

It is done with a **score** $s : \mathcal{X} \to \mathbb{R}$,
learnt with data $\mathcal{D}_n$ that follow **a binary classification model**, *i.e.*

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subset \mathcal{X} \times \mathcal{Y},$$

composed of n i.i.d copies of $(X, Y) \sim P$ where $Y \in \{-1, +1\}$.

Introduce the conditional distributions, for any $t \in \mathbb{R}$,

$$H_s(t) = \mathbb{P}\{s(X) \leqslant t \mid Y = -1\} \text{ and } G_s(t) = \mathbb{P}\{s(X) \leqslant t \mid Y = +1\}.$$

With $\bar{F}(t) = 1 - F(t)$, $\bar{H}_s(t)$ (resp. $\bar{G}_s(t)$) is the FPR (resp. TPR) at t,
and the **ROC curve** writes as:

$$\mathrm{ROC}_{H_s, G_s}(t) : t \mapsto (\bar{H}_s(t), \bar{G}_s(t)).$$

The AUC_{H_s, G_s} is a **scalar summary** of the ROC_{H_s, G_s}.

Bipartite Ranking is NOT Classification

Bipartite ranking separates the neg and pos,
i.e. maximizes (?) the ROC curve, over a family $\mathcal{S}$ of scores.

The **pointwise maximization of the** ROC (**pROC**$_\alpha$) at level α is

$$\max_{s \in \mathcal{S}, t \in \mathbb{R}} \bar{G}_s(t) \quad \text{such that} \quad \bar{H}_s(t) \leqslant \alpha. \tag{1}$$

The **Lemma of Neyman-Pearson** gives optimal solutions for Eq. (1)

$$s^*_\alpha(x) = T \circ \eta(x) \quad \text{and} \quad t^*_\alpha = T \circ Q^*_\alpha,$$

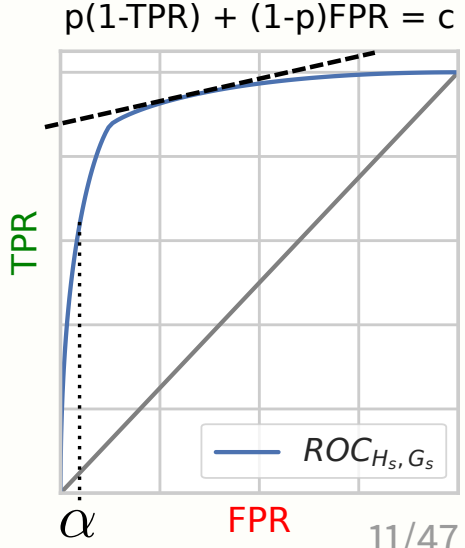
where: Q^*_α is the $(1-\alpha)$-quantile of $\eta(X) \mid Y = -1$,
$\eta(x) = \mathbb{P}\{Y = +1 | X = x\}$ and $T : [0,1] \to \mathbb{R}$ is increasing.

Classification minimizes an error rate $\mathbb{P}\{g(X) \neq Y\}$,
over a family $\mathcal{G}$ of classifiers $g : \mathcal{X} \to \{-1, +1\}$.

The **Bayes classifier** is

$$g^*(x) := 2 \cdot \mathbb{I}\{\eta(x) > 1/2\} - 1.$$

[Devroye et al., 1996]



Related Works and our Contribution

In [Clémençon and Vayatis, 2010, Clémençon and Vayatis, 2009],
$\longrightarrow$ slow and fast finite-sample learning bounds for **pROC**,
$\longrightarrow$ TreeRank, guarantees on its deviation from the optimal ROC.

Metric learning losses are not designed after a ranking objective.

E.g. $$\max_{M \in S_+^d} \sum_{(x_i,x_j)\in\mathcal{D}} d_M(x_i,x_j) \quad \text{s.t.} \quad \sum_{(x_i,x_j)\in\mathcal{S}} d_M^2(x_i,x_j) \leqslant 1.$$

from [Xing et al., 2002]. see [Bellet et al., 2015] for more losses.

> We see metric learning as **pairwise bipartite ranking**, and call that point of view **similarity ranking**.

We extend guarantees for **pROC** to similarity ranking,
as **pROC** is a **natural criterion in biometrics**.

We extend TreeRank to **learning a similarity function**,
as well as the previous guarantees.

Difficulties: Pairwise and functional performance criterion.

Similarity Ranking as Pairwise Bipartite Ranking

Introduce data $\mathcal{D}_n$ that follows a **standard classification model**, i.e.

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subset \mathcal{X} \times \mathcal{Y},$$

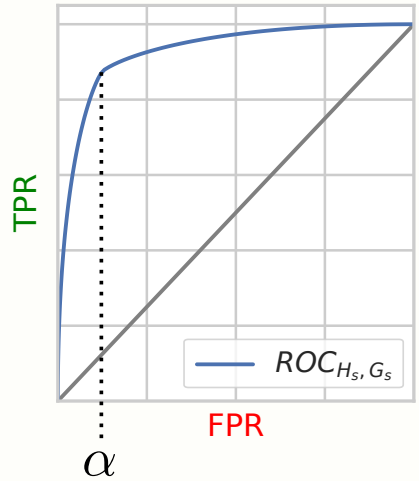
is composed of n i.i.d copies of $(X, Y) \sim P$ where $Y \in \{1, \ldots, K\}$.

- **Bipartite ranking:** Rank $X_1, X_2, \ldots$ by relevance,
- **Similarity ranking:** Rank $(X_1, X_2), (X_1, X_3), \ldots$ by similarity.

Let $(X, Y) \perp (X', Y') \sim P$, the optimal ranking over $\mathcal{X}^2$ is induced by:

$$\eta : x, x' \mapsto \mathbb{P}\{Y = Y' \mid X = x, X' = x'\}.$$

Pointwise ROC Optimization



We introduce the following problem:

$$\max_{K \in \mathcal{K}} R^+(K) \text{ s.t. } R^-(K) \leqslant \alpha, \quad (2)$$

with $R^+(K) = \mathbb{E}\left[K(X, X') \mid Y = Y'\right]$ and $R^-(K) = \mathbb{E}\left[K(X, X') \mid Y \neq Y'\right]$.

Remark: When $K(X, X') = \mathbb{I}\{s(X, X') > t\}$, **Eq. (2)** is **pROC**.

We can estimate the quantities $R^+(K)$ and $R^-(K)$ by:

$$R_n^+(K) := \frac{1}{n_+} \sum_{i<j} K(X_i, X_j) \cdot \mathbb{I}\{Y_i = Y_j\},$$

$$R_n^-(K) := \frac{1}{n_-} \sum_{i<j} K(X_i, X_j) \cdot \mathbb{I}\{Y_i \neq Y_j\},$$

where $n_+ := \sum_{i<j} \mathbb{I}\{Y_i = Y_j\} = n(n-1)/2 - n_-$.

R_n^+ and R_n^- are **ratios of U-statistics**.

Basic Properties of U-statistics

Definition: given an **i.i.d. sample** $\mathcal{Q}_m = \{Z_1, \dots, Z_m\} \subset \mathcal{Z}$, $U_m(K)$ is a U-statistic with **kernel** $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, where:

$$U_m(K) := \frac{2}{m(m-1)} \sum_{i<j} K(Z_i, Z_j).$$

Issue: $K(Z_1, Z_2)$ and $K(Z_1, Z_3)$ are dependent !

First Hoeffding decomposition: [Hoeffding, 1963]

$$U_m(K) = \frac{1}{m!} \sum_{\sigma \in \mathfrak{S}_m} \frac{1}{\lfloor m/2 \rfloor} \sum_{i=1}^{\lfloor m/2 \rfloor} h\left(Z_{\sigma(i)}, Z_{\sigma(i+\lfloor m/2 \rfloor)}\right).$$

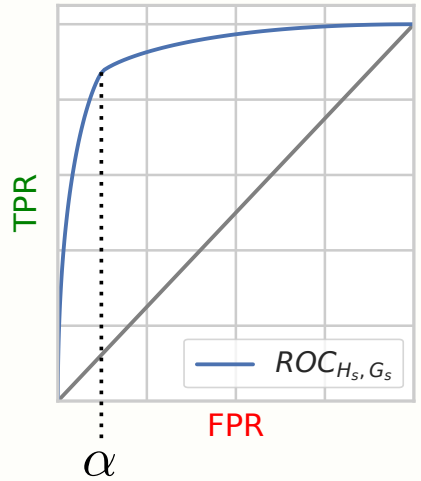
$\Rightarrow$ We can generalize usual results with Jensen's inequality.

Second Hoeffding decomposition: [Hoeffding, 1948]

$$U_m(K) = 2T_m(K) + W_m(K),$$

where $T_m(K)$ is a standard empirical process, $W_m(K)$ a negligible term.

ERM for Pointwise ROC Optimization



Theoretical problem:

$$\max_{K \in \mathcal{K}} R^+(K) \text{ s.t. } R^-(K) \leqslant \alpha, \quad (2)$$

with $R^+(K) = \mathbb{E}\left[K(X, X') \mid Y = Y'\right]$ and $R^-(K) = \mathbb{E}\left[K(X, X') \mid Y \neq Y'\right]$.

Remark: When $K(X, X') = \mathbb{I}\{S(X, X') > t\}$, **Eq. (2)** is **pROC**.

Empirical problem:

$$\max_{K \in \mathcal{K}} R_n^+(K) \text{ s.t. } R_n^-(K) \leqslant \alpha + \Phi. \quad (3)$$

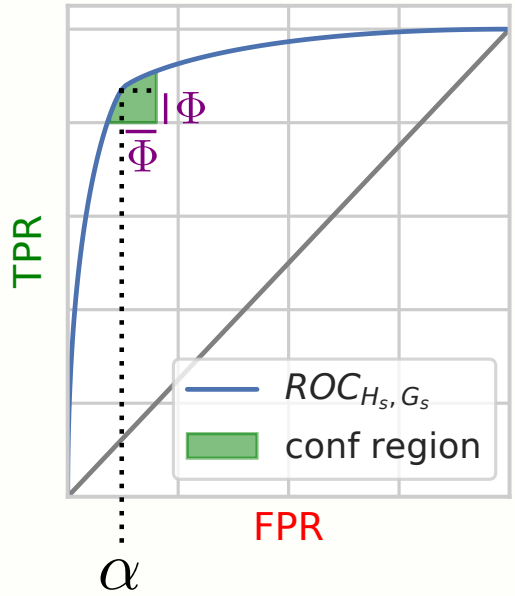
with $\Phi > 0$.

Why a tolerance parameter Φ ?

- Tolerate variations of $R_n^-(K)$ around its expectation $R^-(K)$.

Let K^* and $\widehat{K}_n$ be respectively the solutions of Eq. (2) and Eq. (3).
ERM guarantees prove that $\widehat{K}_n$ is a "decent" solution to Eq. (2).

Universal Guarantee for **pROC**



Theorem 1

Suppose that:

1. $\mathcal{K}$ *has limited complexity (e.g. is a VC-class),*
2. $\mathbb{P}\{Y = Y'\} \geqslant \epsilon > 0.$

Introducing:

$$\Phi_{n,\delta} := C_{\delta,\mathcal{K},\epsilon} \cdot n^{-1/2},$$

we have that, $\forall n \geqslant n_0$*, with probability (w.p.)* $\geqslant 1 - \delta$*,*

$$R^+(\widehat{K}_n) \geqslant R^+(K^*) - \Phi_{n,\delta}, \quad \textit{and} \quad R^-(\widehat{K}_n) \leqslant \alpha + \Phi_{n,\delta}.$$

Proof: Property of [Clémençon and Vayatis, 2010] with results on *U*-statistics.

Novelty:
We adapt Theorem 10 of [Clémençon and Vayatis, 2010]
to similarity ranking, which involves pairwise estimators.

Fast Guarantee for **pROC**

We prove faster bounds under a noise assumption (**NA**) on P.

Theorem 2

*Suppose that assumptions of Theorem 1 and **NA** are verified, we have that,* $\forall a \in (0,1)$ *and* $n \geqslant n_0$*, w.p.* $\geqslant 1-\delta$*,*

$$R^+(\widehat{K}_n) \geqslant R^+(K^*) - C_0 \cdot n^{-(2+a)/4},$$
$$R^-(\widehat{K}_n) \leqslant \alpha + \Phi_{n,\delta}.$$

where C_0 *depends of* $\delta, \epsilon, Q^*_\alpha, a, \mathcal{K}$ *and* $B > 0$*.*

Proof: We combine a bound on the variance of $T_m(K)$ derived from **NA**, Talagrand's inequality, and an usual bound for degenerate U-stats.

Remark: We are slower in n than fast classif. rates (in $n^{-1/(2-a)}$). [Mammen and Tsybakov, 1995]

Novelty:
We adapt Theorem 10 of [Clémençon and Vayatis, 2010] to similarity ranking, which involves U-statistics and a weaker **NA**.

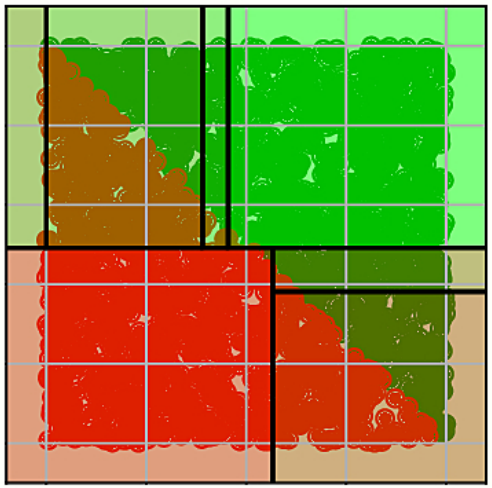
An Extension of TreeRank for Similarity Ranking

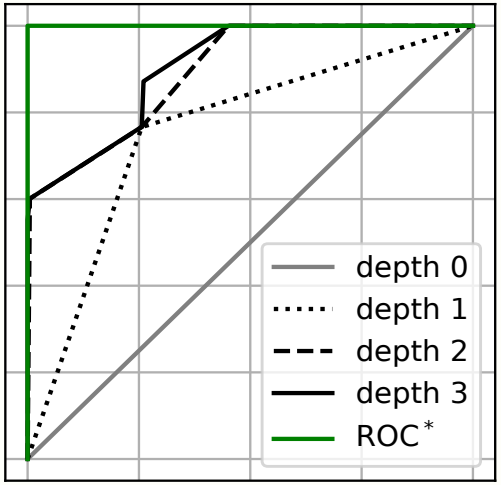
In bipartite ranking,
[Clémençon and Vayatis, 2009] proposed TreeRank, an algorithm that recursively splits the space to optimize for increments in AUC.

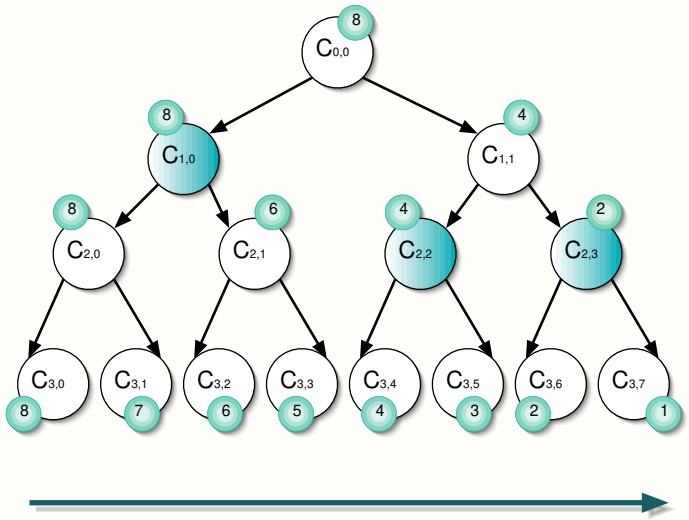
We extend TreeRank to **similarity ranking**, using a symmetric transformation f of the input pair $f : \mathcal{X}^2 \to \mathbb{R}^q$ such that:

$$\text{for all } (x, x') \in \mathcal{X}^2, \quad f(x, x') = f(x', x).$$

and **extend sup norm guarantees** with results on U-statistics.
[Clémençon and Vogel, 2019]







Contributions Recap' - Similarity Ranking

We presented:

1) An extension of the slow and fast finite-sample generalization bounds for **pROC** of bipartite ranking, to similarity ranking.

2) An extension of the TreeRank algorithm to similarity ranking, and of previous finite-sample guarantees on the uniform distance of the ROC of the learnt score to the optimal ROC.

Outline

Introduction

I - Similarity Ranking

II - Fair Scoring Functions
- Fairness in Bipartite Ranking
- ROC -based Fairness
- Experimental Results

III - Label Ranking

Perspectives

Fair ML, beyond Biometrics

Algorithmic decisions are increasingly used in many domains:

Banking (*e.g.* loans) Recruting (*e.g.*, hiring)

Insurance (*e.g.* cars) Judiciary (*e.g.*, bail)

Recently, the fairness of algorithms has gathered lots of attention.

05/2016: The COMPAS system predicts recidivism likelihood for US courts.

Algorithms are designed for the interest of some party,
fairness in ML suggests confronting those to the law.

"Predictive models are really just opinions embedded in math." *C. O'Neil.*

Lack of fairness is not always a consequence of sampling bias.

A good illustration is given by **age performance differences** in biometrics.

See our work on sampling bias [Vogel et al., 2020a].

Fairness Definitions in Binary Classification

A lot of recent works considered fairness in binary classification, with two sensitive groups.
[Donini et al., 2018, Menon and Williamson, 2018, Zafar et al., 2019]

They add a **sensitive variable** $Z \in \{0, 1\}$ to the usual binary classification model (X, Y), and learn $g : \mathcal{X} \to \{-1, +1\}$ from:

$$\mathcal{D}_n = \{(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\} \subset \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}.$$

Many definitions of fairness exist, and apply to specific use-cases.

· Treatment: $g(X, Z) = g(X)$ a.s.
· Impact: $\mathbb{P}\{g(X) = +1 \mid Z = 0\} = \mathbb{P}\{g(X) = +1 \mid Z = 1\}$
· Error: $\mathbb{P}\{g(X) \neq Y \mid Z = 0\} = \mathbb{P}\{g(X) \neq Y \mid Z = 1\}$
· FPR: $\mathbb{P}\{g(X) = +1 \mid Y = -1, Z = 0\} = \mathbb{P}\{g(X) = +1 \mid Y = -1, Z = 1\}$

Fairness Definitions in Bipartite Ranking

Fair ranking is a recent topic, mostly tackled by the IR community.
Authors:

· modify a fixed score to induce fairness [Zehlike et al., 2017, Biega et al., 2018],

· consider fairness in exposure over several rankings [Singh and Joachims, 2019],

· use fairness def° based on AUC's [Borkan et al., 2019, Beutel et al., 2019].

For any $z \in \{0, 1\}$, let:
$$H_s^{(z)}(t) := \mathbb{P}\{s(X) \leqslant t \mid Y = -1, Z = z\},$$
$$G_s^{(z)}(t) := \mathbb{P}\{s(X) \leqslant t \mid Y = +1, Z = z\}.$$

BNSP AUC ([Borkan et al., 2019]): $\mathrm{AUC}_{H_s, G_s^{(0)}} = \mathrm{AUC}_{H_s, G_s^{(1)}}$.

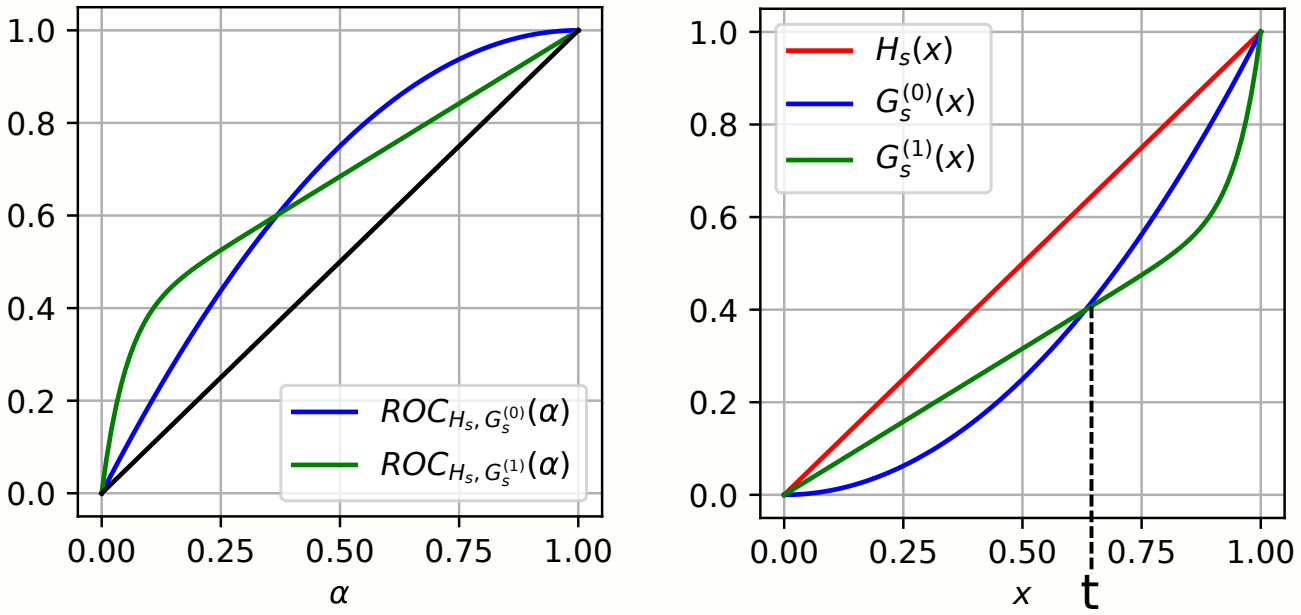
Many similar def° of fairness based on the AUC were proposed.

Contribution:
We proposed an unified framework for learning with AUC constraints.
See section 10.3 of the manuscript.

Limitations of AUC Constraints

Below, with $s \in [0,1]$, we have $\mathrm{AUC}_{H_s, G_s^{(0)}} = \mathrm{AUC}_{H_s, G_s^{(1)}}$.

However, $\sup_{t \in [0,1]} |G_s^{(0)}(t) - G_s^{(1)}(t)| \approx 0.10$.



There exists an **unknown threshold** t, for which the induced classifier $g_{s,t}(x) = \mathrm{sign}(s(x) - t)$ is fair in FNR.

We propose fairness contraints specified on points of ROC's.

Learning with Pointwise ROC Constraints

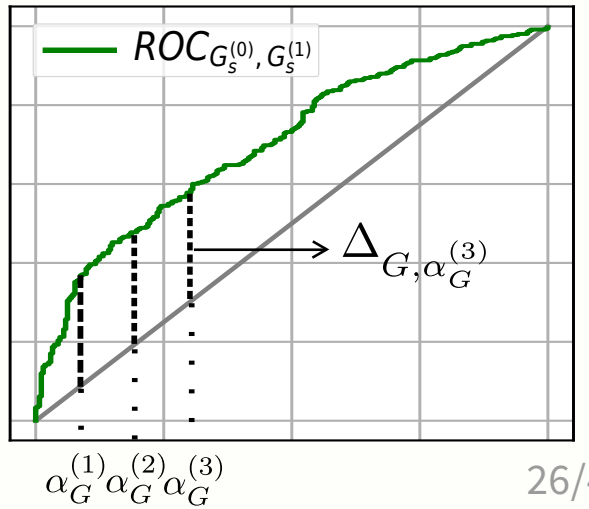
To measure the difference between cdfs for $Z = 0$ and $Z = 1$, let:

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

We introduce a sum of m_H pointwise constraints for $\Delta_{H,.}$ and m_G for $\Delta_{G,.}$ as a penalization, and maximize L_Λ in $\mathcal{S}$, where:

$$L_\Lambda(s) := \mathrm{AUC}_{H_s,G_s} - \sum_{k=1}^{m_H} \lambda_H^{(k)} |\Delta_{H,\alpha_H^{(k)}}(s)| - \sum_{k=1}^{m_G} \lambda_G^{(k)} |\Delta_{G,\alpha_G^{(k)}}(s)|. \quad (4)$$

We prove finite-sample generalization bounds in $O(n^{-1/2})$ for maximizing L_Λ.



Experimental Settings

Here, we report on two datasets from the fairness literature:

· *Compas Dataset*, featured *e.g.* in [Donini et al., 2018],
  Prediction: recidivist or not / sensitive group: ethnicity.

· *Adult Income Dataset*, featured *e.g.* in [Donini et al., 2018],
  Prediction: salary $\geqslant$ \$50K / sensitive group: gender.

AUC-based constraints:

Different constraints are used, depending on the dataset.

ROC-based constraints:

To align the dist. of low FPR's and TPR's between $Z = 0$ and $Z = 1$, we penalize high $|\Delta_{H,1/8}(s)|$, $|\Delta_{H,1/4}(s)|$, $|\Delta_{G,1/8}(s)|$ and $|\Delta_{G,1/4}(s)|$.

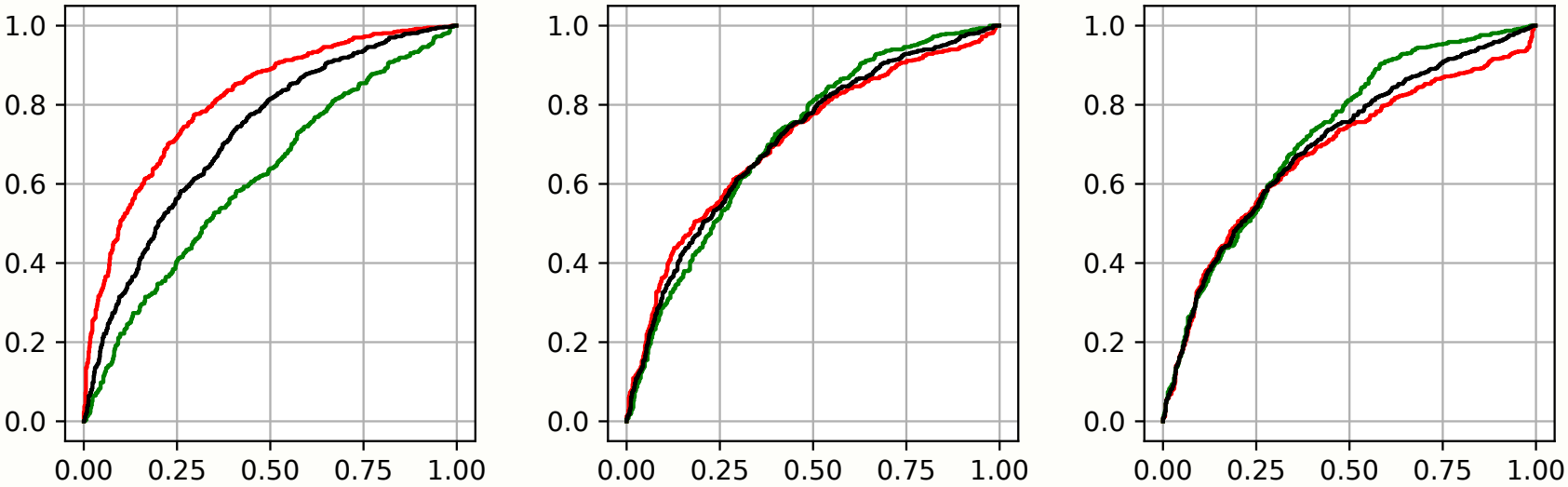
Experimental Results - Compas

0: caucasian
1: ethnic minority

No constraint

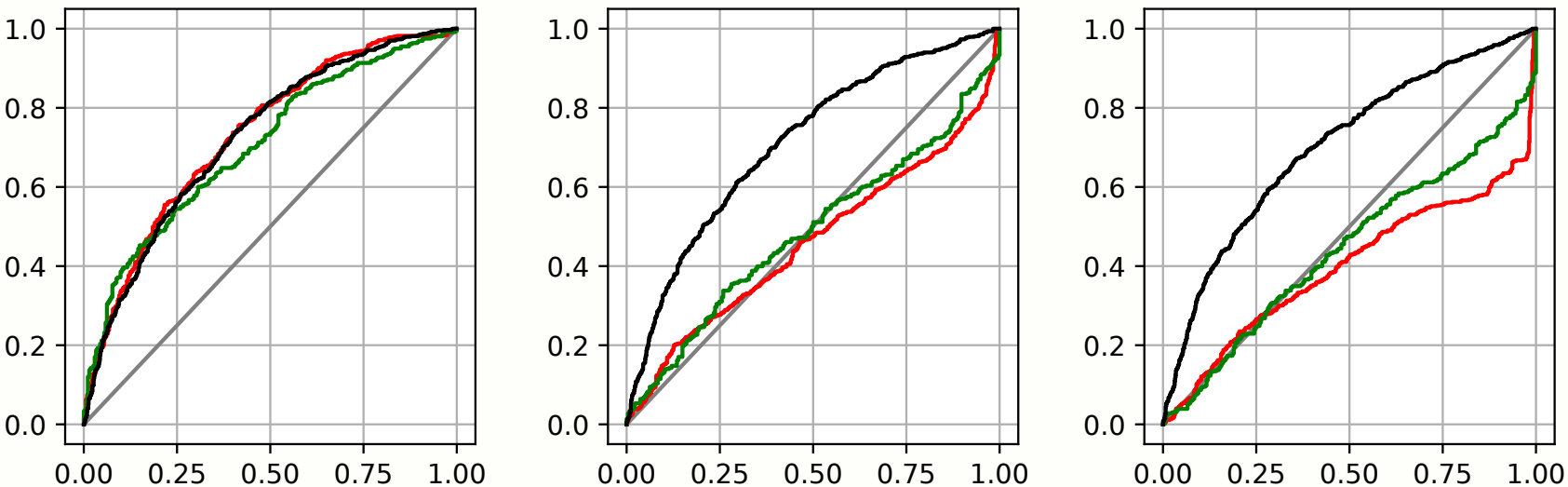
AUC Fairness

ROC Fairness



AUC cons ROCs

ROC_{H_s, G_s} $ROC_{H_s^{(1)}, G_s}$ $ROC_{H_s^{(0)}, G_s}$



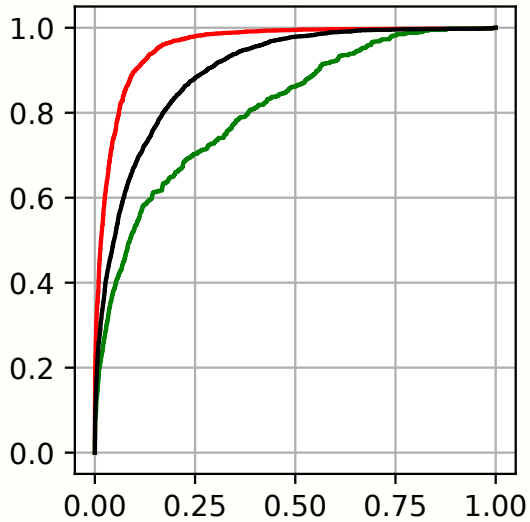
ROC cons ROCs

ROC_{H_s, G_s} $ROC_{G_s^{(0)}, G_s^{(1)}}$ $ROC_{H_s^{(0)}, H_s^{(1)}}$

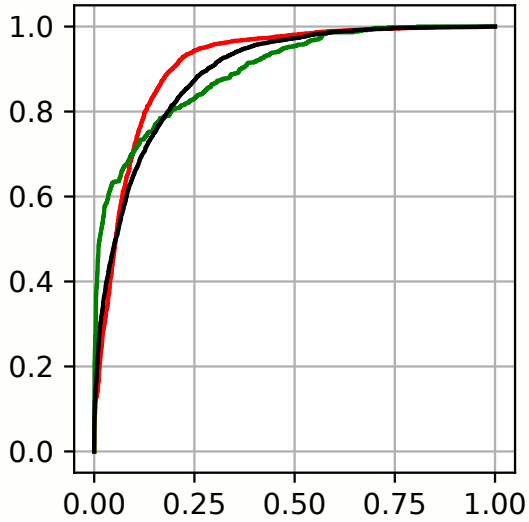
Experimental Results - Adult

0: woman
1: man

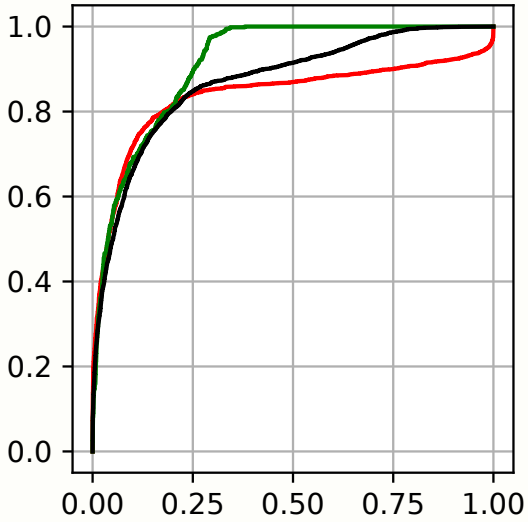
No constraint



AUC Fairness

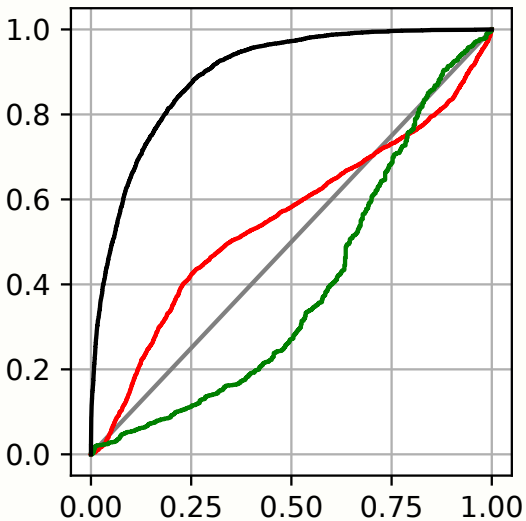


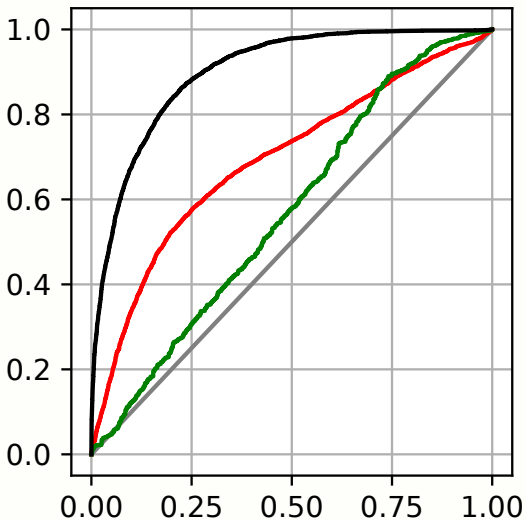
ROC Fairness

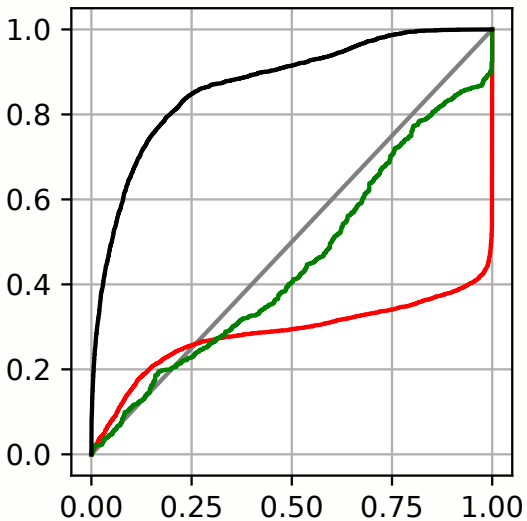


AUC cons ROCs

ROC_{H_s, G_s} $ROC_{H_s^{(1)}, G_s^{(0)}}$ $ROC_{H_s^{(0)}, G_s^{(1)}}$







ROC cons ROCs

ROC_{H_s, G_s} $ROC_{G_s^{(0)}, G_s^{(1)}}$ $ROC_{H_s^{(0)}, H_s^{(1)}}$

Contributions Recap' - Fair Scoring Functions

We presented:

1) A new approach to fairness for bipartite ranking, well-suited to problems that concern specific FPR ranges on the ROC curve.

2) Experimental results for learning score functions under fairness constraints.

Outline

Introduction

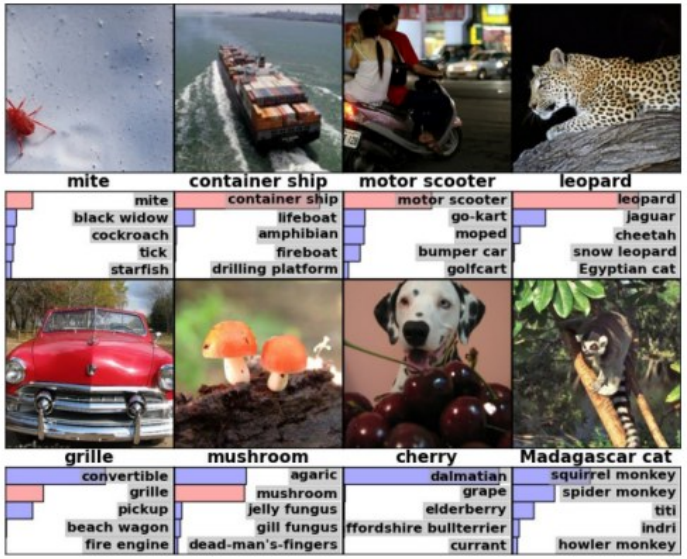
I - Similarity Ranking

II - Fair Scoring Functions

III - Label Ranking
- From Classification to Label Ranking
- Label Ranking as Ranking Median Regression
- Solving Label Ranking with One-vs-One

Perspectives

Most Likely Labels in Classification



[Krizhevsky et al., 2012]

Classification:
Same probabilistic model $(X, Y) \sim P$
as similarity ranking, *i.e.* $\mathcal{Y} = \{1, \ldots, K\}$.
We pick a classifier $g : \mathcal{X} \to \mathcal{Y}$ in a family $\mathcal{G}$.

For hard problems, one returns a list of the **most likely labels**
for an observation $x \in \mathcal{X}$, which concerns many applications.
e.g. biometrics, search engines, ...

One often uses intermediate values of a classification algorithm,
to derive an ordering on $\mathcal{Y}$, *e.g.* softmax probabilities,
evaluates it with a performance indicator, *e.g.* precision at top-k.

Question: How to **explicitly** learn ordered labels from classif. data ?

From Classification to Label Ranking

Introduce the **posterior probabilities** $\eta(x) := (\eta_1(x), \ldots, \eta_K(x))$,
with $\eta_k(x) := \mathbb{P}\{Y = k | X = x\}$.

The classification loss writes: $L(g) := \mathbb{P}\{g(X) \neq Y\}$,
and the **Bayes classifier** g^* writes: $g^*(x) := \underset{k \in \{1,\ldots,K\}}{\arg\max}\, \eta_k(x)$.

Tasks and targets: by order of difficulty:

· **Classification:** the maximum of the $\eta_k(x)$'s, in $\{1, \ldots K\}$.

· **Label Ranking (LR)**: a decreasing order of $\eta(x)$, in $\mathfrak{S}_K$.

· Conditional density estimation: the vector $\eta(x)$, in $\mathbb{R}^K$.

For $x \in \mathcal{X}$, introduce $\sigma_x^* \in \mathfrak{S}_K$, *s.t.*

$$\eta_{\sigma_x^{*-1}(1)}(x) > \eta_{\sigma_x^{*-1}(2)}(x) > \cdots > \eta_{\sigma_x^{*-1}(K)}(x),$$

then $g^*(x) = \sigma_x^{*-1}(1)$.

Our Contributions

LR is related to **Ranking Median Regression** (RMR), i.e. predicting $\Sigma \in \mathfrak{S}_K$ from $X \in \mathcal{X}$, with as data i.i.d. copies of (X, Σ).
e.g. [Vembu and Gärtner, 2010, Tsoumakas et al., 2009].

The One-Versus-One (OVO) is a famous approach in **classification.**.

We see **LR** as **RMR** with only the partial info $Y = \Sigma^{-1}(1)$, and derive a solution to **LR** from the **OVO** procedure.

Our main result is a **fast generalization guarantee** for our solution.

Incidentally, we derive the **first finite-sample bound for OVO**, under assumptions on the distribution of the data.

Papers on RMR with partial info: [Korba et al., 2018, Brinker and Hüllermeier, 2019].

Ranking Median Regression (RMR) (1/2)

Ranking median regression (RMR):
Select $s : \mathcal{X} \to \mathfrak{S}_K$ in the family $\mathcal{S}$ that predicts $\Sigma \in \mathfrak{S}_K$ from X, with $(X, \Sigma) \sim P$. In practice, s minimizes:

$$R(s) := \mathbb{E}[d(s(X), \Sigma)]. \tag{5}$$

where $d : \mathfrak{S}_K \times \mathfrak{S}_K \to \mathbb{R}_+$ symmetric and $d(\sigma, \sigma) = 0, \forall \sigma \in \mathfrak{S}_K$.

Many sensible candidates exist for $d(\sigma, \sigma')$, *e.g.*
· the error: $\mathbb{I}\{\sigma \neq \sigma'\}$,
· the Kendall τ distance d_τ: $\sum_{i<j} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\}$,

Observe that: $$d(\sigma, \sigma') \leqslant \max_{\tau, \tau' \in \mathfrak{S}_K^2} d(\tau, \tau') \times \mathbb{I}\{\sigma \neq \sigma'\}. \tag{6}$$

LR as **RMR** with partial info: we bound $\mathbb{P}\{s(X) \neq \sigma_X^*\}$.

Ranking Median Regression (RMR) (2/2)

Introduce $p_{i,j}(x) = \mathbb{P}\{\Sigma(i) < \Sigma(j) \mid X = x\}$ for $1 \leqslant i < j \leqslant K$.

SST assumption: $\forall (i,k,l), \forall x$, we have $p_{i,j}(x) \neq 1/2$ and

$$p_{i,j}(x) > 1/2 \quad \text{and} \quad p_{j,k}(x) > 1/2 \quad \Rightarrow \quad p_{i,k}(x) > 1/2.$$

Under the **SST assumption** and $d = d_\tau$, the min of Eq. (5) is s^*_X,

$$s^*_X(k) := 1 + \sum_{l \neq k} \mathbb{I}\{p_{k,l}(X) < 1/2\} \qquad \text{for any } k \in \{1, \ldots, K\}.$$

[Clémençon et al., 2018]

LR as **RMR**: we can characterize σ^*_x with the $p_{i,j}$'s.

Label Ranking as Ranking Median Regression

We see **LR** as **RMR** with only the partial info $Y = \Sigma^{-1}(1)$.
Label Ranking: Choose a **ranking rule** $s : \mathcal{X} \to \mathfrak{S}_K$ that minimizes:

$$R(s) := \mathbb{E}[d(s(X), \sigma^*_X)],$$

where σ^*_X **is unobserved.** Same as classif if $d(\sigma, \sigma') = \mathbb{I}\{\sigma^{-1}(1) \neq \sigma'^{-1}(1)\}$.

We can build a r.v. $\Sigma \in \mathfrak{S}_K$, $\Sigma \sim \text{BTLP}(\eta(X))$ and $Y = \Sigma^{-1}(1)$ *a.s.*

$$p_{k,l}(X) = \mathbb{P}\{\Sigma(k) < \Sigma(l) \mid X\} = \frac{\eta_k(X)}{\eta_k(X) + \eta_l(X)} =: \eta_{k,l}(X),$$

where $\eta_{k,l}$ is **the posterior probability for classifying k against l.**

From RMR results: $\sigma^*_X(k) = 1 + \sum_{l \neq k} \mathbb{I}\{\eta_{k,l}(X) < 1/2\}$.

The One-Versus-One Approach for Label Ranking

One-Vs-One (OVO) extends binary classif algos to multiclass classif.
See *e.g.* [Allwein et al., 2000, Wu et al., 2004].

With $g^*_{k,l}(x) := 2 \cdot \mathbb{I}\{\eta_{k,l}(x) \geqslant 1/2\} - 1$,

we have that: $$\sigma^*_X(k) = 1 + \sum_{l \neq k} \mathbb{I}\{g^*_{k,l}(X) = -1\}.$$

OVO procedure for LR: Output $\widehat{\sigma}_X$.

1. Solve all empirical k vs l classification problems, gives $\widehat{g}_{k,l}$'s,
2. Construct $\widehat{\sigma}_X$ with $\widehat{\sigma}_X(k) = 1 + \sum_{k \neq l} \mathbb{I}\{\widehat{g}_{k,l}(X) = -1\}$.

Fast Guarantees for k Versus l Classification

Let $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$ and $Y_{k,l} = \mathbb{I}\{Y = l\} - \mathbb{I}\{Y = k\}$,
$\widehat{L}_{k,l}(g)$ estimates $L_{k,l}(g) := \mathbb{P}\{g(X) \neq Y_{k,l} | Y \in \{k, l\}\}$.

Proposition 3

If: · *the proposed family $\mathcal{G}$ has finite VC-dimension V,*
· *$\exists\ \epsilon > 0$ s.t. for all $k \neq l$ and $x \in \mathcal{X}$, $\eta_k(x) + \eta_l(x) > \epsilon$,*
· *(**NA**) $\exists\ \alpha \in [0, 1]$ and $B > 0$, s.t. for all $k \neq l$ and $t \geqslant 0$,*

$$\mathbb{P}\left\{|2 \cdot \eta_{k,l}(X) - 1| < t\right\} \leqslant B t^{\frac{\alpha}{1-\alpha}}.$$

[Mammen and Tsybakov, 1995]

Then, for any $n \geqslant 1$, w.p. $\geqslant 1 - \delta$,

$$L_{k,l}(\widehat{g}_{k,l}) - L^*_{k,l} \leqslant 2\left(\inf_{g \in \mathcal{G}} L_{k,l}(g) - L^*_{k,l}\right) + r_n(\delta),$$

with $r_n(\delta) = O(n^{-\frac{1}{2-\alpha}})$ and $L^*_{k,l} := L_{k,l}(g^*_{k,l})$.

Proof: Variance control from **NA**, Talagrand's inequality.

Garanties for One-Versus-One for Label Ranking

Remark: NA implies $\mathbb{P}\{g(X) \neq g^*_{k,l}(X)\} \leqslant \beta(L_{k,l}(g) - L^*_{k,l})^\alpha$.

Boole's inequality implies that:

$$\mathbb{P}\{\widehat{\sigma}_X \neq \sigma^*_X\} \leqslant \sum_{k<l} \mathbb{P}\left\{\widehat{g}_{k,l}(X) \neq g^*_{k,l}(X)\right\}, \tag{7}$$

so **NA**, Proposition 3 and Eq. (7) give guarantees for **LR** in $O(n^{-\frac{\alpha}{2-\alpha}})$.
E.g. $\alpha = 1/2$ gives $O(n^{-\frac{1}{3}})$ versus $O(n^{-\frac{1}{2-\alpha}}) = O(n^{-\frac{2}{3}})$.

It gives **guarantees for OVO classifiers**, since: $\bar{g}(X) := \widehat{\sigma}_X{}^{-1}(1)$.

Contributions Recap' - Label Ranking

We presented:

1) A finite-sample bound for a solution to label ranking based on the one-versus-one procedure.

Outline

Introduction

I - Similarity Ranking

II - Fair Scoring Functions

III - Label Ranking

Perspectives

The Optimization of Similarity Ranking Criteria

Limitations:

a) **pROC** is **non-convex** and **non-smooth**, thus hard to optimize.

→ Approximate **pROC** solutions ?

→ Consider other, easier to optimize criteria ? e.g. ranking the best ?

[Rudin, 2006]

b) The **TreeRank** procedure is not well-suited to image data, since it combines several binary classification models.

→ Adapt the procedure with separations that are lighter with depth ?

→ Can bagging correct for the misspecification of the model ?

[Clémençon et al., 2013]

Promising Directions in Fair Scoring

Limitations:
Can we characterize optimal elements for fairness for scoring ?
It would allow us to theoretically study trade-offs between performance and fairness.
[Menon and Williamson, 2018] does it for binary classification.
[Chzhen et al., 2020] gives an optimal solution to fair regression.

Outline

Appendix
- References
- Similarity Ranking
- Fair Scoring Appendix
- Label Ranking Appendix

References I


Allwein, E., Schapire, R., and Singer, Y. (2000).
Reducing multiclass to binary: a unifying approach for margin classifiers.
Journal of Machine Learning Research, 1:113–141.


Bellet, A., Habrard, A., and Sebban, M. (2015).
Metric Learning.
Morgan & Claypool Publishers.


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


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


Blom, G. (1976).
Some properties of incomplete U-statistics.
Biometrika, 63(3):573–580.

References II


Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019).
Nuanced metrics for measuring unintended bias with real data for text classification.
In *Companion of The 2019 World Wide Web Conference, WWW 2019*, pages 491–500. ACM.


Brinker, K. and Hüllermeier, E. (2019).
A reduction of label ranking to multiclass classification.
In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Proceedings, Part III*, volume 11908 of *Lecture Notes in Computer Science*, pages 204–219. Springer.


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


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


Clémençon, S., Korba, A., and Sibony, E. (2018).
Ranking median regression: Learning to order through local consensus.
In *Proceedings of the conference Algorithmic Learning Theory*.


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

References III


Clémençon, S., Depecker, M., and Vayatis, N. (2013).
Ranking Forests.
Journal of Machine Learning Research, 14:39–73.


Clémençon, S. and Vayatis, N. (2010).
Overlaying classifiers: a practical approach for optimal ranking.
In *Constructive Approximation*, number 32, pages 313–320.


Clémençon, S. and Vogel, R. (2019).
On tree-based methods for similarity learning.
In *Machine Learning, Optimization, and Data Science - 5th International Conference, LOD 2019*, volume 11943 of *Lecture Notes in Computer Science*, pages 676–688. Springer.


de la Pena, V. and Giné, E. (1999).
Decoupling: from Dependence to Independence.
Springer.


Devroye, L., Györfi, L., and Lugosi, G. (1996).
A probabilistic theory of pattern recognition.
Springer.


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

References IV


Fisher, R. A. (1936).
The use of multiple measurements in taxonomic problems.
Annals of Eugenics, 7(2):179–188.


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


He, K., Zhang, X., Ren, S., and Sun, J. (2016).
Deep residual learning for image recognition.
In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 770–778. IEEE Computer Society.


Hoeffding, W. (1948).
A class of statistics with asymptotically normal distribution.
The Annals of Mathematical Statistics, 19:293–325.


Hoeffding, W. (1963).
Probability inequalities for sums of bounded random variables.
Journal of the American Statistical Association, 58(301):13–30.

References V


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


Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007).
Labeled faces in the wild: A database for studying face recognition in unconstrained environments.
Technical Report 07-49, University of Massachusetts, Amherst.


Jain, A. K., Ross, A. A., and Nandakumar, K. (2011).
Introduction to Biometrics.
Springer.


Korba, A., Garcia, A., and d'Alché-Buc, F. (2018).
A structured prediction approach for label ranking.
In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 9008–9018.


Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
Imagenet classification with deep convolutional neural networks.
In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, pages 1106–1114.

References VI


Mammen, E. and Tsybakov, A. B. (1995).
Asympotical minimax recovery of the sets with smooth boundaries.
The Annals of Statistics, 23(2):502–524.


Massart, P. and Nédélec, E. (2006).
Risk bounds for statistical learning.
Annals of Statistics, 34(5).


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


Ojala, T., Pietikäinen, M., and Harwood, D. (1994).
Performance evaluation of texture measures with classification based on kullback discrimination of distributions.
In *12th IAPR International Conference on Pattern Recognition, Conference A: Computer Vision & Image Processing, ICPR 1994*, pages 582–585. IEEE.


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

References VII


Shorack, G. and Wellner, J. a. (1989).
Empirical Processes with applications to Statistics.
SIAM.


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


Sung, K.-K. and Poggio, T. (1998).
Example-based learning for view-based human face detection.
IEEE Transactions Pattern Analysis and Machine Intelligence, 20(1):39–51.


Tsoumakas, G., Katakis, I., and Vlahavas, I. (2009).
Mining multi-label data.
In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.


Turk, M. A. and Pentland, A. (1991).
Face recognition using eigenfaces.
In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1991*, pages 586–591. IEEE.


Vembu, S. and Gärtner, T. (2010).
Label ranking algorithms: A survey.
In *Preference learning*, pages 45–64. Springer.

References VIII


Viola, P. A. and Jones, M. J. (2001).
Rapid object detection using a boosted cascade of simple features.
In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, pages 511–518. IEEE Computer Society.


Vogel, R., Achab, M., Clémençon, S., and Tillier, C. (2020a).
Weighted empirical risk minimization: Sample selection bias correction based on importance sampling.
CoRR, abs/2002.05145.


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


Vogel, R., Bellet, A., and Clémençon, S. (2020b).
Learning fair scoring functions: Fairness definitions, algorithms and generalization bounds for bipartite ranking.
CoRR, abs/2002.08159.

References IX


Vogel, R., Bellet, A., Clémençon, S., Jelassi, O., and Papa, G. (2019).
Trade-offs in large-scale distributed tuplewise estimation and learning.
In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Proceedings, Part II*, volume 11907 of *Lecture Notes in Computer Science*, pages 229–245. Springer.


Vogel, R. and Clémençon, S. (2020).
A multiclass classification approach to label ranking.
In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pages 1421–1430. PMLR.


Wu, T., Lin, C., and Weng, R. (2004).
Probability estimates for multi-class classification by pairwise coupling.
Journal of Machine Learning Research, 5:975–1005.


Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. J. (2002).
Distance metric learning with application to clustering with side-information.
In *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002]*, pages 505–512. MIT Press.


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

References X

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

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016).
Joint face detection and alignment using multi-task cascaded convolutional networks.
CoRR, abs/1604.02878.

Detailed Agenda for Today

Part I: We consider metric learning as bipartite ranking on pairs.
(1, 8, 2, 7, 9) [Vogel et al., 2018, Clémençon and Vogel, 2019]

Part II: We address fairness in bipartite ranking.
(11, 8, 9) [Vogel et al., 2020b]

Part III: We propose to learn a ranking rule over classes from classif. data.
(10, 6, 9) [Vogel and Clémençon, 2020]

Topics for today: 1) similarity learning, 2) pairwise learning, 7) subsampling, 8) (bipartite) ranking, 9) theoretical guarantees, 10) performance criteria, 11) fairness.
Topics in other works: 4) large-scale learning, 6) sampling bias.

Topics not covered: 3) low-dimensional repr, 5) cross-domain learning.

Other works (see manuscript):
We extend *U*-statistics approximations strategies to distributed environments. (2, 4, 7, 9) [Vogel et al., 2019]

We propose a reweighting strategy to deal with sampling bias. (6, 9) [Vogel et al., 2020a]

Sketch of Proof of Theorem 1

Based on Theorem 10 of [Clémençon and Vayatis, 2010], we have:

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{K\in\mathcal{K}}|R_n^+(K) - R^+(K)| \leqslant \Phi\right\} + \mathbb{P}\left\{\sup_{K\in\mathcal{K}}|R_n^-(K) - R^-(K)| \leqslant \Phi\right\} - 1 \\
&\quad\leqslant \mathbb{P}\left\{R^+(\widehat{K}_n) \geqslant R^+(K^*) - 2\Phi \quad \text{and} \quad R^-(\widehat{K}_n) \leqslant \alpha + 2\Phi\right\}.
\end{aligned}$$

We need tail bounds for uniform deviations of R_n^+ and R_n^- over $\mathcal{K}$.

We have them ([Hoeffding, 1963]) for both the numerator and denominator of R_n^- and R_n^+.

Careful upper-bounding of:

$$\sup_{K\in\mathcal{K}}|R_n^+(K) - R^+(K)| \quad \text{and} \quad \sup_{K\in\mathcal{K}}|R_n^-(K) - R^-(K)|,$$

conclude the proof.

Sketch of Proof of Theorem 2

NA: for almost every $x \in \mathcal{X}$, $\eta(x, X)$ admits a density bounded by B.
Proof: The proof relies on a refined analysis of R_n^+.

The second Hoeffding decomposition writes U-statistics as $T_n + W_n$, where: [Hoeffding, 1948]
· $T_n = (1/n) \sum_i q_K(X_i, Y_i)$ is a standard mean,
· W_n is a "degenerate" U-statistic.

Degenerate U-statistics decrease quickly in $O(n^{-1})$.
[de la Pena and Giné, 1999]

Under the NA assumption, we have, for any $a \in (0, 1)$:

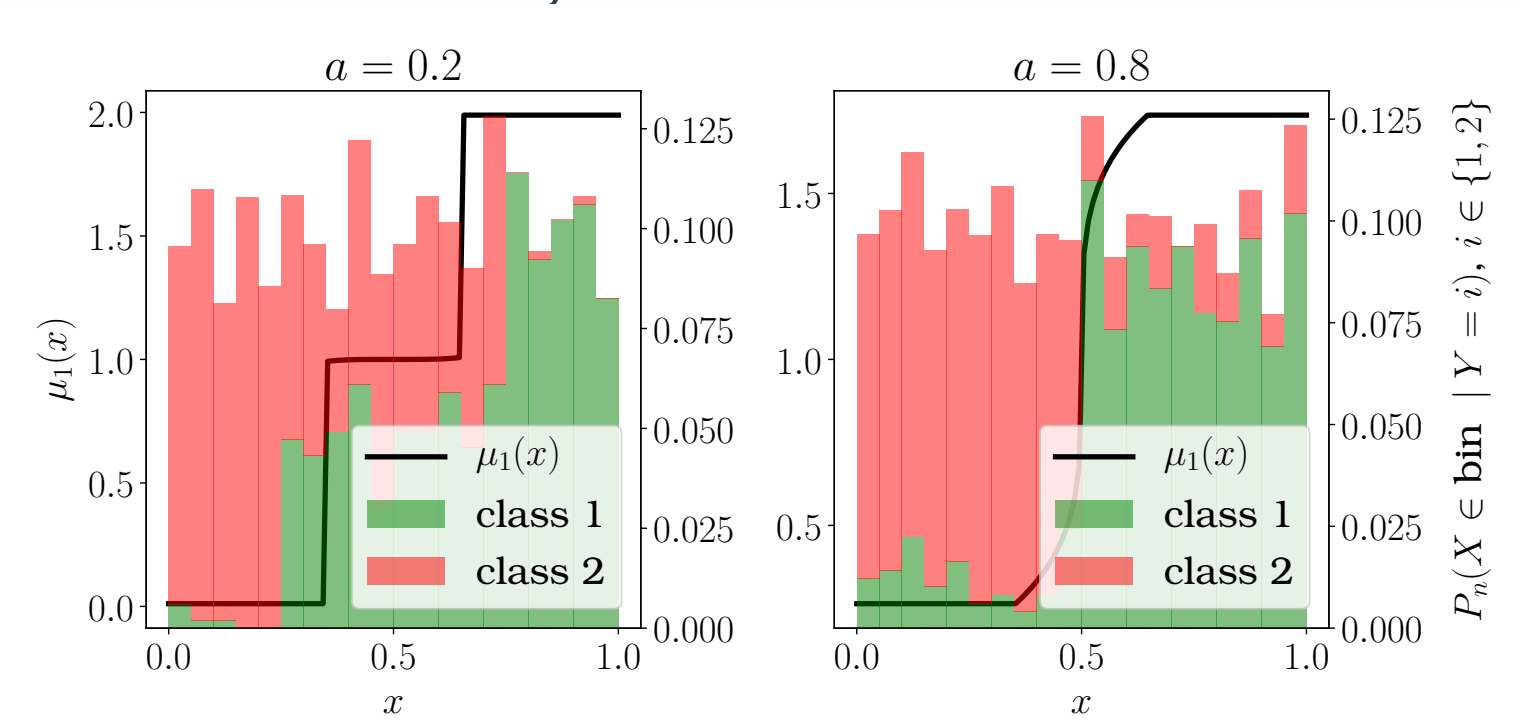
$$\mathbb{E}_{X'} \left[|\eta(X, X') - Q_\alpha^*|^{-a} \right] \leqslant \frac{2B}{1-a} \quad \text{a.s.}$$

The NA assumption implies a bound on the variance of $q_K(X, Y)$, by a function of $R^+(K) - R^+(K^*)$ and $R^-(K) - R^-(K^*)$.

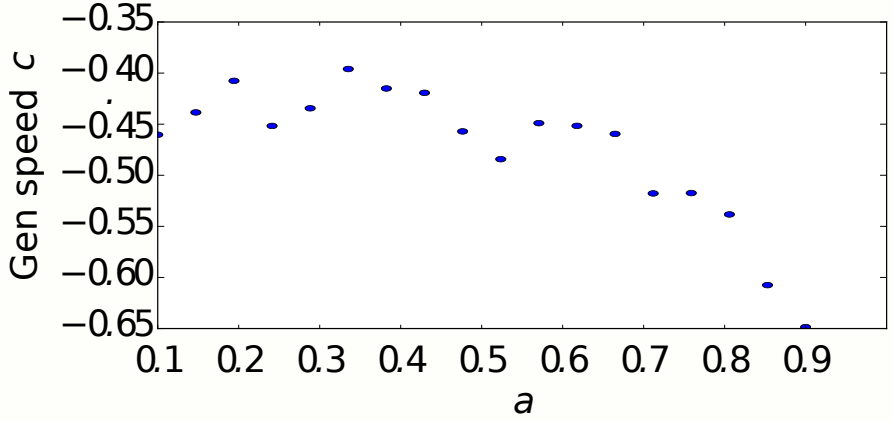
Applying:
· Berstein's inequality for $\mathcal{K}$ finite,
· Talagrand's inequality for more general $\mathcal{K}$,
gives a bound on T_n, which concludes the result.

Illustrating the Fast Rates in Practice

We can choose P to satisfy **NA** with different a's:



And we can compute the generalization rates:



Computational Hurdles to **pROC**

In biometric applications K is high and n/K is constant.
$\Rightarrow R_n^-$ sums $O(n^2)$ pairs.
LFW dataset: $n_+ = 2 \times 10^5$ and $n_- = 9 \times 10^7$.

We could approximate R_n^- on a smaller sample of size $m < n$.
[Blom, 1976] proposed averaging B pairs selected WR from all pairs, to form **incomplete U-statistics**.

[Clémençon et al., 2016] proposed bounds for incomplete U-stats.
Replacing R_n^- in Theorem 1 gives $\Phi = O(B^{-1/2}) + O(n^{-1/2})$.
$\Rightarrow$ Taking $B = O(n)$ gives the usual bound in $O(n^{-1/2})$.

A smaller sample gives $O(n^{-1/4})$ for the same number of pairs.

In [Vogel et al., 2019], we consider the estimation of U-statistics in a distributed environment.

A General Definition for AUC -based Fairness

Introduce all distributions as $D(s) := \left[H_s^{(0)}, H_s^{(1)}, G_s^{(0)}, G_s^{(1)}\right]$.
Any proposed AUC constraint writes as:

$$\mathrm{AUC}_{\alpha^\top D(s), \beta^\top D(s)} = \mathrm{AUC}_{\alpha'^\top D(s), \beta'^\top D(s)}, \tag{8}$$

with α, α', β and β' belong to the 4D probability simplex.

Theorem 4

The following propositions are equivalent:

1. *Eq.* (8) *is verified when* $X|Y = y, Z = 0$ *and* $X|Y = y, Z = 1$ *equal in dist. for any* $y \in \mathcal{Y}$ *and* $\eta(X)$ *is not a.s. constant.*
2. $(e_1 + e_2)^\top [(\alpha - \alpha') - (\beta - \beta')] = 0.$
3. *Eq.* (8) *is equivalent to* $\Gamma^\top C(s) = 0$ *where* $\Gamma \in \mathbb{R}^5$ *and* $C(s) \in \mathbb{R}^5$ *are 5 elementary fairness measures.*

Proof: Decompose carefully Eq. (8) on the 5 elementary fairness measures.

Learning with AUC-based Fairness

Let $\mathcal{S}$ be a proposed family of scores. With an example constraint, integrating the constraint as a penalty gives, with fixed $\lambda > 0$:

$$\max_{s \in \mathcal{S}} L_\lambda(s) \quad \text{with } L_\lambda(s) = \mathrm{AUC}_{H_s, G_s} - \lambda \left| \mathrm{AUC}_{H_s^{(0)}, G_s^{(0)}} - \mathrm{AUC}_{H_s^{(1)}, G_s^{(1)}} \right|,$$

and a solution is written s_λ^*.

The empirical counterpart $\widehat{L}_\lambda$ of L_λ replaces the AUC's above by estimators, using the empirical counterparts of H_s, $H_s^{(z)}$, G_s, $G_s^{(z)}$ on the sample $\mathcal{D}_n = \{(X_i, Y_i, Z_i)\}_{i=1}^n$. Its maximizer is written $\widehat{s}_\lambda$.

Theorem 5

Assume that $\mathcal{S}$ is VC-major ($\{x \mid s(x) > t\}_{s \in \mathcal{S}, t \in \mathbb{R}}$ is VC) with VC-dim V, and there exists $\epsilon > 0$, $\mathbb{P}\{Y = y, Z = z\} \geqslant \epsilon$ for any $y \in \mathcal{Y}, z \in \mathcal{Z}$. Then, for any $n > 1$ and $\delta > 0$, w.p. $\geqslant 1 - \delta$:

$$\epsilon^2 \left[L_\lambda(s_\lambda^*) - L_\lambda(\widehat{s}_\lambda) \right] \leqslant C \sqrt{\frac{V}{n}} + (8\lambda + 2\epsilon) \sqrt{\frac{\log(13/\delta)}{n-1}} + O(n^{-1}).$$

Proof: The empirical AUC's can be dealt with using *U*-statistics.

Guarantee for Learning with Pointwise ROC COnstraints

Maximizing Eq. (4) gives the score s^*_Λ.
The empirical counterpart of L_Λ is $\widehat{L}_\Lambda$, its maximizer is $\widehat{s}_\Lambda$.

Theorem 6

Assume: $\mathcal{S}$ is VC-major ($\{x \mid s(x) > t\}_{s\in\mathcal{S},t\in\mathbb{R}}$ is VC) with VC-dim V,

$$\exists \epsilon > 0, \mathbb{P}\{Y = y, Z = z\} \geqslant \epsilon \text{ for any } y \in \mathcal{Y}, z \in \mathcal{Z},$$

$$\exists M, \kappa > 0, \forall F \in \{H, G\}, z \in \{0, 1\}, s \in \mathcal{S}, M \leqslant F_s^{(z)'} \leqslant M\kappa.$$

Then, for any $n, \delta > 0$, w.p. $\geqslant 1 - \delta$,

$$\epsilon^2 \cdot [L_\Lambda(s^*_\Lambda) - L_\Lambda(\widehat{s}_\Lambda)] \leqslant C_{\lambda,\epsilon,\kappa}\sqrt{\frac{V}{n}} + C'_{\Lambda,\epsilon,\kappa}\sqrt{\frac{\log(19/\delta)}{n-1}} + O(n^{-1}).$$

Sketch of proof of Theorem 6

This result is based on a control of the terms, for both $F \in \{F, G\}$:

$$\sup_{s,\alpha \in \mathcal{S} \times [0,1]} |\widehat{\Delta}_{F,\alpha}(s) - \Delta_{F,\alpha}(s)|,$$

which is the uniform deviation of an empirical ROC.

From [Hsieh and Turnbull, 1996], we have:

$$\widehat{\Delta}_{F,\alpha}(s) - \Delta_{F,\alpha}(s) = \left[F_s^{(1)} \circ F_s^{(0)-1} - F_s^{(1)} \circ \widehat{F}_s^{(0)-1}\right](1-\alpha) \tag{9}$$

$$+ \left[F_s^{(1)} \circ \widehat{F}_s^{(0)-1} - \widehat{F}_s^{(1)} \circ \widehat{F}_s^{(0)-1}\right](1-\alpha). \tag{10}$$

We can bound Eq. (10) by: $\sup_{(s,t) \in \mathcal{S} \times \mathbb{R}} |\widehat{F}_s^{(1)}(t) - F_s^{(1)}(t)|$.

Since the derivative of $F_s^{(1)}$ is bounded, using the mean value theorem, we can bound Eq. (9) by *almost* a quantile process.

The equality of the uniform deviation of a standard and a quantile process [Shorack and Wellner, 1989] implies the result.

A Gradient Descent Approach for Fair Scoring

$\widehat{L}_\lambda / \widehat{L}_\Lambda$ are not continuous, we replace $x \mapsto \mathbb{I}\{x \geqslant 0\}$ by $\sigma(x) = 1/(1+e^{-x})$.
We use ROC criteria, so we normalize s with moving mean and stdev.

For **AUC fairness**, with $c \in [-1, +1]$, we maximize:

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

We change c every n_{adapt} iter, based on stats on a validation dataset:
$\rightarrow$ if the constraint term is positive, then $c = \min(c + \Delta c, 1)$,
$\rightarrow$ otherwise $c = \max(c - \Delta c, -1)$,

For **ROC fairness**, with $c \in [-1, +1]$, we maximize:

$$\widetilde{L}_{\Lambda,c,t}(s) := \widetilde{\mathrm{AUC}}_{H_s, G_s} - \sum_{F \in \{F,G\}} \left(\frac{1}{m_F} \sum_{k=1}^{m_F} \lambda_F^{(k)} \cdot \ell_F^{(k)}(s) \right),$$

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

The $c_F^{(k)}$'s and $t_F^{(k)}$ are changed to have $\widetilde{F}_s^{(0)}\left(t_F^{(k)}\right) = \widetilde{F}_s^{(1)}\left(t_F^{(k)}\right) = \alpha_F^{(k)}$.

Ranking Median Regression (RMR)

With $\mathcal{D}_n = \{(X_i, \Sigma_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$, under a complexity assumption on $\mathcal{S}$, [Clémençon et al., 2018] derives generalization bounds in $O(n^{-1/2})$.

Bounds in $O(n^{-1})$ are derived with an additional noise assumption,

$$\textbf{NA}: \quad H = \text{ess inf } \min_{i<j} |p_{i,j}(X) - 1/2| > 0.$$

NA resembles **Massart's condition** for binary classification. [Massart and Nédélec, 2006]

Under **NA**, $$\mathbb{P}\{s(X) \neq s^*_X\} \leqslant (1/H) \times (R(s) - R(s^*_X)). \tag{11}$$

Bounding the excess risk gives bounds on $\mathbb{P}\{s(X) \neq s^*_X\}$.

BTLP model

Definition 7 (Conditional Bradley-Terry-Luce-Plackett (BTLP))

The conditional distribution of Σ^{-1} given X is defined by recursion: given a **hidden preference vector** $w(X) = (w_1(X), \ldots, w_K(X))$, set $S_1 := \{1, \ldots, K\}$, $S_k := S_1 \backslash \{\Sigma^{-1}(1), \ldots, \Sigma^{-1}(k-1)\}$ and:

$$\Sigma^{-1}(k) \sim \mathcal{M}\left(1, \left\{ \frac{w_l(X)}{\sum_{m \in S_k} w_m(X)} \;\middle|\; l \in S_k \right\}\right),$$

with $\mathcal{M}(n, p)$ is the multinomial dist. We write $\Sigma \sim \mathrm{BTLP}(w(X))$.

Formal Theorem for Label Ranking Guarantees

Theorem 8

Under the same assumptions as those of Proposition 3, for all $\delta \in (0,1)$ and $n \geqslant n_0(\delta, \alpha, \epsilon, B, V)$, w.p. $\geqslant 1-\delta$,

$$\mathbb{P}\{\hat{\sigma}_X \neq \sigma_X^*\} \leqslant \frac{\beta}{\epsilon}\left(\sum_{k<l} 2\left(\inf_{g\in\mathcal{G}} L_{k,l}(g) - L_{k,l}^*\right)^{\alpha} + \binom{K}{2} r_n^{\alpha}\left(\frac{\delta}{\binom{K}{2}}\right)\right).$$

Guarantees with OVO for top-k and classification

The top-k loss writes $\ell_k(y, \sigma) = \mathbb{I}\{y \notin \{\sigma^{-1}(1), \dots, \sigma^{-1}(k)\}\}$,
and the top-k risk of $s : \mathbb{R}^q \to \mathfrak{S}_K$ is: $W_k(s) = \mathbb{E}\left[\ell_k(Y, s(X))\right]$.

We prove a tighter bound than Theorem 8 for top-k classif with **OVO**.

Proposition 9

Let $k \in \{1, \dots, K\}$, $W_k^* := W_k(\sigma_X^*)$. Under the assumptions of Prop 3, for any $\delta \in (0,1)$ and $n \geqslant n_1(\delta, \alpha, \epsilon, B, V)$, w.p. $\geqslant 1 - \delta$,

$$W_k(\hat{\sigma}_X) - W_k^* \leqslant \frac{\beta}{\epsilon} C_{K,k} \left(2 \max_{m \neq l} \left(\inf_{g \in \mathcal{G}} L_{l,m}(g) - L_{l,m}^* \right)^{\alpha} + r_n^{\alpha} \left(\frac{\delta}{\binom{K}{2}} \right) \right).$$

Proof: relies on: $W_k(s) - W_k^* \leqslant \mathbb{P}\left\{\mathrm{Top}_k(\hat{\sigma}_X) \neq \mathrm{Top}_k(\sigma_X^*)\right\}$, reduces again to k vs l.

For $k = 1$, it gives **guarantees for OVO classifiers**: $\bar{g}(X) := \hat{\sigma}_X{}^{-1}(1)$.