Ranking Data with Continuous Labels through Oriented Recursive Partitions

Stephan Clémençon **Mastane Achab**
LTCI, Télécom ParisTech, Université Paris-Saclay
75013 Paris, France
first.last@telecom-paristech.fr

Abstract

We formulate a supervised learning problem, referred to as *continuous ranking*, where a continuous real-valued label Y is assigned to an observable r.v. X taking its values in a feature space $\mathcal{X}$ and the goal is to order all possible observations x in $\mathcal{X}$ by means of a scoring function $s : \mathcal{X} \to \mathbb{R}$ so that $s(X)$ and Y tend to increase or decrease together with highest probability. This problem generalizes *bi/multi-partite ranking* to a certain extent and the task of finding optimal scoring functions $s(x)$ can be naturally cast as optimization of a dedicated functional criterion, called the IROC curve here, or as maximization of the Kendall τ related to the pair $(s(X), Y)$. From the theoretical side, we describe the optimal elements of this problem and provide statistical guarantees for empirical Kendall τ maximization under appropriate conditions for the class of scoring function candidates. We also propose a recursive statistical learning algorithm tailored to empirical IROC curve optimization and producing a piecewise constant scoring function that is fully described by an oriented binary tree. Preliminary numerical experiments highlight the difference in nature between *regression* and *continuous ranking* and provide strong empirical evidence of the performance of empirical optimizers of the criteria proposed.

1 Introduction

The predictive learning problem considered in this paper can be easily stated in an informal fashion, as follows. Given a collection of objects of arbitrary cardinality, $N \geq 1$ say, respectively described by characteristics $x_1, \ \ldots, \ x_N$ in a feature space $\mathcal{X}$, the goal is to learn how to order them by increasing order of magnitude of a certain unknown continuous variable y. To fix ideas, the attribute y can represent the 'size' of the object and be difficult to measure, as for the physical measurement of microscopic bodies in chemistry and biology or the cash flow of companies in quantitative finance and the features x may then correspond to *indirect measurements*. The most convenient way to define a preorder on a feature space $\mathcal{X}$ is to transport the natural order on the real line onto it by means of a (measurable) scoring function $s : \mathcal{X} \to \mathbb{R}$: an object with charcateristics x is then said to be 'larger' ('strictly larger', respectively) than an object described by x' according to the scoring rule s when $s(x') \leq s(x)$ (when $s(x) < s(x')$). Statistical learning boils down here to build a scoring function $s(x)$, based on a *training* data set $\mathcal{D}_n = \{(X_1, Y_1), \ \ldots, \ (X_n, Y_n)\}$ of objects for which the values of all variables (direct and indirect measurements) have been jointly observed, such that $s(X)$ and Y tend to increase or decrease together with highest probability or, in other words, such that the ordering of new objects induced by $s(x)$ matches that defined by their true measures as well as possible. This problem, that shall be referred to as *continuous ranking* throughout the article can be viewed as an extension of *bipartite ranking*, where the output variable Y is assumed to be binary and the objective can be naturally formulated as a functional M-estimation problem by means of the concept of ROC curve, see [7]. Refer also to [4], [11], [1] for approaches based on the optimization

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

of summary performance measures such as the AUC criterion in the binary context. Generalization to the situation where the random label is ordinal and may take a finite number $K \geq 3$ of values is referred to as *multipartite ranking* and has been recently investigated in [16] (see also *e.g.* [14]), where distributional conditions guaranteeing that ROC surface and the VUS criterion can be used to determine optimal scoring functions are exhibited in particular.

It is the major purpose of this paper to formulate the *continuous ranking* problem in a quantitative manner and explore the connection between the latter and bi/multi-partite ranking. Intuitively, optimal scoring rules would be also optimal for any bipartite subproblem defined by thresholding the continuous variable Y with cut-off $t > 0$, separating the observations X such that $Y < t$ from those such that $Y > t$. Viewing this way *continuous ranking* as a continuum of nested bipartite ranking problems, we provide here sufficient conditions for the existence of such (optimal) scoring rules and we introduce a concept of *integrated* ROC *curve* (IROC curve in abbreviated form) that may serve as a natural performance measure for continuous ranking, as well as the related notion of *integrated* AUC *criterion*, a summary scalar criterion, akin to Kendall tau. Generalization properties of empirical Kendall tau maximizers are discussed in the Supplementary Material. The paper also introduces a novel recursive algorithm that solves a discretized version of the empirical *integrated* ROC *curve* optimization problem, producing a scoring function that can be computed by means of a hierarchical combination of binary classification rules. Numerical experiments providing strong empirical evidence of the relevance of the approach promoted in this paper are also presented.

The paper is structured as follows. The probabilistic framework we consider is described and key concepts of bi/multi-partite ranking are briefly recalled in section 2. Conditions under which optimal solutions of the problem of ranking data with continuous labels exist are next investigated in section 3, while section 4 introduces a dedicated quantitative (functional) performance measure, the IROC curve. The algorithmic approach we propose in order to learn scoring functions with nearly optimal IROC curves is presented at length in section 5. Numerical results are displayed in section 6. Some technical proofs are deferred to the Supplementary Material.

2 Notation and Preliminaries

Throughout the paper, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$. The pseudo-inverse of any cdf $F(t)$ on $\mathbb{R}$ is denoted by $F^{-1}(u) = \inf\{s \in \mathbb{R} : \ F(s) \geq u\}$, while $\mathcal{U}([0,1])$ denotes the uniform distribution on the unit interval $[0,1]$.

2.1 The probabilistic framework

Given a continuous real valued r.v. Y representing an attribute of an object, its 'size' say, and a random vector X taking its values in a (typically high dimensional euclidian) feature space $\mathcal{X}$ modelling other observable characteristics of the object (*e.g.* 'indirect measurements' of the size of the object), hopefully useful for predicting Y, the statistical learning problem considered here is to learn from $n \geq 1$ training independent observations $\mathcal{D}_n = \{(X_1, Y_1), \ \ldots, \ (X_n, Y_n)\}$, drawn as the pair (X, Y), a measurable mapping $s : \mathcal{X} \to \mathbb{R}$, that shall be referred to as a *scoring function* throughout the paper, so that the variables $s(X)$ and Y tend to increase or decrease together: ideally, the larger the score $s(X)$, the higher the size Y. For simplicity, we assume throughout the article that $\mathcal{X} = \mathbb{R}^d$ with $d \geq 1$ and that the support of Y's distribution is compact, equal to $[0,1]$ say. For any $q \geq 1$, we denote by λ_q the Lebesgue measure on $\mathbb{R}^q$ equipped with its Borelian σ-algebra and suppose that the joint distribution $F_{X,Y}(dxdy)$ of the pair (X, Y) has a density $f_{X,Y}(x, y)$ w.r.t. the tensor product measure $\lambda_d \otimes \lambda_1$. We also introduces the marginal distributions $F_Y(dy) = f_Y(y)\lambda_1(dy)$ and $F_X(dx) = f_X(x)\lambda_d(dx)$, where $f_Y(y) = \int_{x \in \mathcal{X}} f_{X,Y}(x, y)\lambda_d(dx)$ and $f_X(x) = \int_{y \in [0,1]} f_{X,Y}(x, y)\lambda_1(dy)$ as well as the conditional densities $f_{X|Y=y}(x) = f_{X,Y}(x, y)/f_Y(y)$ and $f_{Y|X=x}(y) = f_{X,Y}(x, y)/f_X(x)$. Observe incidentally that the probabilistic framework of the continuous ranking problem is quite similar to that of *distribution-free regression*. However, as shall be seen in the subsequent analysis, even if the regression function $m(x) = \mathbb{E}[Y \mid X = x]$ can be optimal under appropriate conditions, just like for regression, measuring ranking performance involves criteria that are of different nature than the expected least square error and plug-in rules may not be relevant for the goal pursued here, as depicted by Fig. 2 in the Supplementary Material.

Scoring functions. The set of all scoring functions is denoted by $\mathcal{S}$ here. Any scoring function $s \in \mathcal{S}$ defines a total preorder on the space $\mathcal{X}$: $\forall (x, x') \in \mathcal{X}^2$, $x \preceq_s x' \Leftrightarrow s(x) \leq s(x')$. We also set $x \prec_s x'$ when $s(x) < s(x')$ and $x =_s x'$ when $s(x) = s(x')$ for $(x, x') \in \mathcal{X}^2$.

2.2 Bi/multi-partite ranking

Suppose that Z is a binary label, taking its values in $\{-1, +1\}$ say, assigned to the r.v. X. In bipartite ranking, the goal is to pick s in $\mathcal{S}$ so that the larger $s(X)$, the greater the probability that Y is equal to 1 ideally. In other words, the objective is to learn $s(x)$ such that the r.v. $s(X)$ given $Y = +1$ is as *stochastically larger*[1] as possible than the r.v. $s(X)$ given $Y = -1$: the difference between $\bar{G}_s(t) = \mathbb{P}\{s(X) \geq t \mid Y = +1\}$ and $\bar{H}_s(t) = \mathbb{P}\{s(X) \geq t \mid Y = -1\}$ should be thus maximal for all $t \in \mathbb{R}$. This can be naturally quantified by means of the notion of ROC curve of a candidate $s \in \mathcal{S}$, *i.e.* the parametrized curve $t \in \mathbb{R} \mapsto (\bar{H}_s(t), \bar{G}_s(t))$, which can be viewed as the graph of a mapping $\mathrm{ROC_s} : \alpha \in (0, 1) \mapsto \mathrm{ROC_s}(\alpha)$, connecting possible discontinuity points by linear segments (so that $\mathrm{ROC_s}(\alpha) = \bar{\mathrm{G}}_\mathrm{s} \circ (1 - \mathrm{H_s^{-1}})(1 - \alpha)$ when H_s has no flat part in $H_s^{-1}(1 - \alpha)$, where $H_s = 1 - \bar{H}_s$). A basic Neyman Pearson's theory argument shows that the optimal elements $s^*(x)$ related to this natural (functional) bipartite ranking criterion (*i.e.* scoring functions whose ROC curve dominates any other ROC curve everywhere on $(0, 1)$) are transforms $(T \circ \eta)(x)$ of the posterior probability $\eta(x) = \mathbb{P}\{Z = +1 \mid X = x\}$, where $T : \mathrm{SUPP}(\eta(X)) \to \mathbb{R}$ is any strictly increasing borelian mapping. Optimization of the curve in $\sup$ norm has been considered in [7] or in [8] for instance. However, given its functional nature, in practice the ROC curve of any $s \in \mathcal{S}$ is often summarized by the area under it, which performance measure can be interpreted in a probabilistic manner, as the theoretical rate of *concording pairs*

$$\mathrm{AUC(s)} = \mathbb{P}\left\{\mathrm{s(X)} < \mathrm{s(X')} \mid \mathrm{Z} = -1,\ \mathrm{Z'} = +1\right\} + \frac{1}{2}\mathbb{P}\left\{\mathrm{s(X)} = \mathrm{s(X')} \mid \mathrm{Z} = -1,\ \mathrm{Z'} = +1\right\}, \tag{1}$$

where (X', Z') denoted an independent copy of (X, Z). A variety of algorithms aiming at maximizing the AUC criterion or surrogate pairwise criteria have been proposed and studied in the literature, among which [11], [15] or [3], whereas generalization properties of empirical AUC maximizers have been studied in [5], [1] and [12]. An analysis of the relationship between the AUC and the error rate is given in [9].

Extension to the situation where the label Y takes at least three ordinal values (*i.e.* multipartite ranking) has been also investigated, see *e.g.* [14] or [6]. In [16], it is shown that, in contrast to the bipartite setup, the existence of optimal solutions cannot be guaranteed in general and conditions on (X, Y)'s distribution ensuring that optimal solutions do exist and that extensions of bipartite ranking criteria such as the ROC manifold and the volume under it can be used for learning optimal scoring rules have been exhibited. An analogous analysis in the context of continuous ranking is carried out in the next section.

3 Optimal elements in ranking data with continuous labels

In this section, a natural definition of the set of optimal elements for continuous ranking is first proposed. Existence and characterization of such optimal scoring functions are next discussed.

3.1 Optimal scoring rules for continuous ranking

Considering a threshold value $y \in [0, 1]$, a considerably weakened (and discretized) version of the problem stated informally above would consist in finding s so that the r.v. $s(X)$ given $Y > y$ is as stochastically larger than $s(X)$ given $Y < y$ as possible. This *subproblem* coincides with the *bipartite ranking* problem related to the pair (X, Z_y), where $Z_y = 2\mathbb{I}\{Y > y\} - 1$. As briefly recalled in subsection 2.2, the optimal set $\mathcal{S}_y^*$ is composed of the scoring functions that induce the same ordering as

$$\eta_y(X) = \mathbb{P}\{Y > y \mid X\} = 1 - (1 - p_y)/(1 - p_y + p_y \Phi_y(X)),$$

where $p_y = 1 - F_Y(y) = \mathbb{P}\{Y > y\}$ and $\Phi_y(X) = (dF_{X|Y>y}/dF_{X|Y<y})(X)$.

[1] Given two real-valued r.v.'s U and U', recall that U is said to be *stochastically larger* than U' when $\mathbb{P}\{U \geq t\} \geq \mathbb{P}\{U' \geq t\}$ for all $t \in \mathbb{R}$.

A continuum of bipartite ranking problems. The rationale behind the definition of the set $\mathcal{S}^*$ of optimal scoring rules for continuous ranking is that any element s^* should score observations x in the same order as η_y (or equivalently as Φ_y).

Definition 1. (OPTIMAL SCORING RULE) *An optimal scoring rule for the continuous ranking problem related to the random pair (X, Y) is any element s^* that fulfills: $\forall y \in (0,1)$,*

$$\forall (x, x') \in \mathcal{X}^2, \quad \eta_y(x) < \eta_y(x') \Rightarrow s^*(x) < s^*(x'). \tag{2}$$

In other words, the set of optimal rules is defined as $\mathcal{S}^* = \bigcap_{y \in (0,1)} \mathcal{S}_y^*$.

It is noteworthy that, although the definition above is natural, the set $\mathcal{S}^*$ can be empty in absence of any distributional assumption, as shown by the following example.

Example 1. *As a counter-example, consider the distributions $F_{X,Y}$ such that $F_Y = \mathcal{U}([0,1])$ and $F_{X|Y=y} = \mathcal{N}(|2y-1|, (2y-1)^2)$. Observe that $(X, 1-Y) \stackrel{d}{=} (X, Y)$, so that $\Phi_{1-t} = \Phi_t^{-1}$ for all $t \in (0,1)$ and there exists $t \neq 0$ s.t. Φ_t is not constant. Hence, there exists no s^* in $\mathcal{S}$ such that* (2) *holds true for all $t \in (0,1)$.*

Remark 1. (INVARIANCE) *We point out that the class $\mathcal{S}^*$ of optimal elements for continuous ranking thus defined is invariant by strictly increasing transform of the 'size' variable Y (in particular, a change of unit has no impact on the definition of $\mathcal{S}^*$): for any borelian and strictly increasing mapping $H : (0,1) \to (0,1)$, any scoring function $s^*(x)$ that is optimal for the continuous ranking problem related to the pair (X, Y) is still optimal for that related to $(X, H(Y))$ (since, under these hypotheses, for any $y \in (0,1)$: $Y > y \Leftrightarrow H(Y) > H(y)$).*

3.2 Existence and characterization of optimal scoring rules

We now investigate conditions guaranteeing the existence of optimal scoring functions for the continuous ranking problem.

Proposition 1. *The following assertions are equivalent.*

1. *For all $0 < y < y' < 1$, for all $(x, x') \in \mathcal{X}^2$: $\Phi_y(x) < \Phi_y(x') \Rightarrow \Phi_{y'}(x) \leq \Phi_{y'}(x')$.*
2. *There exists an optimal scoring rule s^* (i.e. $\mathcal{S}^* \neq \emptyset$).*
3. *The regression function $m(x) = \mathbb{E}[Y \mid X = x]$ is an optimal scoring rule.*
4. *The collection of probability distributions $F_{X|Y=y}(dx) = f_{X|Y=y}(x)\lambda_d(dx)$, $y \in (0,1)$ satisfies the monotone likelihood ratio property: there exist $s^* \in \mathcal{S}$ and, for all $0 < y < y' < 1$, an increasing function $\varphi_{y,y'} : \mathbb{R} \to \mathbb{R}_+$ such that: $\forall x \in \mathbb{R}^d$,*

$$\frac{f_{X|Y=y'}}{f_{X|Y=y}}(x) = \varphi_{y,y'}(s^*(x)).$$

Refer to the Appendix section for the technical proof. Truth should be said, assessing that Assertion 1. is a very challenging statistical task. However, through important examples, we now describe (not uncommon) situations where the conditions stated in Proposition 1 are fulfilled.

Example 2. *We give a few important examples of probabilistic models fulfilling the properties listed in Proposition 1.*

• **Regression model.** *Suppose that $Y = m(X) + \epsilon$, where $m : \mathcal{X} \to \mathbb{R}$ is a borelian function and ϵ is a centered r.v. independent from X. One may easily check that $m \in \mathcal{S}^*$.*

• **Exponential families.** *Suppose that $f_{X|Y=y}(x) = \exp(\kappa(y)T(x) - \psi(y))f(x)$ for all $x \in \mathbb{R}^d$, where $f : \mathbb{R}^d \to \mathbb{R}_+$ is borelian, $\kappa : [0,1] \to \mathbb{R}$ is a borelian strictly increasing function and $T : \mathbb{R}^d \to \mathbb{R}$ is a borelian mapping such that $\psi(y) = \log \int_{x \in \mathbb{R}^d} \exp(\kappa(y)T(x))f(x)dx < +\infty$.*

We point out that, although the regression function $m(x)$ is an optimal scoring function when $\mathcal{S}^* \neq \emptyset$, the *continuous ranking* problem does not coincide with *distribution-free regression* (notice incidentally that, in this case, any strictly increasing transform of $m(x)$ belongs to $\mathcal{S}^*$ as well). As depicted by Fig. 2 the least-squares criterion is not relevant to evaluate continuous ranking performance and naive plug-in strategies should be avoided, see Remark 3 below. Dedicated performance criteria are proposed in the next section.

4 Performance measures for continuous ranking

We now investigate quantitative criteria for assessing the performance in the continuous ranking problem, which practical machine-learning algorithms may rely on. We place ourselves in the situation where the set $\mathcal{S}^*$ is not empty, see Proposition 1 above.

A functional performance measure. It follows from the view developped in the previous section that, for any $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$ and for all $y \in (0,1)$, we have:

$$\forall \alpha \in (0,1), \;\; \mathrm{ROC}_{\mathrm{s,y}}(\alpha) \leq \mathrm{ROC}_{\mathrm{s}^*,\mathrm{y}}(\alpha) = \mathrm{ROC}^*_{\mathrm{y}}(\alpha), \tag{3}$$

denoting by $\mathrm{ROC}_{\mathrm{s,y}}$ the ROC curve of any $s \in \mathcal{S}$ related to the bipartite ranking subproblem (X, Z_y) and by $\mathrm{ROC}^*_{\mathrm{y}}$ the corresponding optimal ROC curve, *i.e.* the ROC curve of strictly increasing transforms of $\eta_y(x)$. Based on this observation, it is natural to design a dedicated performance measure by aggregating these 'sub-criteria'. Integrating over y w.r.t. a σ-finite measure μ with support equal to $[0,1]$, this leads to the following definition $\mathrm{IROC}_{\mu,\mathrm{s}}(\alpha) = \int \mathrm{ROC}_{\mathrm{s,y}}(\alpha)\mu(\mathrm{dy})$. The functional criterion thus defined inherits properties from the $\mathrm{ROC}_{\mathrm{s,y}}$'s (*e.g.* monotonicity, concavity). In addition, the curve $\mathrm{IROC}_{\mu,\mathrm{s}^*}$ with $s^* \in \mathcal{S}^*$ dominates everywhere on $(0,1)$ any other curve $\mathrm{IROC}_{\mu,\mathrm{s}}$ for $s \in \mathcal{S}$. However, except in pathologic situations (*e.g.* when $s(x)$ is constant), the curve $\mathrm{IROC}_{\mu,\mathrm{s}}$ is not invariant when replacing Y's distribution by that of a strictly increasing transform $H(Y)$. In order to guarantee that this desirable property is fulfilled (see Remark 1), one should integrate w.r.t. Y's distribution (which boils down to replacing Y by the uniformly distributed r.v. $F_Y(Y)$).

Definition 2. (INTEGRATED $\mathrm{ROC}/\mathrm{AUC}$ CRITERIA) *The integrated* ROC *curve of any scoring rule $s \in \mathcal{S}$ is defined as: $\forall \alpha \in (0,1)$,*

$$\mathrm{IROC}_{\mathrm{s}}(\alpha) = \int_{\mathrm{y}=0}^{1} \mathrm{ROC}_{\mathrm{s,y}}(\alpha)\mathrm{F}_{\mathrm{Y}}(\mathrm{dy}) = \mathbb{E}\left[\mathrm{ROC}_{\mathrm{s,Y}}(\alpha)\right]. \tag{4}$$

The integrated AUC *criterion is defined as the area under the integrated* ROC *curve: $\forall s \in \mathcal{S}$,*

$$\mathrm{IAUC(s)} = \int_{\alpha=0}^{1} \mathrm{IROC}_{\mathrm{s}}(\alpha)\mathrm{d}\alpha. \tag{5}$$

The following result reveals the relevance of the functional/summary criteria defined above for the continuous ranking problem. Additional properties of IROC curves are listed in the Supplementary Material.

Theorem 1. *Let $s^* \in \mathcal{S}$. The following assertions are equivalent.*

1. *The assertions of Proposition 1 are fulfilled and s^* is an optimal scoring function in the sense given by Definition 1.*
2. *For all $\alpha \in (0,1)$, $\mathrm{IROC}_{\mathrm{s}^*}(\alpha) = \mathbb{E}\left[\mathrm{ROC}^*_{\mathrm{Y}}(\alpha)\right]$.*
3. *We have $\mathrm{IAUC}_{\mathrm{s}^*} = \mathbb{E}\left[\mathrm{AUC}^*_{\mathrm{Y}}\right]$, where $\mathrm{AUC}^*_{\mathrm{y}} = \int_{\alpha=0}^{1} \mathrm{ROC}^*_{\mathrm{y}}(\alpha)\mathrm{d}\alpha$ for all $y \in (0,1)$.*

If $\mathcal{S}^* \neq \emptyset$, then we have: $\forall s \in \mathcal{S}$,

$$\begin{aligned} \mathrm{IROC}_{\mathrm{s}}(\alpha) &\leq \mathrm{IROC}^*(\alpha) \stackrel{\mathrm{def}}{=} \mathbb{E}\left[\mathrm{ROC}^*_{\mathrm{Y}}(\alpha)\right], \quad \textit{for any } \alpha \in (0,1,) \\ \mathrm{IAUC(s)} &\leq \mathrm{IAUC}^* \stackrel{\mathrm{def}}{=} \mathbb{E}\left[\mathrm{AUC}^*_{\mathrm{Y}}\right]. \end{aligned}$$

In addition, for any borelian and strictly increasing mapping $H : (0,1) \to (0,1)$, replacing Y by $H(Y)$ leaves the curves $\mathrm{IROC}_{\mathrm{s}}$, $s \in \mathcal{S}$, unchanged.

Equipped with the notion defined above, a scoring rule s_1 is said to be more accurate than another one s_2 if $\mathrm{IROC}_{\mathrm{s}_2}(\alpha) \leq \mathrm{IROC}_{\mathrm{s}_1}(\alpha)$ for all $\alpha \in (0,1)$.The IROC curve criterion thus provides a partial preorder on $\mathcal{S}$. Observe also that, by virtue of Fubini's theorem, we have $\mathrm{IAUC(s)} = \int \mathrm{AUC}_{\mathrm{y}}(\mathrm{s})\mathrm{F}_{\mathrm{Y}}(\mathrm{dy})$ for all $s \in \mathcal{S}$, denoting by $\mathrm{AUC}_{\mathrm{y}}(\mathrm{s})$ the AUC of s related to the bipartite ranking subproblem (X, Z_y). Just like the AUC for bipartite ranking, the scalar IAUC criterion defines a full preorder on $\mathcal{S}$ for continuous ranking. Based on a training dataset $\mathcal{D}_n$ of independent copies of (X, Y), statistical versions of the $\mathrm{IROC}/\mathrm{IAUC}$ criteria can be straightforwardly computed by replacing the distributions F_Y, $F_{X|Y>t}$ and $F_{X|Y<t}$ by their empirical counterparts in (3)-(5), see the Supplementary Material for further details. The lemma below provides a probabilistic interpretation of the IAUC criterion.

Lemma 1. *Let (X', Y') be a copy of the random pair (X, Y) and Y'' a copy of the r.v. Y. Suppose that (X, Y), (X', Y') and Y'' are defined on the same probability space and are independent. For all $s \in \mathcal{S}$, we have:*

$$\mathrm{IAUC(s)} = \mathbb{P}\left\{\mathrm{s(X)} < \mathrm{s(X')} \mid \mathrm{Y} < \mathrm{Y''} < \mathrm{Y'}\right\} + \frac{1}{2}\mathbb{P}\left\{\mathrm{s(X)} = \mathrm{s(X')} \mid \mathrm{Y} < \mathrm{Y''} < \mathrm{Y'}\right\}. \quad (6)$$

This result shows in particular that a natural statistical estimate of $\mathrm{IAUC(s)}$ based on $\mathcal{D}_n$ involves U-statistics of degree 3. Its proof is given in the Supplementary Material for completeness.

The Kendall τ statistic. The quantity (6) is akin to another popular way to measure the tendency to define the same ordering on the statistical population in a summary fashion:

$$\begin{aligned} d_\tau(s) &\overset{def}{=} \mathbb{P}\left\{(s(X) - s(X')) \cdot (Y - Y') > 0\right\} + \frac{1}{2}\mathbb{P}\left\{s(X) = s(X')\right\} \quad (7) \\ &= \mathbb{P}\{s(X) < s(X') \mid Y < Y'\} + \frac{1}{2}\mathbb{P}\left\{X =_s X'\right\}, \end{aligned}$$

where (X', Y') denotes an independent copy of (X, Y), observing that $\mathbb{P}\{Y < Y'\} = 1/2$. The empirical counterpart of (7) based on the sample $\mathcal{D}_n$, given by

$$\widehat{d}_n(s) = \frac{2}{n(n-1)} \sum_{i<j} \mathbb{I}\left\{(s(X_i) - s(X_j)) \cdot (Y_i - Y_j) > 0\right\} + \frac{1}{n(n-1)} \sum_{i<j} \mathbb{I}\left\{s(X_i) = s(X_j)\right\} \quad (8)$$

is known as the *Kendall τ statistic* and is widely used in the context of statistical hypothesis testing. The quantity (7) shall be thus referred to as the (theoretical or true) *Kendall τ*. Notice that $d_\tau(s)$ is invariant by strictly increasing transformation of $s(x)$ and thus describes properties of the order it defines. The following result reveals that the class $\mathcal{S}^*$, when non empty, is the set of maximizers of the theoretical Kendall τ. Refer to the Supplementary Material for the technical proof.

Proposition 2. *Suppose that $\mathcal{S}^* \neq \emptyset$. For any $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$, we have: $d_\tau(s) \leq d_\tau(s^*)$.*

Equipped with these criteria, the objective expressed above in an informal manner can be now formulated in a quantitative manner as a (possibly functional) M-estimation problem. In practice, the goal pursued is to find a reasonable approximation of a solution to the optimization problem $\max_{s \in \mathcal{S}} d_\tau(s)$ (respectively $\max_{s \in \mathcal{S}} \mathrm{IAUC(s)}$), where the supremum is taken over the set of all scoring functions $s : \mathcal{X} \to \mathbb{R}$. Of course, these criteria are unknown in general, just like (X, Y)'s probability distribution, and the empirical risk minimization (ERM in abbreviated form) paradigm (see [10]) invites for maximizing the statistical version (8) over a class $\mathcal{S}_0 \subset \mathcal{S}$ of controlled complexity when considering the criterion $d_\tau(s)$ for instance. The generalization capacity of empirical maximizers of the Kendall τ can be straightforwardly established using results in [5]. More details are given in the Supplementary Material.

Before describing a practical algorithm for recursive maximization of the IROC curve, a few remarks are in order.

Remark 2. (ON KENDALL τ AND AUC) *We point out that, in the bipartite ranking problem (i.e. when the output variable Z takes its values in $\{-1, +1\}$, see subsection 2.2) as well, the* AUC *criterion can be expressed as a function of the Kendall τ related to the pair $(s(X), Z)$ when the r.v. $s(X)$ is continuous. Indeed, we have in this case $2p(1-p)\mathrm{AUC(s)} = \mathrm{d_\tau(s)}$, where $p = \mathbb{P}\{Z = +1\}$ and $d_\tau(s) = \mathbb{P}\{(s(X) - s(X')) \cdot (Z - Z') > 0\}$, denoting by (X', Z') an independent copy of (X, Z).*

Remark 3. (CONNECTION TO DISTRIBUTION-FREE REGRESSION) *Consider the nonparametric regression model $Y = m(X) + \epsilon$, where ϵ is a centered r.v. independent from X. In this case, it is well-known that the regression function $m(X) = \mathbb{E}[Y \mid X]$ is the (unique) solution of the expected least squares minimization. However, although $m \in \mathcal{S}^*$, the least squares criterion is far from appropriate to evaluate ranking performance, as depicted by Fig. 2. Observe additionally that, in contrast to the criteria introduced above, increasing transformation of the output variable Y may have a strong impact on the least squares minimizer: except for linear stransforms, $\mathbb{E}[H(Y) \mid X]$ is not an increasing transform of $m(X)$.*

Remark 4. (ON DISCRETIZATION) *Bi/multi-partite algorithms are not directly applicable to the continuous ranking problem. Indeed a discretization of the interval [0, 1] would be first required but this would raise a difficult question outside our scope: how to choose this discretization based on the training data? We believe that this approach is less efficient than ours which reveals problem-specific criteria, namely* IROC *and* IAUC.

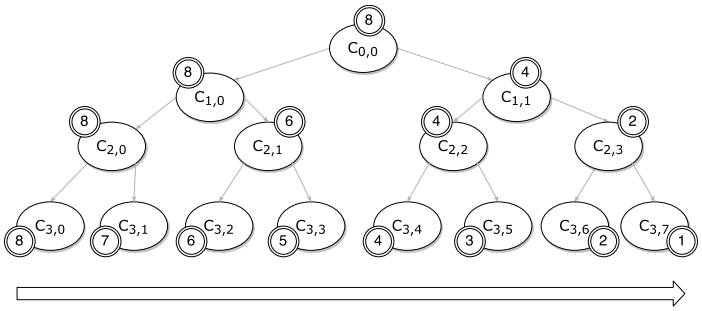

Figure 1: A scoring function described by an oriented binary subtree $\mathcal{T}$. For any element $x \in \mathcal{X}$, one may compute the quantity $s_{\mathcal{T}}(x)$ very fast in a top-down fashion by means of the heap structure: starting from the initial value 2^J at the root node, at each internal node $\mathcal{C}_{j,k}$, the score remains unchanged if x moves down to the left sibling, whereas one subtracts $2^{J-(j+1)}$ from it if x moves down to the right.

5 Continuous Ranking through Oriented Recursive Partitioning

It is the purpose of this section to introduce the algorithm CRANK, a specific tree-structured learning algorithm for continuous ranking.

5.1 Ranking trees and Oriented Recursive Partitions

Decision trees undeniably figure among the most popular techniques, in supervised and unsupervised settings, refer to [2] or [13] for instance. This is essentially due to the visual model summary they provide, in the form of a binary tree graphic that permits to describe predictions by means of a hierachichal combination of elementary rules of the type "$X^{(j)} \leq \kappa$" or "$X^{(j)} > \kappa$", comparing the value taken by a (quantitative) component of the input vector X (the *split variable*) to a certain threshold (the *split value*). In contrast to local learning problems such as classification or regression, predictive rules for a global problem such as *ranking* cannot be described by a (tree-structured) partition of the feature space: cells (corresponding to the terminal leaves of the binary decision tree) must be ordered so as to define a scoring function. This leads to the definition of *ranking trees* as binary trees equipped with a "left-to-right" orientation, defining a tree-structured collection of anomaly scoring functions, as depicted by Fig. 1. Binary ranking trees have been in the context of bipartite ranking in [7] or in [3] and in [16] in the context of multipartite ranking. The root node of a ranking tree $\mathcal{T}_J$ of depth $J \geq 0$ represents the whole feature space $\mathcal{X}$: $\mathcal{C}_{0,0} = \mathcal{X}$, while each internal node (j,k) with $j < J$ and $k \in \{0, \ldots, 2^j - 1\}$ corresponds to a subset $\mathcal{C}_{j,k} \subset \mathcal{X}$, whose left and right siblings respectively correspond to disjoint subsets $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$ such that $\mathcal{C}_{j,k} = \mathcal{C}_{j+1,2k} \cup \mathcal{C}_{j+1,2k+1}$. Equipped with the left-to-right orientation, any subtree $\mathcal{T} \subset \mathcal{T}_J$ defines a preorder on $\mathcal{X}$: elements lying in the same terminal cell of $\mathcal{T}$ being equally ranked. The scoring function related to the oriented tree $\mathcal{T}$ can be written as:

$$s_{\mathcal{T}}(x) = \sum_{\mathcal{C}_{j,k} \text{: terminal leaf of } \mathcal{T}} 2^J \left(1 - \frac{k}{2^j}\right) \cdot \mathbb{I}\{x \in \mathcal{C}_{j,k}\}. \tag{9}$$

5.2 The CRANK algorithm

Based on Proposition 2, as mentioned in the Supplementary Material, one can try to build from the training dataset $\mathcal{D}_n$ a ranking tree by recursive empirical Kendall τ maximization. We propose below an alternative tree-structured recursive algorithm, relying on a (dyadic) discretization of the 'size' variable Y. At each iteration, the local sample (*i.e.* the data lying in the cell described by the current node) is split into two halves (the highest/smallest halves, depending on Y) and the algorithm calls a binary classification algorithm $\mathcal{A}$ to learn how to divide the node into right/left children. The theoretical analysis of this algorithm and its connection with approximation of IROC^* are difficult questions that will be adressed in future work. Indeed we found out that the IROC cannot be

represented as a parametric curve contrary to the ROC, which renders proofs much more difficult than in the bipartite case.

THE CRANK ALGORITHM

1. **Input.** Training data $\mathcal{D}_n$, depth $J \geq 1$, binary classification algorithm $\mathcal{A}$.
2. **Initialization.** Set $\mathcal{C}_{0,0} = \mathcal{X}$.
3. **Iterations.** For $j = 0, \ldots, J-1$ and $k = 0, \ldots, 2^J - 1$,
   (a) Compute a median $y_{j,k}$ of the dataset $\{Y_1, \ldots, , Y_n\} \cap \mathcal{C}_{j,k}$ and assign the binary label $Z_i = 2\mathbb{I}\{Y_i > y_{j,k}\} - 1$ to any data point i lying in $\mathcal{C}_{j,k}$, *i.e.* such that $X_i \in \mathcal{C}_{j,k}$.
   (b) Solve the binary classification problem related to the input space $\mathcal{C}_{j,k}$ and the training set $\{(X_i, Y_i) : 1 \leq i \leq n, X_i \in \mathcal{C}_{j,k}\}$, producing a classifier $g_{j,k} : \mathcal{C}_{j,k} \to \{-1, +1\}$.
   (c) Set $\mathcal{C}_{j+1,2k} = \{x \in \mathcal{C}_{j,k}, g_{j,k} = +1\} = \mathcal{C}_{j,k} \setminus \mathcal{C}_{j+1,2k+1}$.
4. **Output.** Ranking tree $\mathcal{T}_J = \{\mathcal{C}_{j,k} : 0 \leq j \leq J, 0 \leq k < D\}$.

Of course, the depth J should be chosen such that $2^J \leq n$. One may also consider continuing to split the nodes until the number of data points within a cell has reached a minimum specified in advance. In addition, it is well known that recursive partitioning methods fragment the data and the unstability of splits increases with the depth. For this reason, a ranking subtree must be selected. The growing procedure above should be classically followed by a pruning stage, where children of a same parent are progressively merged until the root $\mathcal{T}_0$ is reached and a subtree among the sequence $\mathcal{T}_0 \subset \ldots \subset \mathcal{T}_J$ with nearly maximal IAUC should be chosen using cross-validation. Issues related to the implementation of the CRANK algorithm and variants (*e.g.* exploiting randomization/aggregation) will be investigated in a forthcoming paper.

6 Numerical Experiments

In order to illustrate the idea conveyed by Fig. 2 that the least squares criterion is not appropriate for the continuous ranking problem we compared on a toy example CRANK with CART. Recall that the latter is a regression decision tree algorithm which minimizes the MSE (Mean Squared Error). We also runned an alternative version of CRANK which maximizes the empirical Kendall τ instead of the empirical IAUC: this method is refered to as KENDALL from now on. The experimental setting is composed of a unidimensional feature space $\mathcal{X} = [0, 1]$ (for visualization reasons) and a simple regression model without any noise: $Y = m(X)$. Intuitively, a least squares strategy can miss slight oscillations of the regression function, which are critical in ranking when they occur in high probability regions as they affect the order among the feature space. The results are presented in Table 1. See Supplementary Material for further details.

	IAUC	Kendall τ	MSE
CRANK	0.95	0.92	0.10
KENDALL	0.94	0.93	0.10
CART	0.61	0.58	7.4×10^{-4}

Table 1: IAUC, Kendall τ and MSE empirical measures

7 Conclusion

This paper considers the problem of learning how to order objects by increasing 'size', modeled as a continuous r.v. Y, based on *indirect measurements* X. We provided a rigorous mathematical formulation of this problem that finds many applications (*e.g.* quality control, chemistry) and is referred to as *continuous ranking*. In particular, necessary and sufficient conditions on (X, Y)'s distribution for the existence of optimal solutions are exhibited and appropriate criteria have been proposed for evaluating the performance of scoring rules in these situations. In contrast to distribution-free regression where the goal is to recover the local values taken by the regression function, *continuous*

ranking aims at reproducing the preorder it defines on the feature space as accurately as possible. The numerical results obtained via the algorithmic approaches we proposed for optimizing the criteria aforementioned highlight the difference in nature between these two statistical learning tasks.

Acknowledgments

This work was supported by the industrial chair *Machine Learning for Big Data* from Télécom ParisTech and by a public grant (*Investissement d'avenir* project, reference ANR-11-LABX-0056-LMH, LabEx LMH).

References

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *J. Mach. Learn. Res.*, 6:393–425, 2005.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[3] G. Clémençon, M. Depecker, and N. Vayatis. Ranking Forests. *J. Mach. Learn. Res.*, 14:39–73, 2013.

[4] S. Clémençon, G. Lugosi, and N.Vayatis. Ranking and scoring using empirical risk minimization. In *Proceedings of COLT 2005*, volume 3559, pages 1–15. Springer., 2005.

[5] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of u-statistics. *The Annals of Statistics*, 36:844–874, 2008.

[6] S. Clémençon and S. Robbiano. The TreeRank Tournament algorithm for multipartite ranking. *Journal of Nonparametric Statistics*, 25(1):107–126, 2014.

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

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

[9] Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. In *Advances in neural information processing systems*, pages 313–320, 2004.

[10] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[11] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

[12] Aditya Krishna Menon and Robert C Williamson. Bipartite ranking: a risk-theoretic perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016.

[13] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):1–81, 1986.

[14] S. Rajaram and S. Agarwal. Generalization bounds for k-partite ranking. In *NIPS 2005 Workshop on Learn to rank*, 2005.

[15] A. Rakotomamonjy. Optimizing Area Under Roc Curve with SVMs. In *Proceedings of the First Workshop on ROC Analysis in AI*, 2004.

[16] S. Robbiano S. Clémençon and N. Vayatis. Ranking data with ordinal labels: optimality and pairwise aggregation. *Machine Learning*, 91(1):67–104, 2013.