# Learning Scoring Functions with Order-Preserving Losses and Standardized Supervision

**David Buffoni** DAVID.BUFFONI@LIP6.FR
**Clément Calauzènes** CLEMENT.CALAUZENES@LIP6.FR
**Patrick Gallinari** PATRICK.GALLINARI@LIP6.FR
**Nicolas Usunier** NICOLAS.USUNIER@LIP6.FR
Laboratoire d'Informatique de Paris 6, 4 place Jussieu 75005 Paris, France

## Abstract

We address the problem of designing surrogate losses for learning scoring functions in the context of label ranking. We extend to ranking problems a notion of order-preserving losses previously introduced for multiclass classification, and show that these losses lead to consistent formulations with respect to a family of ranking evaluation metrics. An order-preserving loss can be tailored for a given evaluation metric by appropriately setting some weights depending on this metric and the observed supervision. These weights, called the standard form of the supervision, do not always exist, but we show that previous consistency results for ranking were proved in special cases where they do. We then evaluate a new pairwise loss consistent with the (Normalized) Discounted Cumulative Gain on benchmark datasets.

## 1. Introduction

Learning to rank has attracted a lot of attention in the past decade. Research on this topic is mainly driven by Information Retrieval (IR) applications like search engines which display their results in the form of ranked lists of items (e.g. documents or images). The improvement of such learning techniques, or a better understanding of them, may affect millions of users daily.

In most of these ranking tasks, the items are ordered in two steps. They are first given scores (w.r.t. the user query) and then sorted by decreasing scores. We consider here the problem of learning the scoring function using a training set of queries for which preferences over the items are given. More specifically, we address the problem of designing *surrogate losses* for the scoring function. A surrogate loss defines the risk a training procedure minimizes in practice, in contrast to the evaluation metric which will eventually measure the quality of the predicted rankings but usually leads to intractable optimization problems. For instance, some popular surrogate losses are based on *pairwise comparisons* (Weston & Watkins, 1999; Herbrich et al., 2000; Freund et al., 2003; Joachims, 2002); they are defined as convex upper bounds on the *pairwise disagreement*, an evaluation metric which counts the number of misordered pairs of items in the predicted ranking.

In this paper, we address the problem of designing surrogate losses that are *consistent* with respect to the target evaluation metric: in the large sample limit, an optimal scoring function for the surrogate risk is also optimal for the risk defined by the target evaluation metric. Consistency is certainly a desirable property of the surrogate formulation, and has recently been studied in the context of ranking by Duchi et al. (2010), where, in particular, the authors showed that existing pairwise approaches were *not* consistent w.r.t. the pairwise disagreement. This result casts in light a lack of understanding of surrogate losses for ranking, even of the most extensively studied ones.

We study a class of surrogate losses for scoring suggested by Zhang (2004) in the context of multiclass classification called *order-preserving*, and define the standardization function associated to an evaluation metric. This function maps the supervision to a vector of scores which satisfies certain properties w.r.t. to noise. We show that given an evaluation metric, there is a consistent surrogate formulation with an order-preserving loss *if and only if* there is a standardization function for this metric. Considering some widely used ranking evaluation metrics, namely the (Normalized)

Appearing in *Proceedings of the 28$^{th}$ International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

Discounted Cumulative Gain ((N)DCG), the Average Precision (AP), and the Expected Reciprocal Rank (ERR), we give the standardization function for the (N)DCG, and show that for the AP and the ERR, the standardization function *does not* exist. Consequently, no surrogate formulation with an order-preserving loss can be consistent with these metrics.

We relate our results to prior analyzes of consistent surrogate losses for scoring of Cossock & Zhang (2008), Xia et al. (2008), and Duchi et al. (2010). While these analysis are slightly different from ours, we show that the positive consistency results that were obtained correspond to special cases where the considered evaluation metric has a standardization function – and thus for which consistent formulations with order-preserving losses can be defined.

As a by-product of our analysis, we propose a new order-preserving pairwise loss for ranking. We experimentally show that the surrogate formulations consistent w.r.t. the DCG and (N)DCG using this new loss are competitive with different variants of existing pairwise losses. Since the new loss only incorporates minor changes to the existing pairwise approaches, we suggest its use as soon as the evaluation metric has a standardization function. This new loss may be of interest to design learning-to-rank algorithms, considering that pairwise methods are known to exhibit a rather good trade-off between performance and sophistication of the algorithm[1] compared to the more recent *listwise* approach (Xia et al., 2008; Yue et al., 2007).

The paper is organized as follows. We first describe the formal framework in Section 2, and present the notions of order-preserving loss and standardization functions in Sections 3 and 4. In Section 5, we analyze various evaluation metrics for ranking and make the link with previous results. The experiments are presented in Section 6, followed by a conclusion in Section 7.

## 2. Framework

**Notations** Uppercase (resp. lowercase) letters denote random variables (resp. constant values or functions). Boldface characters denote vectors or vector-valued functions, but the normal font is used when sub-scripting multidimensional quantities. $\mathfrak{S}_n$ denotes the set of permutations of $\{1,\dots,n\}$. For a vector $\mathbf{x}$ (resp. a permutation $\sigma$) and two indexes $i$ and $j$, $\mathbf{x}^{i\leftrightarrow j}$ (resp. $\sigma^{i\leftrightarrow j}$) is obtained from $\mathbf{x}$ (resp. $\sigma$) by swapping the values at indexes $i$ and $j$ (resp. $\sigma(i)$ and $\sigma(j)$).

**Basic Definitions** We consider a framework of ranking similar to subset or label ranking (Cossock & Zhang, 2008; Dekel et al., 2003). The data is modeled by two jointly distributed random variables $Q$ and $\mathbf{T}$ taking values in $\mathcal{Q}$ (the *query space*) and $\mathcal{Y}$ (the *supervision space*) respectively. For an observed pair $(q,\mathbf{t})$, there is a fixed set $\mathcal{S}(q)$ to order. $\mathcal{S}(q)$ only depends on $q$, and we identify it with $\{1,...,|\mathcal{S}(q)|\}$. We assume that $|\mathcal{S}(q)|=n$ for a constant $n>1$ for simplicity, but all our results would hold if we only assumed $\forall q, |\mathcal{S}(q)| \le n$. $\mathbf{t}$ belongs to a finite set $\mathcal{Y} \subset \mathbb{R}^{|q|}$ and represents the desired ranking of $\mathcal{S}(q)$.[2] Such a supervision is widely used in IR, where $t_i$ is the relevance judgment of item $i$. A *scoring function* $f$ assigns a score $f_i(q) \in \mathbb{R}$ to each item $i \in \mathcal{S}(q)$. The quality of the vector of scores $\mathbf{f}(q)$ with respect to the ground truth $\mathbf{t}$ is measured by a *scoring error*. Following existing evaluation measures for ranking (see Section 5), a scoring error depends only on the ordering induced by the scores, and not the scores themselves. We propose the formal definition below, which considers that ties in scores are broken randomly by the sorting algorithm:

**Definition 1** *Let* $\mathtt{R}^r : \mathfrak{S}_n \times \mathcal{Y} \to \mathbb{R}_+$. *A* scoring error *is a function* $\mathtt{R}^s : \mathbb{R}^n \times \mathcal{Y} \to \mathbb{R}_+$ *defined by:*

$$\forall \mathbf{s} \in \mathbb{R}^n, \forall \mathbf{y} \in \mathcal{Y}, \mathtt{R}^s(\mathbf{s},\mathbf{y}) = \frac{1}{|\mathfrak{S}_{[\mathbf{s}]}|} \sum_{\sigma \in \mathfrak{S}_{[\mathbf{s}]}} \mathtt{R}^r(\sigma,\mathbf{y}) \,, \tag{1}$$

$$\textit{with } \mathfrak{S}_{[\mathbf{s}]} = \left\{\sigma \in \mathfrak{S}_n \middle| \forall i, f_{\sigma^{-1}(i)}(q) \ge f_{\sigma^{-1}(i+1)}(q)\right\}.$$

$\mathtt{R}^r$ *is called the* ordering error *associated to* $\mathtt{R}^s$.

(Note that we used the convention that $\sigma(i)$ is the rank of item $i$, and the top-ranked item is $\sigma^{-1}(1)$.) The goal is to learn a scoring function $f$, using a training set of (*query*, *desired ranking*) pairs, with low *scoring risk*:

$$\mathcal{R}^s(f) = \mathbb{E}\big[\mathtt{R}^s(\mathbf{f}(Q),\mathbf{T})\big]\,. \tag{2}$$

**Surrogate Losses and Consistency** In practice, the optimization of the empirical scoring risk on a training set is intractable because it is not continuous. The usual solution is then to define a (preferably convex) continuous, bounded below surrogate loss $\Psi$ and learn $f$ on the training data to minimize the $\Psi$*-risk*:

$$\mathcal{R}^\Psi(f) = \mathbb{E}\big[\Psi(\mathbf{f}(Q),\mathbf{T})\big]\,. \tag{3}$$

The goal of the paper is to give general conditions on $\Psi$ and $\mathtt{R}^s$ such that $\Psi$ is *consistent* with respect to $\mathtt{R}^s$ (see Def. 2 below). Th. 1 of (Duchi et al., 2010) states that

[1] See Letor's website http://research.microsoft.com/en-us/um/people/letor/ for the performances of several approaches on document search benchmarks.

[2] We will make an exception in Sections 5.3 and 5.4, where the supervisions considered are respectively weighted preference graphs and permutations.

consistency is a necessary and sufficient condition for a minimizer of $\mathcal{R}^{\Psi}(f)$ to be a minimizer of $\mathcal{R}^{s}(f)$. More precisely, it is equivalent to the following statement: for any sequence of functions $(f_p)_{p\geq 0}$, we have:

$$\left(\mathcal{R}^{\Psi}(f_p) \underset{p\to\infty}{\rightarrow} \inf_{g} \mathcal{R}^{\Psi}(g)\right) \Rightarrow \left(\mathcal{R}^{s}(f_p) \underset{p\to\infty}{\rightarrow} \inf_{g} \mathcal{R}^{s}(g)\right)$$

where the infima are over all measurable functions. The consistency is a pointwise property of the surrogate loss and the scoring error, which requires that for any distribution over the supervision space, any score vector that minimizes the expected loss minimizes the expected scoring error as well:

**Definition 2** *Suppose $\Psi : \mathbb{R}^n \times \mathcal{Y}$ is bounded-below and $\Psi(.,\mathbf{y})$ is continuous for all $\mathbf{y} \in \mathcal{Y}$. Then, $\Psi$ is* consistent *with respect to a scoring error $\mathtt{R}^s$ if, for random variable $\mathbf{Y}$ taking values in $\mathcal{Y}$, we have:*

$$\inf_{\mathbf{s}\in\mathbb{R}^n, \mathbf{s}\notin \arg\min\limits_{\mathbf{s}'\in\mathbb{R}^n} \mathbb{E}\left[\mathtt{R}^s(\mathbf{s}',\mathbf{Y})\right]} \mathbb{E}\left[\Psi(\mathbf{s},\mathbf{Y})\right] > \inf_{\mathbf{s}\in\mathbb{R}^n} \mathbb{E}\left[\Psi(\mathbf{s},\mathbf{Y})\right]. \quad (4)$$

## 3. Order-Preserving Losses

The surrogate losses we will consider are the *order-preserving* losses. This notion was introduced by Zhang (2004) in the context of multiclass classification. We give here a somewhat different formulation more suitable to ranking:

**Definition 3** *A function $\Phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is order-preserving if, for any bounded $\Omega \subset \mathbb{R}^n_+$ and any random variable $\mathbf{A}$ taking values in $\Omega$, we have:*

1. $\min\limits_{\mathbf{s}\in\mathbb{R}^n} \mathbb{E}\left[\Phi(\mathbf{s},\mathbf{A})\right] = \inf\limits_{\mathbf{s}\in\mathbb{R}^n} \mathbb{E}\left[\Phi(\mathbf{s},\mathbf{A})\right]$
2. *$\forall i, j$ such that $\mathbb{E}\left[A_i\right] > \mathbb{E}\left[A_j\right]$, we have:*
   $\mathbf{s}^* \in \arg\min_{\mathbf{s}\in\mathbb{R}^n} \mathbb{E}\left[\Phi(\mathbf{s},\mathbf{A})\right] \Rightarrow s^*_i > s^*_j$

The first condition is only here to simplify the proofs in the paper (it allows us to avoid dealing with scores that have to go to $-\infty$ to reach the infimum). The second condition is crucial: given a set of weights associated to each item $i$, the loss has the property of being minimal only on the score vectors which strictly follow the total preorder imposed by the weights. We may note here that Xia et al. (2008) propose a notion of order-sensitivity for surrogate losses which plays a similar role in their analysis as the order-preserving property here. The two notions are however different since order-sensitivity applies to losses based on a supervision that takes the form of permutations.

The squared loss $\Phi(\mathbf{s},\boldsymbol{\alpha}) = \sum_{i=1}^{n}\left(s_i - \alpha_i\right)^2$ is obviously order-preserving since the minimizer of $\mathbb{E}\left[\Phi(\mathbf{s},\mathbf{A})\right]$ is given by $s^*_i = \mathbb{E}\left[A_i\right]$. However, the following popular loss, which we call the *preorder loss* (Joachims, 2002; Cohen et al., 1997; Freund et al., 2003):

$$\Phi^{preorder}(\mathbf{s},\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j:\alpha_j<\alpha_i} \varphi(s_i - s_j) \quad (5)$$

is *not* order-preserving for convex $\varphi$[3]. However, Zhang (2004) suggests a class of order-preserving losses based on pairwise comparisons to replace the preorder loss:

**Theorem 1** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex, non-increasing, differentiable function satisfying:*

1. *$\varphi(t) < \varphi(-t)$ for all $t > 0$ and $\varphi'(0) < 0$,*
2. *$\varphi'(t_0) = 0$ for some $t_0 > 0$.*

*Then, $\Phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by:*

$$\Phi(\mathbf{s},\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{n} \varphi(s_i - s_j) \quad (6)$$

*is order-preserving.*

The proof of the Theorem is exactly that of Th. 5 of Zhang (2004). The first condition is mandatory to prove that the loss of (6) satisfies the second point of Definition 3. The second condition makes sure that the infimum is reached for finite values of the scores. To the best of our knowledge, the loss (6) has not been proposed for ranking. The difference between (6) and (5) is that (6) makes a comparison for every possible pair of scores $(s_i, s_j)$ (but each individual item has its own weight), while the loss of (5) only consider pairs with different weights. The problem of (5) become visible when we write the expectation: $\mathbb{E}\left[\Phi^{preorder}(\mathbf{s},\mathbf{A})\right] = \sum_{i,j} \mathbb{P}\left(A_i > A_j\right)\varphi(s_i - s_j)$. The loss has then a very complex structure, and a minimizer will not induce the same relative ordering as $\mathbb{E}\left[\mathbf{A}\right]$ in general. In contrast, the loss (6) keeps the same structure when taking the expectation over $\mathbf{A}$. This is the feature that makes it order-preserving.

The order-preserving property is rather desirable for scoring losses, in particular when the supervision takes the form of relevance judgments on each item. Yet, the exact values of the weights $A_i$ to provide to the loss must be carefully chosen depending on the observed supervision and the scoring error. The next sections investigate this issue.

[3] The proof of the inconsistency of the preorder loss w.r.t. the pairwise disagreement error (see Eq. 12) of (Duchi et al., 2010, Th. 11) works in the special case when we restrict to binary-valued $A_i$s. However, in that case, we can show that $\mathbb{E}\left[\mathbf{A}\right]$ is optimal for the pairwise disagreement, so the inconsistency w.r.t. the pairwise disagreement implies that the loss is not order preserving.

## 4. Standardization for Scoring Errors

We now consider a structural property of scoring errors: the existence of a *standardization function*, which maps the observed supervision to an optimal vector of scores, such that the expectation, over any distribution of the supervision, of these optimal vectors remains optimal for the expected scoring error:

**Definition 4** *Let* $\mathtt{R}^s : \mathfrak{S}_n \times \mathcal{Y} \to \mathbb{R}$ *be a scoring error. A* standardization function *of* $\mathcal{Y}$ *for* $\mathtt{R}^s$ *is a function* $\mathbf{r} : \mathcal{Y} \to \mathbb{R}_+^n$ *which, for any* $\mathcal{Y}$*-valued r.v.* $\mathbf{Y}$*, satisfies:*

$$\mathbb{E}\big[\mathbf{r}(\mathbf{Y})\big] \in \operatorname*{arg\,min}_{\mathbf{s}\in\mathbb{R}^n} \mathbb{E}\big[\mathtt{R}^s(\mathbf{s},\mathbf{Y})\big] \,. \tag{7}$$

The existence of such a function obviously provides a consistent surrogate loss with respect to the scoring risk using any order-preserving loss $\Phi$, by setting:

$$\forall \mathbf{s} \in \mathbb{R}^n, \forall \mathbf{y} \in \mathcal{Y}, \Psi(\mathbf{s},\mathbf{y}) = \Phi(\mathbf{s},\mathbf{r}(\mathbf{y}))\,. \tag{8}$$

In fact, the converse is true as well, as long as we consider convex and symmetric order-preserving losses:

**Theorem 2** *Let* $\mathtt{R}^s$ *be a scoring error with standardization function* $\mathbf{r}$*. Let* $\Phi$ *be an order-preserving loss. Then* $\Psi$ *defined by* (8) *is consistent with respect to* $\mathtt{R}^s$*.*

*Conversely, let* $\mathtt{R}^s$ *be a scoring error and* $\Phi$ *be an order-preserving function such that* ***(1)*** $\Phi(.,\boldsymbol{\alpha})$ *is convex for any* $\boldsymbol{\alpha}$ *and* ***(2)*** $\Phi$ *is symmetric in the following sense:*

$$\forall \boldsymbol{\alpha} \in \mathbb{R}_+^n, \forall \mathbf{s} \in \mathbb{R}^n, \forall i,j, \Phi(\mathbf{s},\alpha) = \Phi(\mathbf{s}^{i\leftrightarrow j},\alpha^{i\leftrightarrow j})\,.$$

*If there is a bounded function* $\mathbf{r} : \mathcal{Y} \to \mathbb{R}_+^n$ *such that* $\Psi$ *defined by* (8) *is consistent with respect to* $\mathtt{R}^s$*, then* $\mathbf{r}$ *is a standardization function for* $\mathtt{R}^s$*.*

*Proof* Let $\mathbf{Y}$ be a $\mathcal{Y}$-valued random variable. By our definition of order-preserving, the infimum of $\mathbb{E}\big[\Phi(.,\mathbf{r}(\mathbf{Y}))\big]$ is a minimum. So, by the definition of consistency, the first point is proved if we show that a minimizer $\mathbf{s}^*$ of $\mathbb{E}\big[\Phi(.,\mathbf{r}(\mathbf{Y}))\big]$ also minimizes $\mathbb{E}\big[\mathtt{R}^s(.,\mathbf{Y})\big]$. Given the order-preserving property, we know that $s_i^* > s_j^*$ whenever $\mathbb{E}\big[r_i(\mathbf{Y})\big] > \mathbb{E}\big[r_j(\mathbf{Y})\big]$. We do not know the relative ordering of $s_i^*$ and $s_j^*$ if $\mathbb{E}\big[r_i(\mathbf{Y})\big] = \mathbb{E}\big[r_j(\mathbf{Y})\big]$. However, by the definition of a scoring error, it is clear that the relative ordering between items $i$ and $j$ does not matter in case of a tie in an optimal score vector. Therefore $s^*$ is optimal for $\mathbb{E}\big[\mathtt{R}^s(.,\mathbf{Y})\big]$ so $\Psi$ defined by (8) is consistent w.r.t. $\mathtt{R}^s$.

For the second point, notice that for any bounded r.v. $\mathbf{A}$, $\mathbb{E}\big[\Phi(\mathbf{s},\mathbf{A})\big]$ is convex in $\mathbf{s}$ and symmetric. This implies that for any $i,j$ such that $\mathbb{E}\big[r_i(\mathbf{Y})\big] = \mathbb{E}\big[r_j(\mathbf{Y})\big]$, there is a score vector $\mathbf{s}^*$ minimizing $\mathbb{E}\big[\Phi(.,\mathbf{r}(\mathbf{Y}))\big]$ with $s_i^* = s_j^*$. Since $\Psi$ is consistent w.r.t. $\mathtt{R}^s$, $s^*$ minimizes $\mathbb{E}\big[\mathtt{R}^s(.,\mathbf{Y})\big]$. $\mathbb{E}\big[\mathbf{r}(\mathbf{Y})\big]$ and $\mathbf{s}^*$ have exactly the same ties, and the same relative ordering of $i$ and $j$ whenever $s_i^* > s_j^*$. Thus, they induce exactly the same orderings, which implies that $\mathbb{E}\big[\mathbf{r}(\mathbf{Y})\big]$ also minimizes $\mathbb{E}\big[\mathtt{R}^s(.,\mathbf{Y})\big]$.Since this implication is true for any r.v. $\mathbf{Y}$, $\mathbf{r}$ is a standardization function for $\mathtt{R}^s$. □

For a given scoring error, the theorem allows us to reduce the problem of finding a consistent surrogate formulation to the analysis of the scoring error itself: if one can explicitly find a standardization function, then we have a consistent surrogate formulation. While the theorem follows naturally from the definitions, it has a number of important consequences.

First, even if the standardization function exists and the observed supervision $\mathbf{y}$ takes the form of relevance scores, the weights $\mathbf{A}$ of the definition of an order-preserving loss should *not* be the relevance scores themselves, as demonstrated in Section 5.1. The standardization function gives us the how to set these weights, depending on the scoring error. Secondly, the theorem also points out the limitations of order-preserving surrogate losses, in the sense that they cannot help in designing consistent losses if the scoring error does not have a standardization function, as we shall see in Section 5.2. Finally, the notion of standardization function can be trivially extended to any other forms of supervision (not only relevance scores). Order-preserving losses may then be used in new contexts. This issue is discussed in Sections 5.3 and 5.4.

## 5. Special Cases

We now turn on to the specific analysis of existing scoring errors, in terms of existence of a standardization function. We first focus on well-known evaluation metrics used in IR, and then relate the notion of standardization function to positive consistency results obtained by (Duchi et al., 2010) and (Xia et al., 2008), under specific noise assumption on a supervision set which is not composed of vectors of relevance scores.

### 5.1. Discounted Cumulative Gain

The DCG (Manning et al., 2008) is a widely used evaluation metric in search engines applications. Cossock & Zhang (2008) proved the consistency of a regression approach w.r.t. the DCG. We show here that this evaluation metric has a standardization function. The usual definition of the DCG is:

$$\forall \sigma \in \mathfrak{S}_n, \mathtt{DCG@}K(\sigma,\mathbf{y}) = \sum_{k=1}^{K} \frac{2^{y_{\sigma^{-1}(k)}}-1}{\log(1+k)}\,. \tag{9}$$

$K$ is a cutoff to ignore the predicted ordering after the first $K$ items. To balance the influence of individual queries when averaging over multiple queries, The normalized DCG $\texttt{NDCG@}K(\sigma,\mathbf{y})=\frac{\texttt{DCG@}K(\sigma,\mathbf{y})}{\max\limits_{\sigma'}\texttt{DCG@}K(\mathbf{y},\sigma')}$ is used.

Th. 1 of Cossock & Zhang (2008) states that the vector of scores defined by $s_i^* = \mathbb{E}\big[2^{Y_i} - 1\big]$ maximizes the DCG@$K$ for any $K$. A similar result holds for the NDCG, since adding the normalization factor is equivalent to changing the value of the relevance scores. Thus, both the DCG@$K$ and the NDCG@$K$ have standardization functions, respectively given by:

$$\begin{aligned} r_i^{DCG@K}(\mathbf{y}) &= 2^{y_i} - 1 \\ r_i^{NDCG@K}(\mathbf{y}) &= \big(2^{y_i} - 1\big)\big(\max_{\sigma'\in\mathfrak{S}_{|q|}} \texttt{DCG@}K(\mathbf{y},\sigma')\big)^{-1} \end{aligned}$$

Consequently, any order-preserving loss leads to a consistent formulation with respect to the (N)DCG, as soon as one uses these weights.

### 5.2. Expected Reciprocal Rank and Average Precision

The ERR (Chapelle et al., 2009) was recently used as the official evaluation metric in a learning to rank challenge (Chapelle & Chang, 2011). The AP is another popular metric for binary (0/1) relevance judgments (Manning et al., 2008). Their formulas are:

$$\texttt{ERR}(\sigma,\mathbf{y}) = \sum_{k=1}^{n} \frac{R_k}{k} \prod_{j=1}^{k-1} (1 - R_j), R_k = \frac{2^{y_{\sigma^{-1}(k)}} - 1}{2^{gMax}} \quad (10)$$

$$\texttt{AP}(\sigma,\mathbf{y}) = (\sum_{i=1}^{n} y_i)^{-1} \sum_{i:y_i=1} \frac{1}{\sigma(i)} \sum_{k:y_k=1} I_{\{\sigma(k)\le\sigma(i)\}} \quad (11)$$

where $gMax$ in (10) is the greatest possible relevance score. Both have a highly complex structure, and there is no consistent surrogate loss w.r.t. the ERR or the AP when considering order-preserving losses:

**Theorem 3** *There is no standardization function for the ERR and for the AP.*

The proof of the theorem is based on the following lemma, which gives some properties of a standardization function on binary relevance judgments.

**Lemma 4** *Fix $n \ge 3$ and $\mathcal{Y} = \{0,1\}^n$. Let $\mathtt{R}^s$ be a scoring error with associated ordering error $\mathtt{R}^r$ such that for any $\sigma \in \mathfrak{S}_n$, any two indexes $i, j$ and any $\mathbf{y} \in \mathcal{Y}$:*

1. $\mathtt{R}^r(\sigma,\mathbf{y}) = \mathtt{R}^r(\sigma^{i\leftrightarrow j}, \mathbf{y}^{i\leftrightarrow j})$ *(symmetry),*
2. $(y_i > y_j$ *and* $\sigma(i) < \sigma(j)) \Rightarrow \mathtt{R}^r(\sigma,\mathbf{y}) < \mathtt{R}^r(\sigma^{i\leftrightarrow j}, \mathbf{y})$ *(strict monotonicity),*

*Then, for any standardization function* $\mathbf{r}$ *for* $\mathtt{R}^s$*:*

1. $\forall \mathbf{y} \in \mathcal{Y}, \forall i, j, y_i > y_j \Rightarrow r_i(\mathbf{y}) > r_j(\mathbf{y})$,
2. $\forall \mathbf{y} \in \mathcal{Y}, \forall i, j, y_i = y_j \Rightarrow r_i(\mathbf{y}) = r_j(\mathbf{y})$,
3. $\forall \mathbf{y}, \mathbf{y}' \in \mathcal{Y}$ *s.t.* $\sum_k y_k = \sum_k y'_k$*, for any* $i, j$*:* $y_i + y'_i = y_j + y'_j \Rightarrow r_i(\mathbf{y}) + r_i(\mathbf{y}') = r_j(\mathbf{y}) + r_j(\mathbf{y}')$.

Before proving the lemma, we prove the theorem:

*Proof of Theorem 3* Consider the binary relevance case with 4 items to rank, the two supervision vectors $\mathbf{y} = (1,1,0,0)$ and $\mathbf{y}' = (0,0,1,1)$, and the r.v. $\mathbf{Y}$ which gives 1/2 probability to each of them. Aiming for a contradiction, suppose the ERR (resp. the AP) has a standardization function $\mathbf{r}^{ERR}$ (resp. $\mathbf{r}^{AP}$). Then, by the third point of Lemma 4, $\mathbb{E}\big[r_i^{ERR}(\mathbf{Y})\big]$ (resp. $\mathbb{E}\big[r_i^{AP}(\mathbf{Y})\big]$) does not depend on $i$. If this is an optimal score vector, then any permutation of the four items is optimal. Computing the ERR and the AP for the rankings $1\succ 2\succ 3\succ 4$ and $1\succ 3\succ 2\succ 4$, we find:

$$\mathbb{E}\big[\texttt{ERR}(1\succ 3\succ 2\succ 4, \mathbf{Y})\big] = \mathbb{E}\big[\texttt{ERR}(1\succ 2\succ 3\succ 4, \mathbf{Y})\big] + \frac{1}{24}$$

$$\mathbb{E}\big[\texttt{AP}(1\succ 3\succ 2\succ 4, \mathbf{Y})\big] = \mathbb{E}\big[\texttt{AP}(1\succ 2\succ 3\succ 4, \mathbf{Y})\big] - \frac{1}{12}$$

Thus, some permutations are suboptimal, which contradicts the existence of a standardization function. □

*Proof of Lemma 4* Point 1 follows directly from the strict monotonicity. In the rest of the proof, for any $i$, $\mathbf{y}^i$ is the vector defined by $y_i^i = 1$ and $y_k^i = 0$ for $k \neq i$.

We prove the second point by contradiction. Suppose $\mathbf{y}$ is such that $y_i = y_j$ and $r_i(\mathbf{y}) > r_j(\mathbf{y})$. Define $\mathbf{Y}^\alpha$ by $\mathbb{P}\big(\mathbf{Y}^\alpha = \mathbf{y}^j\big) = \alpha$ and $\mathbb{P}\big(\mathbf{Y}^\alpha = \mathbf{y}\big) = 1\text{-}\alpha$, for $\alpha > 0$ such that $(1-\alpha)r_i(\mathbf{y}) + \alpha r_i(\mathbf{y}^j) > (1-\alpha)r_j(\mathbf{y}) + \alpha r_j(\mathbf{y}^j)$. By strict monotonicity ($\mathbf{y}^j$ requires $j$ ranked before $i$) and symmetry (the relative ordering of $i$ and $j$ does not matter for $\mathbf{y}$), the small probability $\alpha$ implies $\mathbb{E}\big[r_j(\mathbf{Y}^\alpha)\big] > \mathbb{E}\big[r_i(\mathbf{Y}^\alpha)\big]$, which is impossible considering our choice of $\alpha$.

We also prove the third point by contradiction. Suppose there are $\mathbf{y}, \mathbf{y}'$ and two indexes $i$ and $j$ such that $y_i + y'_i = y_j + y'_j$ and $r_i(\mathbf{y}) + r_i(\mathbf{y}') > r_j(\mathbf{y}) + r_j(\mathbf{y}')$. Notice that by the second point, we necessarily have $y_i \neq y_j$, thus $y_i + y'_i = 1$. Without loss of generality, assume $y_i = 1$ (thus, $y_j = y'_i = 1$ and $y'_i = 0$). Let $\mathbf{Y}^\beta$ s.t. $\mathbb{P}\big(\mathbf{Y}^\beta = \mathbf{y}^j\big) = \beta$, $\mathbb{P}\big(\mathbf{Y}^\beta = \mathbf{y}\big) = \mathbb{P}\big(\mathbf{Y}^\beta = \mathbf{y}'\big) = \frac{1\text{-}\beta}{2}$, and $\beta > 0$ small enough so that $\mathbb{E}\big[r_i(\mathbf{Y}^\beta)\big] > \mathbb{E}\big[r_j(\mathbf{Y}^\beta)\big]$. Since $\sum_k y_k = \sum_k y'_k$ and using the symmetry of the ranking risk, we can claim that $\mathbf{y}$ and $\mathbf{y}'$ do not impose any constraint on the relative ordering of any two items for which $y_i + y'_i = y_j + y'_j$. The probability $\beta$ imposes $\mathbb{E}\big[r_j(\mathbf{Y}^\beta)\big] > \mathbb{E}\big[r_i(\mathbf{Y}^\beta)\big]$ by strict monotonicity. This is impossible considering our choice of $\beta$. □

### 5.3. Pairwise Disagreement with Low-Noise

We considered throughout the paper that the supervision takes the form of scores. However, a popular form of supervision for ranking is a (weighted) preference graph (Cohen et al., 1997; Freund et al., 2003; Dekel et al., 2003; Duchi et al., 2010). Consider $\mathcal{Y}$ to be the set of directed acyclic graphs with non-negative weights, and consider the weighted pairwise disagreement (WPD) defined by Duchi et al. (2010)[4]:

$$\forall \sigma \in \mathfrak{S}_n, \mathtt{WPD}(\sigma, \mathbf{y}) = \sum_{i \to j \in \mathbf{y}} a_{i,j}(\mathbf{y}) I_{\{\sigma(j) > \sigma(i)\}} \qquad (12)$$

where $a_{i,j}(\mathbf{y})$ is the cost of ordering $i$ before $j$ given the preference graph $\mathbf{y}$. Duchi et al. (2010) showed that many pairwise comparisons losses, including the preorder loss of Equation 5 and weighted versions of it, are not consistent in general with respect to the WPD. They however proposed a new, consistent loss (a linear loss with a penalty term on the predicted scores) for the case where the conditional distribution of the supervision given the query $q$ $\mathbf{T}_{|Q=q}$ has *low noise*. The exact definition of low noise for a $\mathcal{Y}$-valued r.v. $\mathbf{Y}$ is the following (Duchi et al., 2010, Def. 8):

$$\forall i, j, k, \mathbb{E}\big[a_{i,j}(\mathbf{Y}) - a_{j,i}(\mathbf{Y})\big] \geq \\ \mathbb{E}\big[a_{i,k}(\mathbf{Y}) - a_{k,i}(\mathbf{Y})\big] + \mathbb{E}\big[a_{k,j}(\mathbf{Y}) - a_{j,k}(\mathbf{Y})\big]$$

The interested reader may refer to the original paper for discussions on how restrictive the assumption is. In this low-noise setting, the proof of Th. 13 of (Duchi et al., 2010) states that an optimal score vector for the scoring error associated to the WPD is $\mathbf{s}^*$ defined by:

$$s_i^* = \mathbb{E}\big[\sum_{k=1}^{n} (a_{i,k}(\mathbf{Y}) - a_{k,i}(\mathbf{Y}))\big].$$

After trivial modifications to our definitions to handle the case of preference graphs for the supervision, if the weights $a_{i,j}(\mathbf{y})$ are bounded by some constant $C$, the WPD has a standardization function defined by:

$$r_i^{WPD}(\mathbf{y}) = nC + \sum_{k=1}^{n} \big(a_{i,k}(\mathbf{y}) - a_{k,i}(\mathbf{y})\big) .$$

The term $nC$ is just added to guarantee the non-negativity of $r_i^{WPD}(\mathbf{y})$. As such, the pairwise loss defined in Th. 1 1 leads to a consistent formulation with respect to the WPD if one uses $\mathbf{r}^{WPD}$ to set the weight of each item. The result is somewhat surprising, since order-preserving losses are not really intuitive when the observed supervision is a preference graph.

### 5.4. $0/1$-Error and Order-Preserving Permutation Probability Spaces

Xia et al. (2008) prove the consistency of various algorithms when the supervision set $\mathcal{Y} = \mathfrak{S}_n$. The consistency was proved w.r.t. the $0/1$ error on permutations:

$$\forall \sigma \in \mathfrak{S}_n, \forall \mathbf{y} \in \mathcal{Y} = \mathfrak{S}_n, \mathtt{ZO}(\sigma, \mathbf{y}) = I_{\{\sigma = \mathbf{y}\}} \qquad (13)$$

and require an additional constraint on the conditional distributions of the supervision given the query $q$ $\mathbf{T}_{|Q=q}$: for each query, the distribution must belong to an *order preserving permutation probability space* (see Def. 5 below) with respect to $n-1$ pairs of items $(j_1^q, j_2^q), (j_2^q, j_3^q), ..., (j_{n-1}^q, j_n^q)$, where each $j_k^q \neq j_p^q$ for $k \neq p$. This constraint is similar to the Plackett-Luce model (up to the possibility of assigning a probability of 0 to some permutations), for which consistency w.r.t. the $0/1$-error was also pointed out by (Cheng et al., 2010). The definition of an order-preserving permutation probability space is (Xia et al., 2008, Def. 3):

**Definition 5** *Let $i, j \in \{1, ..., n\}$ and $\mathbf{Y}$ be a $\mathfrak{S}_n$-valued random variable. $\mathbf{Y}$ defines a permutation probability space which is* order-preserving *with respect to $(i, j)$ if, for any $\sigma$ s.t. $\sigma(i) < \sigma(j)$, we have $\mathbb{P}(\mathbf{Y} = \sigma) > \mathbb{P}(\mathbf{Y} = \sigma^{i \leftrightarrow j})$ or $\mathbb{P}(\mathbf{Y} = \sigma) = \mathbb{P}(\mathbf{Y} = \sigma^{i \leftrightarrow j}) = 0$.*

Let us now consider $\mathbf{r}^{ZO} : \mathfrak{S}_n \times \mathcal{Y} \to \mathbb{R}_+^n$ defined by:

$$r_i^{ZO}(\mathbf{y}) = \sum_{k=1}^{n} I_{\{\mathbf{y}(i) \leq k\}}$$

(recall that $\mathbf{y}$ is a permutation, and $\mathbf{y}(i)$ is the rank of item $i$ by convention). Note that even though the same term *order-preserving* is used for permutation probabilities and the surrogate losses, these correspond to totally different notions. We now give a sketch of proof that if $\mathbf{Y}$ defines an order-preserving probability space with respect to $(j_k, j_{k+1}), k = 1..n-1$ for distinct indexes $j_k$, then $\mathbb{E}\big[\mathbf{r}^{ZO}(\mathbf{Y})\big]$ is an optimal score vector.

First, let us recall that in the proof of Theorem 5 of (Xia et al., 2008), the authors showed that the optimal permutation for the $0/1$ error is $\sigma^*$ defined by $\sigma^*(j_k) = k$ for all $k$. We can take $j_k = k$ to simplify the notations without loss of generality. For two indexes $i < j$, it is easy to see that the order-preserving property (of the probability distribution of the labels) guarantees that for any $k$, $\mathbb{P}\big(\mathbf{Y}(i) \leq k\big) \geq \mathbb{P}\big(\mathbf{Y}(j) \leq k\big)$, and the inequality must be strict for $k = i$ (because there is at least one permutation with $\mathbf{y}(i) = i$ which has non-zero probability: the optimal one). This proves $\mathbb{E}\big[r_i^{ZO}(\mathbf{Y})\big] > \mathbb{E}\big[r_j^{ZO}(\mathbf{Y})\big]$, and we recover the optimal ordering for the $0/1$-loss.

[4] There is a minor difference with the definition of Eq. 7 of Duchi et al. (2010) in the way we handle a tie between $i$ and $j$, but it does not affect our result.

Thus, if the supervision set is the set of permutations and with the same distributional assumption on the supervision as (Xia et al., 2008), we can find a consistent surrogate formulation (with respect to the 0/1-error on permutations) with order-preserving losses.

## 6. Experiments

Theorems 1 and 2 show that we can define consistent, pairwise surrogate losses w.r.t. any scoring error which has a standardization function. We compare, in this section, the new consistent pairwise loss with existing variants of the preorder loss when a standardization function is given. In that case, the new pairwise loss has stronger theoretical guarantees than preorder losses while keeping the simplicity (in terms of implementation) of existing pairwise approaches. We believe that if the new loss achieves similar or better empirical performances to the preorder loss, it should be considered as a valid replacement for the latter.

We carry out experiments on two benchmark datasets: MQ2007 from Letor 4.0[5] and YLTRC[6], the dataset used in the Yahoo! Learning to Rank Challenge (Chapelle & Chang, 2011). The first one contains 1700 queries with three relevance levels $(0, 1, 2)$ for the supervision. YLTRC contains $30,000$ queries with five relevance levels. Both datasets have about 25 documents per query. We carry on two types of experiments. The first one compares the "blind" application of both the preorder and consistent pairwise loss, while the second type compares the two losses with respect to well-known tuning of the pairwise losses. The two losses we compare have the following form:

Preorder loss a general form of the preorder loss:

$$\Phi^{preorder}(\mathbf{s}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j:\alpha_j<\alpha_i} w_{i,j}(\alpha)\varphi(s_i - s_j) \quad (14)$$

Consistent loss based on Equation 6 for which $\alpha$ is set to the standard form of the supervision for the (non-truncated) DCG or NDCG. The experimental setup is as follows: we use a differentiable version of the hinge loss (see Chapelle (2007)) for $\varphi$, to keep the order-preserving property (see Th. 1) of the consistent losses (the same $\varphi$ is used for all algorithms). We consider linear scoring functions trained by minimizing the loss together with a regularization term $\frac{\lambda}{2}||\mathbf{w}||^2$ where $\mathbf{w}$ is the parameter vector. Minimization is carried out with Stochastic Gradient Descent (Bottou, 2004). The use of linear functions is motivated by their widespread use in conjunction with pairwise losses. The regularization hyperparameter $\lambda$ is chosen with a crude search on the validation set, using either the DCG (9), NDCG or the ERR (10 depending on the final evaluation metric. We also evaluate in ERR since it was the official measure of the Yahoo! Challenge. We use non-truncated (N)DCG because our losses are designed and learned to be consistent with them, but results with truncated versions are similar.

*Table 1.* Test performances on YLTRC (top) and Letor (bottom) datasets. ↓ means significantly worse performance than the best run.

| | ERR | DCG | NDCG |
|---|---|---|---|
| PreOrder | 0.4392↓ | 1.6894↓ | 0.8332↓ |
| Consist$^{\text{DCG}}$ | 0.4469 | **1.7025** | 0.8349↓ |
| Consist$^{\text{NDCG}}$ | **0.4472** | 1.7021 | **0.8374** |
| PreOrder | 0.31392 | 1.7678 | 0.66696 |
| Consist$^{\text{DCG}}$ | **0.316** | **1.76806** | 0.66714 |
| Consist$^{\text{NDCG}}$ | 0.3134 | 1.76566 | **0.66858** |

**Consistent Loss vs Preorder Loss** We report here on the results of the "blind" application of the algorithms. For the preorder loss, the weights $w_{i,j}(\alpha) = 1$, which corresponds to the standard preorder loss (5). The test performances obtained with this loss and with the consistent pairwise loss (w.r.t. the DCG, and to the NDCG) are shown in Table 1 top (resp. 1 bottom) for YLTRC (resp. Letor). The results show that the performances of the preorder loss are significantly worse than those of the consistent losses. The choice of the standardization function also plays an important role for optimizing a particular measure: the consistent loss w.r.t. the NDCG gives a significant improvement, in terms of NDCG, over both the consistent loss w.r.t. the DCG and the preorder loss (on YLTRC, the differences are not significant on Letor).

**Comparison with tuned Preorder Losses** While the preorder loss is the most popular algorithm based on pairwise comparisons, some weighting heuristics have been proposed to improve its performance on IR tasks. For instance, Cao et al. (2006) proposed to balance the influence of each query in the empirical risk by normalizing the loss for each query, and to weight each pair depending on the relevance of its members. We study here the performance of our methods against such tuned preorder losses, one using query normalization ($w_{i,j}(\boldsymbol{\alpha}) = \frac{1}{\#comparisons}$ in (14)) only and another one with both query normalization and relevance weighting ($w_{i,j}(\boldsymbol{\alpha}) = \frac{2^{\alpha_i}-2^{\alpha_j}}{\#comparisons}$)

[5] http://research.microsoft.com/en-us/um/beijing/projects/letor/; we only considered MQ2007 because MQ2008 is much smaller.

[6] http://learningtorankchallenge.yahoo.com

*Table 2.* Test performances on YLTRC (top) and Letor (bottom) datasets.

| | ERR | DCG | NDCG |
|---|---|---|---|
| PreOrder$^{\text{Norm}}$ | 0.4458↓ | 1.6988↓ | 0.8370 |
| PreOrder$^{\text{Norm+DCG}}$ | 0.4499 | 1.7072 | **0.8376** |
| Consist$^{\text{Norm+DCG}}$ | **0.4511** | **1.7082** | 0.8370 |
| PreOrder$^{\text{Norm}}$ | 0.31278 | 1.76412 | 0.66794 |
| PreOrder$^{\text{Norm+DCG}}$ | **0.31776** | 1.77058 | **0.6714** |
| Consist$^{\text{Norm+DCG}}$ | 0.31746 | **1.77124** | 0.67032 |

$(2^{y_i} - 2^{y_j})$. The loss consistent with the DCG also has a query-normalization term $(\frac{1}{n(n-1)})$, which does not affect consistency. The results are reported in Table 2. While all results seem similar, it still validates the consistent loss which enjoys better theoretical guarantees than the heuristics for the preorder loss.

## 7. Conclusion

We presented an analysis of the consistency of order-preserving losses with respect to various ranking metrics. The consistency of the surrogate formulation is bound to the existence of a standardization function for the scoring error. In contrast to previous analysis on the consistency of ranking algorithms, we were able to prove (or disprove) the existence of consistent formulations with respect to various scoring errors by only analyzing the scoring error itself.

Our work also suggests a new approach to building pairwise surrogate losses. The new loss achieves similar or better results compared to existing pairwise loss with highly tuned heuristic weights on the pairs of items. While many heuristics can be envisioned for the preorder loss, our analysis provides a simple and theoretically grounded rule for designing the pairwise loss depending on the final evaluation metric, in the special case where a standardization function is given.

### Acknowledgments

This work was partially funded by the French DGA and the FUI project OpenWay III.

## References

Bottou, Léon. Stochastic learning. In Bousquet, Olivier and von Luxburg, Ulrike (eds.), *Advanced Lectures on Machine Learning*, Lecture Notes in Artif. Intel. 3176, pp. 146–168. Springer Verlag, 2004.

Cao, Yunbo, Xu, Jun, Liu, Tie-Yan, Li, Hang, Huang, Yalou, and Hon, Hsiao-Wuen. Adapting ranking svm to document retrieval. In *Proc. of the 29th SIGIR Conf. on Research and Development in Inf. Ret.*, 2006.

Chapelle, Olivier. Training a support vector machine in the primal. *Neural Comput.*, 19:1155–1178, 2007.

Chapelle, Olivier and Chang, Yi. Yahoo! learning to rank challenge overview. *J. of Mach. Learn. Res*, 14:1–24, 2011.

Chapelle, Olivier, Metlzer, Donald, Zhang, Ya, and Grinspan, Pierre. Expected reciprocal rank for graded relevance. In *Proc. of the Conference on Information and Knowledge Management*, pp. 621–630, 2009.

Cheng, Weiwei, Dembczynski, Krzysztof, and Hüllermeier, Eyke. Label Ranking Methods based on the Plackett-Luce Model. In *Proc. of the Intl. Conf. on Mach. Learn.*, pp. 215–222, 2010.

Cohen, William W., Schapire, Robert E., and Singer, Yoram. Learning to order things. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 1997.

Cossock, David and Zhang, Tong. Statistical analysis of bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, 2008.

Dekel, Ofer, Manning, Christopher D., and Singer, Yoram. Log-linear models for label ranking. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2003.

Duchi, John, Mackey, Lester W., and Jordan, Michael I. On the consistency of ranking algorithms. In *Proc. of the Intl. Conf. on Mach. Learn.*, pp. 327–334, 2010.

Freund, Yoav, Iyer, Raj, Schapire, Robert E., and Singer, Yoram. An efficient boosting algorithm for combining preferences. *J. of Mach. Learn. Res.*, 4:933–969, 2003.

Herbrich, Ralf, Graepel, Thore, and Obermayer, Klaus. Large margin rank boundaries for ordinal regression. In Smola, Bartlett, Schoelkopf, and Schuurmans (eds.), *Advances in Large Margin Classifiers*. 2000.

Joachims, Thorsten. Optimizing search engines using clickthrough data. In *Proc. of Knowledge Disc. and Data Mining (KDD)*, 2002.

Manning, Christopher D., Raghavan, Prabhakar, and Schtze, Hinrich. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

Weston, Jason and Watkins, Christopher. Support vector machines for multi-class pattern recognition. In *Eur. Symp. On Art. Neural Net.*, pp. 219–224, 1999.

Xia, Fen, Liu, Tie-Yan, Wang, Jue, Zhang, Wensheng, and Li, Hang. Listwise approach to learning to rank: theory and algorithm. In *Proc. of the Intl. Conf. on Mach. Learn.*, 2008.

Yue, Yisong, Finley, Thomas, Radlinski, Filip, and Joachims, Thorsten. A support vector method for optimizing average precision. In *Proc. of the 30th SIGIR Conf. on Research and Development in Inf. Ret.*, 2007.

Zhang, Tong. Statistical analysis of some multi-category large margin classification methods. *J. of Mach. Learn. Res.*, 5:1225–1251, 2004.