Calibration and regret bounds for order-preserving surrogate losses in learning to rank

Clément Calauzènes, Nicolas Usunier, Patrick Gallinari

Noname manuscript No.
(will be inserted by the editor)

Calibration and Regret Bounds for Order-Preserving Surrogate Losses in Learning to Rank*

Clément Calauzènes · Nicolas Usunier · Patrick Gallinari

Received: date / Accepted: date

Abstract Learning to rank is usually reduced to learning to score individual objects, leaving the "ranking" step to a sorting algorithm. In that context, the surrogate loss used for training the scoring function needs to behave well with respect to the target performance measure which only sees the final ranking. A characterization of such a good behavior is the notion of calibration, which guarantees that minimizing (over the set of measurable functions) the surrogate risk allows us to maximize the true performance.

In this paper, we consider the family of order-preserving (OP) losses which includes popular surrogate losses for ranking such as the squared error and pairwise losses. We show that they are calibrated with performance measures like the Discounted Cumulative Gain (DCG), but also that they are *not* calibrated with respect to the widely used Mean Average Precision and Expected Reciprocal Rank. We also derive, for some widely used OP losses, quantitative surrogate regret bounds with respect to several DCG-like evaluation measures.

Keywords Learning to rank; calibration; surrogate regret bounds

1 Introduction

Learning to rank has emerged as a major field of research in machine learning due to its wide range of applications. Typical applications include creating the query-dependent document ranking in search engines, where one learns to order sets of documents, each of these sets being attached to a query, using relevance

* This work contains material from Sections 3, 4 and 5 of [3]. It substantially extends its framework and mostly contains new results.

C. Calauzènes, N. Usunier, P.Gallinari
University Pierre et Marie Curie
Department of Computer Science (Laboratoire d'Informatique de Paris 6)
E-mail: first.last@lip6.fr
Present address: of N. Usunier
Heudiasyc, Université Technologique de Compiègne
E-mail: nicolas.usunier@hds.utc.fr

judgments for each document as supervision. This task is known as subset ranking [14]. Another application is label ranking (see e.g. [16, 31]), where one learns to order a fixed set of labels depending on an input with a training set composed of observed inputs and the corresponding weak or partial order over the label set. Label ranking is a widely used framework to deal with multiclass/multilabel classification when the application accepts a ranking of labels according to the posterior probability of class membership instead of a hard decision about class membership.

In a similar way to other prediction problems in discrete spaces like classification, the optimization of the empirical ranking performance over a restricted class of functions is most frequently an intractable problem. Just like one optimizes the hinge loss or the log-loss in binary classification as surrogate for the classification error, the usual approach in learning to rank is to replace the original performance measure by a continuous, preferably differentiable and convex function of the predictions. This has lead many researchers to reduce the problem of learning to rank to learning a *scoring function* which assigns a real value to each individual item of the input set. The final ranking is then produced with a sorting algorithm. Many existing learning algorithms follow this approach both for label ranking (see e.g. [33, 15, 16]) and subset ranking [21, 4, 34, 14, 25, 5, 9]. This relaxation has two advantages. First, the sorting algorithm is a very efficient way to obtain a ranking (without scores, obtaining a full ranking is usually a very difficult problem). Second, defining a continuous surrogate loss on the space of predicted scores is a much easier task than defining one in the space of permutations.

While the computational advantage of such surrogate formulations is clear, one needs to have guarantees that minimizing the surrogate formulation (i.e. what the learning algorithm actually does) also enables us to maximize the ranking performance (i.e. what we want the algorithm to do). That is, we want the learning algorithm to be consistent with the true ranking performance we want to optimize. [28] presents general definitions and results showing that an asymptotic guarantee of consistency is equivalent to a notion of *calibration* of the surrogate loss with respect to the ranking performance measure, while the existence of non-asymptotic guarantees in the form of surrogate regret bounds are equivalent to a notion of *uniform calibration*. A surrogate regret bound quantifies how fast the evaluation measure is maximized as the surrogate loss is minimized. We note here that such non-asymptotic guarantees are critical in machine learning where it is delusive to hope for learning near-optimal functions in a strong sense. The calibration and uniform calibration have been extensively studied in (cost-sensitive) binary classification (see e.g. [2, 35, 28, 27]) and multiclass classification [35, 29]. In particular, under natural continuity assumptions, it was shown that calibration and uniform calibration of a surrogate loss are equivalent for margin losses. In the context of learning to rank with pairwise preferences, the non-calibration of many existing surrogate losses with respect to the pairwise disagreement was studied in depth in [18]. On the other hand, in the context of learning to rank for information retrieval, surrogate regret bounds for square-loss regression with respect to a ranking performance measure called the Discounted Cumulative Gain (DCG, see [20]) was shown in [14]. These bounds were further extended in [26] to a larger class of surrogate losses.

In this paper, we analyze the calibration and uniform calibration of losses that possess an *order-preserving* property. This property of a surrogate loss implies

(and, to some extend, is equivalent to) the calibration with respect to a ranking performance measure in the family what we call the *generalized positional performance measures* (GPPMs). A GPPM is a performance measure which, up to a suitable parametrization, can be written like a DCG. The study of GPPMs offer, in particular, the possibility to extend the DGC to arbitrary supervision spaces (e.g. linear orders, pairwise preferences) by first mapping the supervision to scores for each item – scores that can be interpreted as utility values. The family of GPPM includes widely known performance measures for ranking like the DCG and its normalized version the NDCG, the precision-at-rank-K as well as the recall-at-rank-K, and Spearman's rank correlation coefficient (when the supervision is a linear ordering). We also give practical examples of *template* order-preserving losses, which can be instantiated for any specific GPPM to obtain a calibrated surrogate loss.

To go further, we investigate conditions under which the stronger notion of uniform calibration holds in addition to the simple calibration. Under natural continuity conditions for the loss function, we show that any loss calibrated with a GPPM is uniformly calibrated when the supervision space is finite, which stands for the existence of a regret bound. Finally, we prove explicit regret bounds for several convex template order-preserving losses. These bounds can be instantiated for any GPPM, such as the (N)DCG, and recall/precision-at-rank-K. In particular, we obtain the first regret bounds with respect to GPPMs for losses based on pairwise comparisons, and recover the surrogate regret bounds of [14] and [26]. Our proof technique is different though, and we are able to slightly improve the constant factor in the bounds.

As a by-product of our analysis, we investigate whether a loss with some order-preserving property can be calibrated with two other measures than GPPMs, namely the Expected Reciprocal Rank [10] used as reference in the recent Yahoo! Learning to Rank Challenge [9] and the Average Precision which was used in past Text REtrieval Conferences (TREC) competitions [32]. We surprisingly show a negative result – even though these measures assume that the supervision itself takes the form of real values (relevance scores). Our result implies that for any transformation of the relevance scores given as supervision, the regression function of these transformations is not optimal for the ERR or the AP in general. We do believe that this result can help understand the limitations of the score-and-sort approach to ranking, and put emphasis on an often neglected fact: the choice of the surrogate formulation is not really a matter of supervision space, but should be carefully chosen depending on the target measure. For example, one can use an appropriate regression approach to optimize Pearson's correlation coefficient when the supervision is a full ordering of the set, but for the ERR or the AP, regression approaches cannot be calibrated even though one gets real values for each item as supervision.

The rest of the paper is organized as follows, Section 2 describes the framework and the basic definitions. In Section 3, we introduce a family of surrogate losses called the *order-preserving losses*, and we study their calibration with respect to a wide range of performance measures. Then, we propose an analysis of sufficient conditions under which calibration is equivalent to the existence of a surrogate regret bound; this analysis is carried out by studying the stronger notion of uniform calibration in Section 4. In Section 5, we describe several methods to find explicit formulas of surrogate regret bounds, and we exhibit some examples for

common surrogate losses. The related work is discussed in Section 6, where we also summarize our main contributions.

2 Ranking Performance Measures and Surrogate Losses

Notation A boldface character always denotes a function taking values in $\mathbb{R}^n$ or an element of $\mathbb{R}^n$ for some $n > 1$. If $\mathbf{f}$ is a function of x, then $f_i(x)$, using normal font and subscript, denotes the i-th component of $\mathbf{f}(x)$. Likewise, if $\mathbf{x} \in \mathbb{R}^n$, x_i denotes its i-th component.

2.1 Definitions and Examples

Scoring Functions and Scoring Performance Measures The prediction task we consider is the following: considering a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$ and some integer $n > 1$, the goal is to predict an ordering of a fixed set of n objects, which we identify with the set of indices $\{1, ..., n\}$, for any $x \in \mathcal{X}$. This ordering is predicted with a score-and-sort approach. A *scoring function* $\mathbf{f}$ is any measurable function $\mathbf{f} : \mathcal{X} \to \mathbb{R}^n$, and the ordering of the set $\{1, ..., n\}$ given $x \in \mathcal{X}$ is obtained by sorting the integers i by decreasing values of $f_i(x)$. Identifying the linear orders of $\{1, ..., n\}$ with the set $\mathfrak{S}_n$ of permutations of $\{1, ..., n\}$, the predicted ordering can thus be any permutation in $\arg\operatorname{sort}(\mathbf{f}(x))$, where:

$$\arg\operatorname{sort} : \mathbf{s} \in \mathbb{R}^n \mapsto \left\{ \sigma \in \mathfrak{S}_n \,\middle|\, \forall 0 < k < n, s_{\sigma(k)} \geq s_{\sigma(k+1)} \right\}$$

Notice that with these definitions, for $\sigma \in \arg\operatorname{sort}(\mathbf{f}(x))$, $\sigma(k)$ denotes the integer of $\{1, ..., n\}$ whose predicted rank is k. Following the tradition in information retrieval, "item i has better rank than item j according to σ" means "$\sigma^{-1}(i) < \sigma^{-1}(j)$", i.e. low ranks are better. Likewise, the top-d ranks stands for the set of ranks $\{1, ...d\}$. Also notice that $\arg\operatorname{sort}$ is a set-valued function because of possible ties.

Predicting a total order over a finite set of objects is of use in many applications. This can be found in information retrieval, where x represents a tuple (query, set of documents) and the score $f_i(x)$ is the score given to the i-th document in the set given the query. In practice, x contains joint feature representations of (query, document) pairs and $f_i(x)$ is the predicted relevance of document i with respect to the query. The learning task associated to this prediction problem has been called subset ranking in [14]. Note that in practice, the set of documents for different queries may vary in size, while in our work it is supposed to be constant. Anyway, all our results hold if one allows to vary the set size, keeping it uniformly bounded. Another example of application is label ranking (see e.g. [16]) where x is some observed object such as a text document or an image, and the set to order is a fixed set of class labels. In that case, x usually contains a feature representation of the object, and a function f_i is learnt for each label index i; higher values of $f_i(x)$ represent higher predicted class-membership probabilities. Large-margin approaches to multiclass classification (see e.g. [33, 15]) are special cases of a score-and-sort approach to label ranking, where the prediction is the top-ranked label.

In the supervised learning setting, the prediction function is trained using a set of examples for which some feedback, or supervision, indicative of the desired ordering is given. In order to measure the quality of a predicted ordering of $\{1, ..., n\}$, a *ranking performance measure* is used. It is a measurable function $r : \mathcal{Y} \times \mathfrak{S}_n \to \mathbb{R}_+$, where $(\mathcal{Y}, \Sigma_\mathcal{Y})$ is a measurable space which will be called the *supervision space*, and each $y \in \mathcal{Y}$ provides information about which ordering are desired. We take the convention that larger values of $r(y, \sigma)$ mean that σ is an ordering of $\{1, ..., n\}$ in accordance to y. The supervision space may differ from one application to the other. In search engine applications, the Average Precision (AP) used in past TREC competitions [32], the Expected Reciprocal Rank (ERR) used in the Yahoo! Learning to Rank Challenge [10, 9] or the (Normalized) Discounted Cumulative Gain ((N)DCG) [20], assume the supervision space $\mathcal{Y}$ is defined as $\{0, ..., p\}^n$ for some integer $p > 0$, where the i-th component of $y \in \mathcal{Y}$ is a judgment of the relevance of the i-th item to rank w.r.t. the query (these performance measures always favor better ranks for items of higher relevance). Other forms of supervision spaces may be used though, for instance in recommendation tasks we may allow user ratings on a continuous scale (yet usually bounded), or allow the supervision to be a preference relation over the set of items, as proposed in one of the earliest papers on learning to rank [13].

Most ranking performance measures used in practical application have the form described above: they are defined on a prediction space which is exactly the set of linear orderings, and do not directly take ties into account. In order to define performance measures for scoring functions, we take the convention that ties are broken randomly. Thus, given a ranking performance measure r, we overload the notation r (the context is clear given the arguments' names) to define a *scoring performance measure* as follows:

$$\forall \mathbf{s} \in \mathbb{R}^n, \forall y \in \mathcal{Y}, r(y, \mathbf{s}) = \frac{1}{|\arg\operatorname{sort}(\mathbf{s})|} \sum_{\sigma \in \arg\operatorname{sort}(\mathbf{s})} r(y, \sigma)$$

where $|S|$ denotes the cardinal of a set. Of particular importance in our work will be the following family of ranking performance measures, which we call *generalized positional performance measures* (GPPM):

Definition 2.1 (Generalized Positional Performance Measure) *Let $r : \mathcal{Y} \times \mathfrak{S}_n \to \mathbb{R}_+$ be a ranking performance measure. We say that r is a* $(\boldsymbol{u}, \phi)$-generalized positional performance measure *(abbreviated $(\boldsymbol{u}, \phi)$-GPPM) if $\phi : \{1..n\} \to \mathbb{R}_+$, $\boldsymbol{u} : \mathcal{Y} \to \mathbb{R}_+^n$ is measurable, and:*

1. *$\phi(1) > 0$ and $\forall 0 < k < n, \phi(k) \geq \phi(k+1) \geq 0$*
2. *$\exists b : \mathcal{Y} \to \mathbb{R}$, such that $r : (y, \sigma) \mapsto b(y) + \sum_{k=1}^n \phi(k) u_{\sigma(k)}(y)$.*

The most popular example of a GPPM is the DCG, for which $\phi(i) = \frac{\mathbf{1}_{\{i \leq k\}}}{\log(1+i)}$ and $u_i(\mathbf{y}) = 2^{y_i} - 1$. The function $\boldsymbol{u}$ can be seen as mapping the supervision y to utility scores for each individual item to rank. We may therefore refer to $\boldsymbol{u}$ as the *utility function* of the $(\boldsymbol{u}, \phi)$-GPPM r. The name we use for this family of measures comes from the positional models described in [10] in which one orders the documents according to the utility (i.e. the relevance) of a document w.r.t. the query. We call them "generalized", because the utility function is, in our case, only a means to transform the supervision so that a given performance measure can be

assimilated to a positional model. However, in a specific context, this transformation is not necessarily the relevance of a document with respect to the query as a user may define it. In particular, in information retrieval, the relevance of a document is usually defined independently of the other documents, while in the case of GPPMs, the utility score may depend on the relevance of the other documents. The Normalized DCG is a typical example of such a GPPM.

Table 1 summarizes the formulas for several, widely used GPPMs: the (Normalized) Discounted Cumulative Gain (N)DCG, the precision at rank K (Prec@k), recall at rank K (Rec@k) and Area Under the ROC Curve (AUC). All of these performance measure assume $\mathcal{Y} \subset \mathbb{R}_+^n$. We also provide the formula of Spearman's rank correlation coefficient (Spearman) as a GPPM to give an example where $\mathcal{Y} \not\subset \mathbb{R}^n$, but the set of total orders of $\{1, ..., n\}$ instead. For completeness, we also give the formula and the supervision space for the Expected Reciprocal Rank (ERR) and the Average Precision (AP). Note that for GPPMs, the utility function may not be unique, Table 1 only gives one possible formulation.

Table 1 Summary of common Performance Measures. The function b is equal to zero for all measures except for the AUC ($b(\mathbf{y}) = \frac{(\|\mathbf{y}\|_1 - 1)}{2(n - \|\mathbf{y}\|_1)}$) and for Spearman Rank Correlation Coefficient ($b(\mathbf{y}) = -\frac{3(n-1)}{(n+1)}$). The details of the calculations are given in Appendix B.1.

$\mathcal{Y}$	Name	Formula	$\phi(i)$	$u_i(\mathbf{y})$
$\mathbf{y} \in \{0..p\}^n$	`DCG@k`	$\sum_{i=1}^{k} \frac{2^{y_{\sigma(i)}} - 1}{\log(1+i)}$	$\frac{\mathbf{1}_{\{i \leq k\}}}{\log(1+i)}$	$2^{y_i} - 1$
	`NDCG@k`	$\frac{\texttt{DCG@k}(\mathbf{y}, \sigma)}{\max_{\sigma' \in \mathfrak{S}_n} \texttt{DCG}(\mathbf{y}, \sigma')}$	$\frac{\mathbf{1}_{\{i \leq k\}}}{\log(1+i)}$	$\frac{2^{y_i} - 1}{\max_{\sigma' \in \mathfrak{S}_n} \texttt{DCG@k}(\mathbf{y}, \sigma')}$
	`ERR`	$\sum_{i=1}^{n} \frac{R_i}{i} \prod_{k=1}^{i-1} (1 - R_k)$ with $R_i = \frac{2^{y_{\sigma(i)}} - 1}{2^p}$	$\times$	$\times$
$\mathbf{y} \in \{0,1\}^n$	`Prec@k`	$\sum_{i=1}^{k} \frac{y_{\sigma(i)}}{k}$	$\frac{1}{k}\mathbf{1}_{\{i \leq k\}}$	y_i
	`Rec@k`	$\sum_{i=1}^{k} \frac{y_{\sigma(i)}}{\|\mathbf{y}\|_1}$	$\mathbf{1}_{\{i \leq k\}}$	$\frac{y_i}{\|\mathbf{y}\|_1}$
	`AP`	$\frac{1}{\|\mathbf{y}\|_1} \sum_{i:y_i=1} \texttt{Prec@}\sigma^{-1}(i)$	$\times$	$\times$
	`AUC`	$\sum_{\substack{i:y_i=1 \\ j:y_j=0}} \frac{\mathbf{1}_{\{\sigma^{-1}(i) < \sigma^{-1}(j)\}}}{\|\mathbf{y}\|_1 (n - \|\mathbf{y}\|_1)}$	$n - i$	$\frac{y_i}{\|\mathbf{y}\|_1 (n - \|\mathbf{y}\|_1)}$
$y \in \mathfrak{S}_n$	`Spearman`	$1 - Z_n \sum_{i=1}^{n} (\sigma^{-1}(i) - y^{-1}(i))^2$ with $Z_n = \frac{6}{n(n^2\text{-}1)}$	$\frac{12(n-i)}{n(n^2-1)}$	$n - y^{-1}(i)$

Learning Objective and Surrogate Scoring Loss In the remaining of the paper, we will use many results due to Steinwart [28], and thus follow its basic assumptions that $\mathcal{Y}$ is a polish space (i.e. a separable completely metrizable space) and $\mathcal{X}$ is complete in the sense of [28, p. 3]. These assumptions are purely technical and allow to deal with ranking tasks in their full generality (note, in particular, that any open or closed subset or $\mathbb{R}^n$ is a polish space, as well as any finite set). Consider a probability measure P on $\mathcal{X} \times \mathcal{Y}$, which is unknown, and a ranking (or, equivalently, scoring) performance measure r. Given a sample drawn i.i.d. from P, the goal of learning to rank is to find a scoring function $\mathbf{f}$ with high ranking performance $\mathcal{R}(P, \mathbf{f})$ on P, defined by:

$$\mathcal{R}(\mathrm{P}, \mathbf{f}) = \int_{\mathcal{X}\times\mathcal{Y}} r\left(y, \mathbf{f}(x)\right) \mathrm{d}P(x, y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} r\left(y, \mathbf{f}(x)\right) \mathrm{d}P(y|x) \mathrm{d}P_{\mathcal{X}}(x)$$

where $P_{\mathcal{X}}$ is the marginal distribution of P over $\mathcal{X}$ and $P(.|.)$ is a regular conditional probability. As usual in learning, the performance measure we intend to maximize is neither continuous nor differentiable. The optimization of the empirical performance is thus intractable, and the common practice is to minimize a *surrogate scoring risk* as a substitute for directly optimizing ranking performance. This surrogate is chosen to ease the optimization of its empirical risk. A natural way to obtain computationally efficient algorithm is to consider as surrogate a continuous and differentiable function of the predicted scores. More generally, we define a *scoring loss* as a measurable function $\ell : \mathcal{Y} \times \mathbb{R}_+^n \to \mathbb{R}_+$. We use the convention that scoring losses, which are substitutes for the ranking/scoring performance, are minimized while the latter are maximized. This will avoid ambiguities about which function is the surrogate and which one is the target.

A major issue of the field of learning to rank is the design of surrogate scoring losses that are, in some sense, well-behaved with respect to the target ranking performance measure. The next subsection will define criteria that should be satisfied for a reasonable surrogate loss. But before going into more details and in order to give a concrete example of a family of losses that may be useful when the performance measure is a GPPM, we define the following family of *template losses*:

Definition 2.2 (Template Scoring Loss) *Let Γ be a subset of $\mathbb{R}_+^n$. A* template scoring loss *is a measurable function $\ell : \Gamma \times \mathbb{R}^n \to \mathbb{R}_+$. For any measurable function $\boldsymbol{u} : \mathcal{Y} \to \mathbb{R}_+^n$ with $\boldsymbol{u}(\mathcal{Y}) \subset \Gamma$, the $\boldsymbol{u}$-instance of ℓ, denoted $\ell^{\boldsymbol{u}}$, is defined by:*

$$\forall y \in \mathcal{Y}, \forall \mathbf{s} \in \mathbb{R}^n, \ell^{\boldsymbol{u}}\left(y, \mathbf{s}\right) = \ell\left(\boldsymbol{u}(y), \mathbf{s}\right)$$

A typical example of template loss is the squared loss defined by $\ell\left(\mathbf{v}, \mathbf{s}\right) = \sum_{i=1}^{n} (v_i - s_i)^2$ on $\Gamma = \mathbb{R}^n$, as proposed in [14]. Other examples of template losses will be given in Section 3.

Note that many surrogate losses have been proposed for learning to rank (see [25] for an exhaustive review), and many of them are actually not template losses. SVM^{map} [34], and many other instances of the structural SVM approach to ranking are good examples [23, 8]. Their advantage is to be designed for a specific performance measure, which may work better in practice when this performance measure is used for evaluation. On the other hand, template losses have the algorithmic advantage of providing an interface that can easily be specialized for a specific GPPM.

2.2 Calibration and Surrogate Regret Bounds

We now describe some natural properties that surrogate loss functions should satisfy. This subsection defines the notations, and briefly summarizes the definitions and results from [28] which we are the basis of our work. The notations defined in this section are used in the rest of the paper without further notice.

Calibration A natural property that a surrogate loss should satisfy is that if one achieves, in some way, to find a scoring function minimizing its associated risk, then the ranking performance of this ranking function should be optimal as well. More formally, consider any scoring function $\mathbf{f}$ and let us denote:

- $\mathcal{L}(P, \mathbf{f}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(y, \mathbf{f}(x)) \mathrm{d}P(y|x) \mathrm{d}P_{\mathcal{X}}(x)$ the *scoring risk* of $\mathbf{f}$;
- $\underline{\mathcal{L}}(P) = \inf_{\substack{\mathbf{f}:\mathcal{X}\to\mathbb{R}^n \\ \mathbf{f} \text{ measurable}}} \mathcal{L}(P, \mathbf{f})$ the optimal scoring risk;
- $\overline{\mathcal{R}}(P) = \sup_{\substack{\mathbf{f}:\mathcal{X}\to\mathbb{R}^n \\ \mathbf{f} \text{ measurable}}} \mathcal{R}(P, \mathbf{f})$ the optimal ranking performance.

Then, we want the following proposition to be true for any sequence $(\mathbf{f}_k)_{k\geq 0}$ of scoring functions:

$$\mathcal{L}(P, \mathbf{f}_k) \underset{k\to\infty}{\longrightarrow} \underline{\mathcal{L}}(P) \quad \Rightarrow \quad \mathcal{R}(P, \mathbf{f}_k) \underset{k\to\infty}{\longrightarrow} \overline{\mathcal{R}}(P) \tag{1}$$

Condition (1) is, in fact, equivalent to the notion of *calibration* (see [28, Definition 2.7]), which we describe now.

Let $\mathcal{D}$ denote the set of probability distributions over $\mathcal{Y}$, and let $\Delta \subset \mathcal{D}$. Following [28, Definition 2.6], we say that P is a distribution of type Δ if $P(.|x) \in \Delta$ for all x. Then, [28, Theorem 2.8] shows that (1) hold for any distribution of type Δ such that $\overline{\mathcal{R}}(P) < +\infty$ and $\underline{\mathcal{L}}(P) < +\infty$ if and only if ℓ is r-calibrated on Δ, according to the following definition:

Definition 2.3 (Calibration) *Let r be a ranking performance measure, ℓ a scoring loss and $\Delta \subset \mathcal{D}$ where $\mathcal{D}$ is the set of probability distributions over $\mathcal{Y}$.*

We say that ℓ is r-calibrated on Δ if for any $\varepsilon > 0$ and any $\triangle \in \Delta$, there exists $\delta > 0$ such that:

$$\forall \mathbf{s} \in \mathbb{R}^n, L(\triangle, \mathbf{s}) - \underline{L}(\triangle) < \delta \implies \overline{R}(\triangle) - R(\triangle, \mathbf{s}) < \varepsilon$$

where $(\triangle, \mathbf{s}) \mapsto L(\triangle, \mathbf{s})$ and $(\triangle, \mathbf{s}) \mapsto R(\triangle, \mathbf{s})$ are called respectively the inner risk *and* inner performance*, and the quantities $L(\triangle, \mathbf{s})$, $\underline{L}(\triangle)$, $R(\triangle, \mathbf{s})$ and $\overline{R}(\triangle)$ are respectively defined by:*

- $\forall \mathbf{s} \in \mathbb{R}^n, L(\triangle, \mathbf{s}) = \int_{\mathcal{Y}} \ell(y, \mathbf{s}) \mathrm{d}\triangle(y)$ *and* $\underline{L}(\triangle) = \inf_{\mathbf{s}\in\mathbb{R}^n} L(\triangle, \mathbf{s})$*;*
- $\forall \mathbf{s} \in \mathbb{R}^n, R(\triangle, \mathbf{s}) = \int_{\mathcal{Y}} r(y, \mathbf{s}) \mathrm{d}\triangle(y)$ *and* $\overline{R}(\triangle) = \sup_{\mathbf{s}\in\mathbb{R}^n} R(\triangle, \mathbf{s})$.

The definition of calibration allows us to study the implication (1), which considers risks and performance defined on the whole data distribution, to the study of the *inner risks* which are much easier to deal with since they are only functions of the distribution over the supervision space and a score vector. Thus, the inner risk and the inner performance are the essential quantities we investigate in this paper. The study of the calibration w.r.t. GPPMs of some surrogate losses will be treated in Section 3.

Remark 1 The criterion given by Equation 1 studies the convergence to performance of the best possible scoring function, even though reaching this function is practically unfeasible on a finite training set since we need to consider a restricted class of functions. Nonetheless, as discussed in [35] in the context of multiclass classification, the best possible performance can be achieved asymptotically as the number of examples grows to infinity, using the method of sieves or structural risk minimization, that is by progressively increasing the models complexity as the training set size increases.

Uniform Calibration and Surrogate Regret Bounds While the calibration gives us an asymptotic relation between the minimization of the surrogate loss and the maximization of the performance, it does not give us any information on how fast the *regret* of $\mathbf{f}_k$ in terms of performance defined by $\overline{\mathcal{R}}(P) - \mathcal{R}(P, \mathbf{f}_k)$ decreases to 0 when the *surrogate regret* of $\mathbf{f}_k$, defined by $\mathcal{L}(P, \mathbf{f}_k) - \underline{\mathcal{L}}(P)$, tends to 0. An answer to this question can be given by a *surrogate regret bound*, which is a function $\Upsilon : \mathbb{R}_+ \to \mathbb{R}_+$ with $\Upsilon(0) = 0$ and continuous in 0, such that, for any distribution P of type Δ satisfying $\overline{\mathcal{R}}(P) < +\infty$ and $\underline{\mathcal{L}}(P) < +\infty$, we have, for any scoring function $\mathbf{f}$:

$$\overline{\mathcal{R}}(P) - \mathcal{R}(P, \mathbf{f}) \leq \Upsilon\left(\mathcal{L}(P, \mathbf{f}) - \underline{\mathcal{L}}(P)\right)$$

[28, Theorems 2.13 and 2.17] show that the existence of such a surrogate regret bound is equivalent to a notion stronger than calibration called *uniform calibration* (see [28, Definition 2.15]):

Definition 2.4 (Uniform Calibration) *With the notations of Definition 2.3, we say that ℓ is uniformly r-calibrated on Δ if, for any $\varepsilon > 0$, there exists $\delta > 0$ such that for any $\triangle \in \Delta$ and any $\mathbf{s} \in \mathbb{R}^n$:*

$$L\left(\triangle, \mathbf{s}\right) - \underline{L}\left(\triangle\right) < \delta \implies \overline{R}\left(\triangle\right) - R\left(\triangle, \mathbf{s}\right) < \varepsilon$$

Some criteria to establish the uniform calibration of scoring losses w.r.t. GPPMs are provided in Section 4. Quantitative regret bounds for specific template scoring losses will then be given in Section 5.

3 Calibration of Order-Preserving Losses

In this section, we address the following question: which surrogate losses are calibrated w.r.t. GPPMs. This leads us to define the order-preserving property for surrogate losses. Since there is no reason to believe that these losses are calibrated only with respect to GPPMs, we address the question of whether they can be calibrated with two other popular performance measures, namely the ERR and the AP. In the remaining of the paper, we make extensive use of the notations of 2.3.

Notation We introduce now additional notations. For a ranking performance measure r, we denote $\mathcal{D}_r = \{\triangle \in \mathcal{D} | \forall \mathbf{s} \in \mathbb{R}^n, R\left(\triangle, \mathbf{s}\right) < +\infty\}$. Likewise, we define $\mathcal{D}_\ell = \{\triangle \in \mathcal{D} | \forall \mathbf{s} \in \mathbb{R}^n, L\left(\triangle, \mathbf{s}\right) < +\infty\}$ for a scoring loss ℓ, and denote $\mathcal{D}_{\ell,r}$ the intersection of $\mathcal{D}_r$ and $\mathcal{D}_\ell$. Finally, given let r be $(\boldsymbol{u}, \phi)$-GPPM and let $\triangle \in \mathcal{D}_r$. We denote $\boldsymbol{U}(\triangle) = \int_{\mathcal{Y}} \boldsymbol{u}(y) d\triangle(y)$ the expected value of $\boldsymbol{u}$. One may notice that $\mathcal{D}_r = \{\triangle | \|\boldsymbol{U}(\triangle)\|_\infty < +\infty\}$.

3.1 Order-Preserving Scoring Losses

As the starting point of our analysis, we first notice that by definition of a $(\boldsymbol{u}, \phi)$-GPPM, the function ϕ is a non-increasing function of the rank. Thus, for any given value of the supervision, the $(\boldsymbol{u}, \phi)$-GPPM is maximized by predicting items of higher utility values at better ranks by the rearrangement inequality[1]; and considering the additive structure of a $(\boldsymbol{u}, \phi)$-GPPM, the expected value of the $(\boldsymbol{u}, \phi)$-GPPM over a distribution $\triangle \in \mathcal{D}_r$, is maximized by ranking the items according to their expected utility values. More formally, for any $(\boldsymbol{u}, \phi)$-GPPM and any $\triangle \in \mathcal{D}_r$:

$$\arg\operatorname{sort}(\mathbf{s}) \subseteq \arg\operatorname{sort}(\boldsymbol{U}(\triangle)) \Rightarrow R(\triangle, \mathbf{s}) = \overline{R}(\triangle) \tag{2}$$

Moreover, the reverse implication holds when ϕ is strictly decreasing (i.e. $\phi(i) > \phi(i+1)$ for any $0 < i < n$).

This result was already noticed by [14], where the authors advocated for regression approaches for optimizing the DCG and in [26] where the authors studied a generalization of regression losses based on Bregman divergences (see Eq. 5 below). This result emphasizes on the fact that optimizing a GPPM is still a much weaker objective in general than regressing the utility values – preserving the ordering induced by the utility function is sufficient. Consequently, it is natural to look for surrogate losses for which the inner risk is minimized only by scores which order the items like $\boldsymbol{U}$: by the definition of calibration, any such loss is r-calibrated with any $(\boldsymbol{u}, \phi)$-GPPM r. This leads us to the following definition:

Definition 3.1 (Order-Preserving Loss) *Let $\boldsymbol{u} : \mathcal{Y} \to \mathbb{R}_+^n$ be a measurable function, ℓ be scoring loss and $\Delta \subset \mathcal{D}_\ell$. We say that the scoring loss $\ell : \mathcal{Y} \times \mathbb{R}^n \to \mathbb{R}$ is order-preserving w.r.t. $\boldsymbol{u}$ on Δ if, for any $\triangle \in \Delta$, we have:*

$$\underline{L}(\triangle) < \inf\left\{ L(\triangle, \mathbf{s}) \middle| \, \mathbf{s} \in \mathbb{R}^n, \arg\operatorname{sort}(\mathbf{s}) \not\subseteq \arg\operatorname{sort}(\boldsymbol{U}(\triangle)) \right\}$$

Moreover, a template scoring loss ℓ (see Definition 2.2) is called order-preserving *if it is order-preserving with respect to the identity function of $\mathbb{R}$ on $\mathcal{D}_\ell$.*

It is clear that in general, if a loss is order-preserving w.r.t. some function $\boldsymbol{u}$, then it is not be order-preserving w.r.t. another utility function $\boldsymbol{u}'$ unless their is a strong relationship between the two functions (e.g. they are equal up to a constant additive or multiplicative factor). As such, in order to obtain loss functions calibrated with any GPPM, template scoring losses are a natural choice. We provide here some examples of such losses, for which surrogate regret bounds are given in Section 5:

– Pointwise template scoring losses:

$$\forall \mathbf{v} \in \Gamma \subset \mathbb{R}^n, \mathbf{s} \in \mathbb{R}^n, \ell(\mathbf{v}, \mathbf{s}) = \sum_{i=1}^{n} \lambda(v_i, s_i) \,. \tag{3}$$

[1] The rearrangement inequality states that for any real numbers $x_1 \geq ... \geq x_n \geq 0$ and $y_1 \geq ... \geq y_n$, and for any permutation $\sigma \in \mathfrak{S}_n$, we have $x_1 y_{\sigma(1)} + ... + x_n y_{\sigma(n)} \leq x_1 y_1 + ... + x_n y_n$. (the dot product is maximized by pairing greater x_is with greater y_is). Moreover, if the x_is are strictly decreasing, then the equality holds if and only if $y_{\sigma(1)} \geq ... \geq y_{\sigma(n)}$.

As mentioned in Section 2.1, one may take $\Gamma = \mathbb{R}^n$ and $\lambda(v_i, s_i) = (v_i - s_i)^2$ as in [14]. This template loss is obviously order-preserving since the optimal value of the scores is precisely the expected value of $\mathbf{v}$ (and thus $\boldsymbol{U}(\triangle)$ when the template loss is instantiated).
We may also consider, given $\eta > 0$, the form $\lambda(v_i, s_i) = v_i\varphi(s_i)+(\eta-v_i)\varphi(-s_i)$, which is convex with respect to $\mathbf{s}$ for any $\mathbf{v}$ in $\Gamma = [0, \eta]^n$ if φ is convex. As we shall see in Section 5, this loss is order-preserving for many choices of φ, including the log-loss ($t \mapsto \log(1 + e^{-t})$), the exponential loss ($t \mapsto e^{-t}$) or differentiable versions of the Hinge loss. The log-loss proposed in [22] in the context of bipartite instance ranking for optimizing the AUC follows the same idea as the latter pointwise losses. The surrogate regret bounds proved in [17] in the same ranking framework than the one we consider here apply to pointwise losses of a similar form, although with a value of η that depends on the supervision at hand.

– Pairwise template scoring losses:

$$\ell(\mathbf{v}, \mathbf{s}) = \sum_{i<j} \lambda(v_i, v_j, s_i - s_j) \tag{4}$$

with $\Gamma = \mathbb{R}^n_+$. For example, taking $\lambda(v_i, v_j, s_i - s_j) = (s_i - s_j - v_i + v_j)^2$ also obviously leads to an order-preserving template loss. But we may also take $\lambda(v_i, v_j, s_i - s_j) = v_i\varphi(s_i - s_j) + v_j\varphi(s_j - s_i)$ (the latter being convex with respect to $\mathbf{s}$ for any $\mathbf{v}$ in Γ whenever φ is so). Such a choice leads to an order preserving template loss whenever φ is non-increasing, differentiable with $\phi'(0) < 0$ and the infimum (over $\mathbf{s} \in \mathbb{R}^n$) of $L(\triangle, .)$. is achieved for any $\triangle$ (see Remark 2 below). Pairwise losses are natural candidates for surrogate scoring losses because they share a natural invariant with the scoring performance measure (invariance by translation of the scores).

– Listwise scoring losses: as proposed in [26], we may consider a general form of surrogate losses defined by Bregman divergences. Let $\Psi : \Gamma \subset \mathbb{R}^n \to \mathbb{R}$ be a strictly convex, differentiable function on a set Γ and define the Bregman divergence associated to Ψ by $B_\psi(\mathbf{v}\|\mathbf{s}) = \psi(\mathbf{v}) - \psi(\mathbf{s}) - \langle\nabla\psi(\mathbf{s}), \mathbf{v} - \mathbf{s}\rangle$. Let $\mathbf{g} : \mathbb{R}^n \to \Gamma$ be invertible and such that for any $\mathbf{s} \in \mathbb{R}^n$, $s_i > s_j \Rightarrow \mathbf{g}_i(\mathbf{s}) > \mathbf{g}_j(\mathbf{s})$. Then, we can use the following template loss:

$$\ell(\mathbf{v}, \mathbf{s}) = B_\psi\big(\mathbf{v}\|\mathbf{g}(\mathbf{s})\big)\,, \tag{5}$$

which is an order-preserving template loss [26] as soon as the closure of $g(\mathbb{R}^n)$ contains Γ. This is due to a characterization of Bregman divergences due to [1] that the expectation of Bregman divergences (for a distribution over the left-hand argument) is uniquely minimized over the right-hand argument when the latter equals the expected value of the former.

Remark 2 (Pairwise Losses) The categorization of surrogate scoring losses into "pointwise", "pairwise" and "listwise" we use here is due to [7]. Note, however, that the pairwise template loss we consider in (4) with $\lambda(v_i, v_j, s_i - s_j) = v_i\varphi(s_i - s_j) + v_j\varphi(s_j - s_i)$ does *not* correspond to what is usually called the "pairwise comparison approach" to ranking and used in many algorithms, including the very popular RankBoost [19] and Ranking SVMs (see e.g. [21, 6]). Indeed, the latter can be written as $\ell(\mathbf{v}, \mathbf{s}) = \sum_{i,j} \mathbf{1}_{v_i > v_j}\varphi(s_i - s_j)$ (or some weighted versions of

this formula). This usual loss was shown to be non-calibrated with respect to the pairwise disagreement error for ranking by [18] for any convex φ in many general settings. This result shows that the loss is not order-preserving in general (because the pairwise disagreement error, when the supervision space is $\{0,1\}^n$ is minimized when we order the items according to their probability of belonging to class 1). On the other hand, with the form we propose in this paper, the inner risk for the $\boldsymbol{u}$-instance of ℓ can be written as $\mathcal{L}^{\boldsymbol{u}}(\triangle, \mathbf{s}) = \sum_{i=1}^{n} U_i(\triangle) \sum_{j \neq i} \varphi(s_i - s_j)$, which has the same form as the inner risk of the multiclass pairwise loss studied in [35] and is order-preserving under the appropriate assumptions [35, Theorem 5].

Remark 3 (A note on terminology) We use the qualifier *order-preserving* for scoring losses in a sense similar to what the author of [35] used in the context of multiclass classification. We may note that [26] use the term *order-preserving* to qualify a function $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ such that $s_i > s_j \Rightarrow g_i(\mathbf{s}) > g_j(\mathbf{s})$, which corresponds to a different notion than the one used here.

3.2 Calibration of Order-Preserving Losses

As already noticed, it follows from the definitions that if r is a $(\boldsymbol{u}, \phi)$-GPPM, then any loss order-preserving w.r.t. $\boldsymbol{u}$ is r-calibrated. The reverse implication is also true: given a measurable $\boldsymbol{u}$, then only a loss order-preserving w.r.t. $\boldsymbol{u}$ is calibrated with any $(\boldsymbol{u}, \phi)$-GPPM (that is, for any ϕ). This latter claim can be found with different definitions in [26, Lemma 3 and Lemma 4]. We summarize these results in the following theorem and give the proof for completeness:

Theorem 3.2 *Let $\boldsymbol{u} : \mathcal{Y} \to \mathbb{R}_+^n$ be a measurable function, ℓ be a scoring loss, and r be a $(\boldsymbol{u}, \phi)$-GPPM and $\Delta \subset \mathcal{D}_{\ell,r}$. The following claims are true:*

1. *If ℓ is order-preserving w.r.t. $\boldsymbol{u}$ on Δ, then ℓ is r-calibrated on Δ.*
2. *If ϕ is strictly decreasing and ℓ is r-calibrated on Δ, then ℓ is order-preserving w.r.t. $\boldsymbol{u}$ on Δ.*

Moreover, if ℓ is an order-preserving template loss, then the $\mathbf{u}$-instance of ℓ is r-calibrated on $\mathcal{D}_{\ell^{\boldsymbol{u}},r}$.

Proof The first claim and the remark on the template loss essentially follows from the definitions and from (2). For point 2, it is sufficient to show that for a given $\triangle \in \boldsymbol{\Delta}$, if $\arg\operatorname{sort}(\mathbf{s}) \not\subseteq \arg\operatorname{sort}(\boldsymbol{U}(\triangle))$, then there is a $c > 0$ such that $\overline{R}(\triangle) - R(\triangle, \mathbf{s}) \geq c$.

Notice that if $\arg\operatorname{sort}(\mathbf{s}) \not\subseteq \arg\operatorname{sort}(\boldsymbol{U}(\triangle))$, then there is a pair (i,j) with $U_i(\triangle) > U_j(\triangle)$ but $s_i \leq s_j$. Then, there is at least one permutation σ in $\arg\operatorname{sort}(\mathbf{s})$ with $\sigma^{-1}(i) > \sigma^{-1}(j)$. If $\tau_{ij} \in \mathfrak{S}_n$ is the transposition of i and j, we then have:

$$\overline{R}(\triangle) - R(\triangle, \sigma) = \underbrace{\overline{R}(\triangle) - R(\triangle, \tau_{ij} \circ \sigma)}_{\geq 0} + \underbrace{R(\triangle, \tau_{ij} \circ \sigma) - R(\triangle, \sigma)}_{= \left(U_i(\triangle) - U_j(\triangle)\right)\left(\phi(\sigma^{-1}(j)) - \phi(\sigma^{-1}(i))\right)} \tag{6}$$

Since $|\arg\operatorname{sort}(\mathbf{s})| \leq n!$, we have:

$$\overline{R}(\triangle) - R(\triangle, \mathbf{s}) \geq \min_{k<n} |\phi(k) - \phi(k+1)| \times \min_{i,j : U_i(\triangle) \neq U_j(\triangle)} \frac{|U_i(\triangle) - U_j(\triangle)|}{n!},$$

which proves the result. □

Note that obviously, if a scoring loss is order-preserving w.r.t. $\boldsymbol{u}$, then it is calibrated with any ranking performance measure such that $\arg\operatorname{sort}(\boldsymbol{U}(\triangle)) \subset \arg\min_\sigma \mathcal{R}(\triangle, \sigma)$. This gives us a full characterization of ranking performance measures with respect to which order-preserving losses are calibrated.

While the order-preserving property is all we need for the calibration w.r.t. to a GPPM, one may then ask if it can be of use for the two other widely used performance measures: the AP and the ERR. The question is important because, apart from the usual precision/recall at rank K and the (N)DCG, these are the most widely used measures in search engine evaluation. Unfortunately, the answer is negative:

Theorem 3.3 *Let $\mathcal{Y} = \{0,1\}^n$, $\boldsymbol{u} : \mathcal{Y} \to \mathbb{R}_+^n$ and ℓ an order-preserving loss w.r.t. $\boldsymbol{u}$ on $\mathcal{D}$. Then ℓ is not calibrated with the* `ERR` *and the* `AP`*.*

The proof of Theorem 3.3 can be found in the Appendix B.2. To the best of our knowledge, there is no existing study of the calibration of any surrogate scoring loss w.r.t. the ERR or the AP.

The theorem, in particular, implies that a regression approach is necessarily not calibrated with the ERR or the AP – whatever function of the relevance measure we are trying to regress. We do believe that the theorem does not cast lights on any weakness of the class of order-preserving (template) losses, but rather provides some strong evidence that these measures are difficult objective for learning and are that score-and-sort approaches are probably not suited for optimizing such performance measures.

4 Calibration and Uniform Calibration

We provided a characterization of surrogate losses calibrated with GPPM, as well as a characterization of the performance measures with respect to which these losses are calibrated. In this section, we are interested in investigating the stronger notion of *uniform calibration* which gives a non-asymptotic guarantee and implies the existence of a surrogate regret bound [28, Theorems 2.13 and 2.17]. Afterwards, in Section 5, we explicit regret bounds for some specific popular surrogate losses. In fact, we express conditions on the supervision space under which the uniform calibration w.r.t. a GPPM is equivalent to the simple calibration w.r.t. the same GPPM for learning to rank.

The equivalence between calibration and uniform calibration with respect to the classification error has already been proved in [2] for the binary case, and in [35, 29] in the multiclass case. Both studies concerned to margin losses, which are similar to the scoring losses we consider in the paper except that $\mathcal{Y}$ is restricted (in our notations) to be the canonical basis of $\mathbb{R}^n$ and $\boldsymbol{u}$ is the identity function. We extend these results to the case of GPPMs, but will not obtain an equivalence between calibration and uniform in general because of the more general form of scoring loss functions and the possible unboundedness of $\boldsymbol{u}$. Yet, we are able to present a number of special cases depending on the loss function and the considered set of distribution over the supervision space where such an equivalence holds.

The existence of a surrogate regret bound independent of the data distribution (even without an explicit statement of the bound) is critical tool in the proof of the consistency of structural risk minimization of the surrogate formulation in

[2, 35, 29]. Indeed, if one performs the empirical minimization of the surrogate risk in function classes that grow (sufficiently slowly) with the number of examples so that the surrogate risk tends to its infimum, the surrogate regret bound is sufficient to show that the sequence of surrogate risk minimizers tend to have maximal performance. The major tool used in [2, 27] for deriving explicit regret bounds also precisely corresponds to proving the uniform calibration. In our case, the criterion we develop for showing the equivalence between calibration and uniform calibration (Theorem 4.2) unfortunately does not lead to tight regret bounds. However, the following technical lemma, which we need to prove this criterion, will also appear crucial for the statement of explicit regret bounds.

Lemma 4.1 *Let r be a $(\boldsymbol{u}, \phi)$-GPPM, $\triangle \in \Delta \subset \mathcal{D}_r$, and $\nu \in \arg\operatorname{sort}(\boldsymbol{U}(\triangle))$. Then, for any $\sigma \in \mathfrak{S}_n$, there is a set $C_\sigma \subset \{1..n\}^2$ satisfying:*

1. $\forall (i,j) \neq (z,t) \in C_\sigma$, *we have* $\{i,j\} \cap \{z,t\} = \emptyset$,
2. $\forall (i,j) \in C_\sigma, U_i(\triangle) > U_j(\triangle)$ *and* $\sigma^{-1}(i) > \sigma^{-1}(j)$,
3. $\overline{R}(\triangle) - R(\triangle, \sigma) \leq \sum_{(i,j)\in C_\sigma} (U_i(\triangle) - U_j(\triangle))\big(\phi(\nu^{-1}(i)) - \phi(\nu^{-1}(j))\big)$.

Consequently, for any $\mathbf{s} \in \mathbb{R}^n$, if we take $C_{\mathbf{s}} = C_\sigma$ for some $\sigma \in \underset{\sigma' \in \arg\operatorname{sort}(\mathbf{s})}{\arg\min} R(\triangle, \sigma')$, we have $\forall (i,j) \in C_{\mathbf{s}}, U_i(\triangle) > U_j(\triangle), s_i \leq s_j$ and:

$$\overline{R}(\triangle) - R(\triangle, \mathbf{s}) \leq \sum_{(i,j)\in C_{\mathbf{s}}} \big(U_i(\triangle) - U_j(\triangle)\big)\big(\phi(\nu^{-1}(i)) - \phi(\nu^{-1}(j))\big).$$

Proof For the proof, we will use the notation $C^{\boldsymbol{u},\phi}_{\triangle,\sigma}(\nu)$ for the set C_σ to make all the dependencies clear. We prove the existence of $C^{\boldsymbol{u},\phi}_{\triangle,\sigma}(\nu)$ by induction on n, the number of items to rank. Let $n > 2$. It is easy to see that the result holds for $n \in \{1,2\}$. Assume that the same holds for any $k \leq n$.

Let $\nu \in \arg\operatorname{sort}(\boldsymbol{U}(\triangle))$ be an optimal ordering and $\sigma \in \mathfrak{S}_n$. The idea of the proof is build a permutation consisting in a set of non-overlapping transpositions $C^{\boldsymbol{u},\phi}_{\triangle,\sigma}(\nu)$ which inner-risk is worse or equal to the one of σ. For clarity, Figure 1 illustrates the exchanges that we now present. To simplify the proof, we make the following abuses of language: the "true rank of i" stands for $\nu^{-1}(i)$, i.e. for the rank of item i according to the optimal ν and the "predicted rank of i" stands for $\sigma^{-1}(i)$. Take $i = \nu(1)$ the true top-ranked item, and denote $d = \sigma^{-1}(i)$ its predicted rank. Now, consider the items in the top-d predicted ranks, and, in that set, denote j the one with worst true rank. Denote p its predicted rank, that is $p = \sigma^{-1}(j)$ with $j = \underset{q:\sigma^{-1}(q)\leq d}{\arg\max} \left\{\nu^{-1}(q)\right\}$. Notice that we have $U_i(\triangle) \geq U_j(\triangle)$ and $\nu^{-1}(j) \geq d$.

Since j is the item with the worst true rank among the top-d predicted items, we can only decrease the performance by exchanging it, in the predicted ranking, with the top-ranked item. In more formal terms, denoting $\tau_{wz} \in \mathfrak{S}_n$ the transposition of w and z, we thus have $R(\triangle, \sigma) \geq R(\triangle, \sigma \circ \tau_{1p})$ ($\sigma \circ \tau_{1p}$ is thus the ranking created by exchanging the items at (predicted) rank 1 and $p = \sigma^{-1}(j)$).

Likewise, since the true rank of j is greater than d and i is the true top-ranked element, we can, as well, only decrease performance if, in the predicted ranking, we exchange i with the item whose predicted rank is the true rank of j. More formally, we have:

$$R(\triangle, \sigma) \geq R\left(\triangle, \sigma \circ \tau_{1p} \circ \tau_{d\nu^{-1}(\sigma(p))}\right)$$

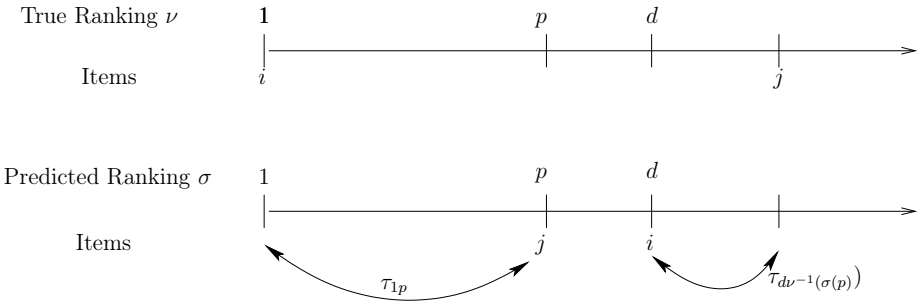

Fig. 1 Pictorial representation of $\sigma \circ \tau_{1p} \circ \tau_{d\nu^{-1}(\sigma(p))}$. The item j is put at first rank (i.e. the rank of i), and the item i isput at the true rank of j (i.e. the rank of j according to the true ranking ν. By the definition of j, this rank is greater than or equal to d.

In words, using $\sigma \circ \tau_{1p} \circ \tau_{d\nu^{-1}(\sigma(p))}$, we put i at the true rank of j (which is worse than the predicted rank of i), and put j at rank 1 (i.e. at the true rank of i). Even though we may have moved some other items, the important point is that the exchanges only decrease performance.

The interest of these exchanges is that i and j in $\sigma \circ \tau_{1p} \circ \tau_{d\nu^{-1}(\sigma(p))}$ have exchanged their position compared to the true optimal ranking ν. Consequently, we have:

$$\begin{aligned} R(\triangle, \nu) - R(\triangle, \sigma) &\leq R(\triangle, \nu) - R\left(\triangle, \sigma \circ \tau_{1p} \circ \tau_{d\nu^{-1}(\sigma(p))}\right) \\ &= (U_i(\triangle) - U_j(\triangle))(\phi(1) - \phi(\nu^{-1}(j))) \\ &\quad + \underbrace{\sum_{k \notin \{i,j\}} U_{\nu(k)}(\triangle)\,\phi(k)}_{=R'(\triangle,\nu')} - \underbrace{\sum_{k \notin \{i,j\}} U_{\sigma(k)}(\triangle)\,\phi(k)}_{=R'(\triangle,\sigma')} \end{aligned}$$

where we define r' as a $(\boldsymbol{u}', \phi')$-GPPM on lists of items of size $n-1$ or $n-2$ depending on i and j:

Case $i \neq j$ In that case, define r' as a $(\boldsymbol{u}', \phi')$-GPPM on lists of items of size $n-2$, such that $\boldsymbol{u}'$, ϕ', ν' and σ' are equal to $\boldsymbol{u}$, ϕ, ν and σ on indices different from i and j up to an appropriate re-indexing of the remaining $n-2$ items. Using the induction assumption, we can find a set $C^{\boldsymbol{u}',\phi'}_{\triangle,\sigma'}(\nu')$ satisfying the three conditions of the lemma, which we add to the pair (i,j) after re-indexing to build $C^{\boldsymbol{u},\phi}_{\triangle,\sigma}(\nu)$. Notice that for now, we do not exactly meet condition (ii) since we have $\forall (i,j) \in C^{\boldsymbol{u},\phi}_{\triangle,\sigma}(\nu), U_{\sigma(i)}(\triangle) \geq U_{\sigma(j)}(\triangle)$, while condition (ii) requires a strict inequality. However, if $U_{\sigma(i)}(\triangle) = U_{\sigma(j)}(\triangle)$ for some pair (i,j) in $C^{\boldsymbol{u},\phi}_{\triangle,\sigma}(\nu)$, then the pair has no influence on the bound and can thus simply be discarded.

Case i=j: In that case, define r' as a $(\boldsymbol{u}', \phi')$-GPPM on lists of items of size $n-1$. We then directly use the induction, ignoring the top-ranked element and considering the set of pairs on the remaining $n-1$ elements.

□

An important characteristic of the set C_σ in the lemma is condition 1, which ensures that the pairs (i,j) are independent (each index i appears in at most one

pair). This condition is critical in the derivation of explicit surrogate bounds of the next section. Another important technical feature of the bound is that is based on misordered pairs, and thus can be applied to any loss. In contrast, the bounds on DCG suboptimality used in [14] or [26] depend on how much (a function of) the score vector $\mathbf{s}$ approximates $\boldsymbol{U}(\triangle)$ – a bound which, consequently, can only be used for regression-like template losses.

We are now ready to give a new characterization of uniform calibration w.r.t. GPPMs. This characterization is easier to deal with than the initial definition of uniform calibration. We note here that it perfectly applies to losses of arbitrary structure, and thus also to non-template losses.

Theorem 4.2 *Let r be a $(\boldsymbol{u}, \phi)$-GPPM, ℓ be a scoring loss and $\Delta \subseteq \mathcal{D}_{\ell,r}$. For any $\varepsilon > 0$, and $i, j \in \{1, ..., n\}$, define:*

$$\Delta_{i,j}(\varepsilon) = \left\{ \triangle \in \Delta \middle| U_i(\triangle) - U_j(\triangle) \geq \varepsilon \right\} \tag{7}$$

and denote

$$\Omega_{i,j} = \{ \mathbf{s} \in \mathbb{R}^n \,|s_i \leq s_j \} \,. \tag{8}$$

Consider the two following statements:

(a) There is a function $\delta : \mathbb{R}_+ \to \mathbb{R}_+$ s.t. $\forall \varepsilon > 0, \delta(\varepsilon) > 0$ and:

$$\forall i \neq j, \forall \triangle \in \Delta_{i,j}(\varepsilon), \underline{L}(\triangle) + \delta(\varepsilon) \leq \inf_{\mathbf{s} \in \Omega_{i,j}} L(\triangle, \mathbf{s}) \,.$$

(b) ℓ is uniformly r-calibrated on Δ.

We have (a) $\Rightarrow$ (b) and, if $\forall 0 < i < n, \phi(i) > \phi(i+1)$ then (b) $\Rightarrow$ (a).

Proof We start with *(a) $\Rightarrow$ (b)*. Fix $\varepsilon > 0$, $\mathbf{s} \in \mathbb{R}^n$ and $\triangle \in \Delta$. From *(a)*, we know that if $L(\triangle, \mathbf{s}) - \underline{L}(\triangle) < \delta(\varepsilon)$ then for any i, j satisfying $\big(U_i(\triangle) - U_j(\triangle)\big)(s_i - s_j) \leq 0$, we have $|U_i(\triangle) - U_j(\triangle)| < \varepsilon$. By Lemma 4.1, we obtain $\overline{R}(\triangle) - R(\triangle, \mathbf{s}) < \frac{n}{2}\phi(1)\varepsilon$, since there are less than $n/2$ non-overlapping pairs of indexes $(i, j), i \neq j$ in $\{1, ..., n\}^n$ and $|\phi(i) - \phi(j)| \leq \phi(1)$ for any i, j. This bound being independent on $\triangle$, this proves the uniform calibration of ℓ w.r.t. r on Δ.

We now prove *(b) $\Rightarrow$ (a)* when $\forall 0 < i < n, \phi(i) > \phi(i+1)$ by contrapositive. Suppose *(a)* does not hold. Then, we can find $\varepsilon > 0$, a sequence $(i_k, j_k)_{k \geq 0}$ with $i_k \neq j_k$ for all k and a sequence $(\triangle_k)_{k \geq 0}$ with $\forall k, \triangle_k \in \Delta_{i_k, j_k}(\varepsilon)$ satisfying $\inf_{\mathbf{s} \in \Omega_{i_k, j_k}} L(\triangle_k, \mathbf{s}) - \underline{L}(\triangle_k) \underset{k \to +\infty}{\longrightarrow} 0$.

Thus, for any $\eta > 0$, we can find $i \neq j$, $\triangle \in \Delta_{i,j}(\varepsilon)$ and $s \in \mathbb{R}^n$ with $s_i \leq s_j$ such that $L(\triangle, \mathbf{s}) - \underline{L}(\triangle) < \eta$. However, if one considers the lower bound of (6), we obtain, for some permutation σ in $\arg\operatorname{sort}(\mathbf{s})$ s.t. $\sigma^{-1}(i) > \sigma^{-1}(j)$:

$$\overline{R}(\triangle) - R(\triangle, \sigma) \geq \epsilon \min_{j < n} |\phi(j) - \phi(j+1)|$$

Finally, since $|\arg\operatorname{sort}(\mathbf{s})| \leq n!$, we obtain $\overline{R}(\triangle) - R(\triangle, \mathbf{s}) \geq \epsilon \frac{\min_{j<n} |\phi(j) - \phi(j+1)|}{n!}$. This lower bound holds for any $i \neq j$, any $\triangle \in \Delta_{i,j}(\varepsilon)$ and any $\mathbf{s}$ such that $\big(U_i(\triangle_k) - U_j(\triangle_k)\big)(s_i - s_j) \leq 0$, and thus ℓ is not uniformly r-calibrated on Δ. □

Using this new characterization, we now address the problem of finding losses ℓ and sets of distributions Δ such that if ℓ is r-calibrated on Δ for some GPPM r, then condition *(a)* of Theorem 4.2 holds, implying the uniform calibration and the existence of the regret bound. The interest of the characterization of Theorem 4.2 is that in some cases, it is implied by large families of losses. Before going to some examples, we provide here the main corollary. Examples for more specific losses or supervision spaces are given in Corollaries 4.5 and in Appendix A.

Corollary 4.3 *Let r be a $(\boldsymbol{u}, \phi)$-GPPM, ℓ be a scoring loss, and $\Delta \subseteq \mathcal{D}_{\ell,r}$. Assume Δ can be given a topology such that :*

1. *Δ is compact;*
2. *the map $\begin{pmatrix} \Delta \to \mathbb{R}_+^n \\ \triangle \mapsto \boldsymbol{U}(\triangle) \end{pmatrix}$ is continuous;*
3. *$\forall i,j, \begin{pmatrix} \Delta \to \mathbb{R} \\ \triangle \mapsto \inf\limits_{\mathbf{s} \in \Omega_{i,j}} L(\triangle, \mathbf{s}) - \underline{L}(\triangle) \end{pmatrix}$ is continuous, with $\Omega_{i,j}$ defined by (8).*

Then, ℓ is r-calibrated on Δ if and only if it is uniformly r-calibrated on Δ.

Proof Since uniform calibration implies calibration, we only have to show the "only if" part.

First, we show using conditions 1 and 2 that for any $\varepsilon > 0$ and any i, j, the set $\Delta_{i,j}(\varepsilon)$ defined by (7) is compact. Since $\boldsymbol{U}$ is continuous on Δ and Δ is compact, $\boldsymbol{U}(\Delta)$ is a compact subset of $\mathbb{R}_+^n$. Therefore, $\boldsymbol{U}(\Delta)$ is bounded. Let $B = \sup_{\triangle \in \Delta} \|\boldsymbol{U}(\triangle)\|_\infty$ and consider now the function $h_{i,j}(\triangle) = U_i(\triangle) - U_j(\triangle)$. $h_{i,j}$ is continuous from Δ to $\mathbb{R}^n$ with Δ compact. Therefore, $h_{i,j}$ is a proper map, i.e. the preimage of any compact is compact (see e.g. [24, Lemma 2.14, p.45]). Thus, $\Delta_{i,j}(\varepsilon) = h_{i,j}^{-1}([\varepsilon, B])$ is compact in Δ.

We now go on to the proof of the result. Let $i \neq j$ and denote $g_{i,j} : \Delta \to \mathbb{R}$ the function defined in condition 3. Since ℓ is r-calibrated on Δ, we have $g_{i,j}(\triangle) > 0$ for any $\triangle \in \Delta_{i,j}(\varepsilon)$ as soon as $\varepsilon > 0$. Since $g_{i,j}$ is continuous and $\Delta_{i,j}(\varepsilon)$ is compact, $g_{i,j}(\Delta_{i,j}(\varepsilon))$ is a compact of $\mathbb{R}$ and the minimum is attained. Defining $\delta(\epsilon) = \min_{i \neq j} \min g_{i,j}(\Delta_{i,j}(\varepsilon))$, we thus have $\delta(\epsilon) > 0$. Using Theorem 4.2, it proves that ℓ is uniformly r-calibrated.

□

Corollary 4.3 gives conditions on the accepted form of supervision (conditions 1 and 2) and on the loss structure (condition 3) which are important to verify that r-calibration on Δ for a GPPM r implies uniform r-calibration on Δ. Conditions 1 and 2 are obviously satisfied when the supervision space is finite, and, as we shall see later, condition 3 is then automatically satisfied as well. Also, we may expect the same result to hold when we restrict $\boldsymbol{U}$ to be bounded. The cases of a finite supervision space is treated below. The more technical cases where the supervision space is infinite is more technical, and is detailed in Appendix A. We first remind the following result which will help us discuss these special cases:

Lemma 4.4 [35, Lemma 27] *Let $K > 0$, and let $\psi_k : \mathbb{R} \to \mathbb{R}_+, k = 1..K$ be K continuous functions. Let $\Omega \subseteq \mathbb{R}^n$, $\Omega \neq \emptyset$ and $\mathcal{Q}$ be a compact subset of $\mathbb{R}_+^K$. Then, the function $\underline{\Psi}$ defined as $\begin{pmatrix} \mathcal{Q} \to \mathbb{R}_+ \\ \mathbf{q} \mapsto \inf\limits_{\mathbf{s} \in \Omega} \sum\limits_{k=1}^{K} q_k \psi_k(\mathbf{s}) \end{pmatrix}$ is continuous.*

From now on, we suppose that *the supervision space $\mathcal{Y}$ is finite*. Then, $\Delta = \mathcal{D}$ can be identified with the $|\mathcal{Y}|$-simplex which is compact using its natural topology. Moreover, in that case, $\boldsymbol{U}$ is necessarily continuous with respect to this topology on Δ and thus conditions 1 and 2 of Corollary 4.3 are satisfied. Thus, the only question which remains is whether the class of loss functions we consider satisfies 3 – a question which is solved by Lemma 4.4. We can now give a full answer to the question of the uniform calibration w.r.t. a GPPM when the supervision space is finite:

Corollary 4.5 *Suppose that $\mathcal{Y}$ is finite. Let r be a $(\boldsymbol{u}, \phi)$-GPPM and ℓ a scoring loss such that $\ell(y, .)$ is continuous on $\mathbb{R}^n$ for any $y \in \mathcal{Y}$. Take $\Delta = \mathcal{D}$ (notice that $\mathcal{D} = \mathcal{D}_\ell = \mathcal{D}_r$). Then, the following claims are true:*

1. *ℓ is r-calibrated on Δ if and only if it is uniformly calibrated on Δ.*
2. *if ℓ is order-preserving w.r.t. $\boldsymbol{u}$ on Δ, it is uniformly r-calibrated on Δ.*
3. *If $\phi(i) > \phi(i+1)$ for all $0 < i < n$, then, ℓ is order-preserving w.r.t. $\boldsymbol{u}$ on Δ if and only if it is uniformly r-calibrated on Δ.*

Proof Since $\mathcal{Y} = \{y_1, ..., y_K\}$ is finite ($K = |\mathcal{Y}|$), we already showed that both conditions 1 and 2 of Corollary 4.3 are satisfied, identifying $\mathcal{D}$ with the K-simplex. Then, for any scoring loss, we have: $L(\triangle, \mathbf{s}) = \sum_{k=1}^{K} \triangle(\{y_k\})\ell(y_k, \mathbf{s})$ which satisfies condition 3 of Corollary 4.3 using Lemma 4.4. Thus, using Corollary 4.3, we know that for any $(\boldsymbol{u}, \phi)$-GPPM r, ℓ is r-calibrated if and only if it is uniformly r-calibrated, giving us the first claim of the corollary.

The second claim comes from the fact that an order-preserving loss is calibrated with any GPPM. The third claim comes from the fact that only order-preserving losses are calibrated w.r.t. $(\boldsymbol{u}, \phi)$-GPPM with $\phi(i) > \phi(i+1)$ for all $0 < i < n$, and the equivalence of r-calibration and uniform r-calibration when the supervision space is finite. □

This result shows that when the supervision space is finite, then any surrogate loss calibrated with respect to a GPPM has a regret bound. Thus, any loss calibrated with a GPPM has non-asymptotic guarantees. Since the exact form of the regret bound depends on the loss at hand, Corollary 4.5 is the stronger result we can obtain for arbitrary losses. We refer to Appendix A for a similar result concerning template losses in a special case where the supervision space is infinite. In the next section, we provide more quantitative surrogate regret bounds for specific template losses.

5 Surrogate Regret Bounds

The previous section deals with the existence of surrogate regret bounds by the study of uniform calibration. Now we propose to practically derive surrogate regret bounds for commonly-used template surrogate losses pointwise, pairwise or that can be written as a Bregman divergence. Like in classification the main idea is to find a convex lower bound of the surrogate regret as a function of the performance regret. However, contrary to classification, computing the calibration function like Steinwart [28] or the Ψ-transform like Bartlett [2] is actually unfeasible. If one tries to find the function δ of Theorem 4.2, the bound will be worse than the ones we reach in this section. Indeed, it doesn't use non-overlapping pairs of indexes.

In this section, we first present regret bounds for specific template losses in Table 2, then we describe three methods of proof for achieving theses bounds for either pointwise losses (3), Bregman divergences (5) or pairwise losses (4). Before starting the analysis, we introduce new notation for the (inner) risks associated to template losses. We first recall that given a template scoring loss $\ell : \Gamma \times \mathbb{R}^n \to \mathbb{R}_+$ with $\Gamma \subset \mathbb{R}_+^n$, its $\boldsymbol{u}$-instance is denoted by $\ell^{\boldsymbol{u}}$. Using a similar notation with a superscript $\boldsymbol{u}$, the *scoring risk* and *inner risk* of $\ell^{\boldsymbol{u}}$ are respectively denoted by:

- For any distribution P on $\mathcal{X} \times \mathcal{Y}$ and prediction function $\mathbf{f}$,

$$\mathcal{L}^{\boldsymbol{u}}(P, \mathbf{f}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell^{\boldsymbol{u}}(y, \mathbf{f}(x)) \mathrm{dP}(y, x) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(\boldsymbol{u}(y), \mathbf{f}(x)) \mathrm{dP}(y, x),$$

- For any $\triangle \in \mathcal{D}$, and $\mathbf{s} \in \mathbb{R}^n$,

$$L^{\boldsymbol{u}}(\triangle, \mathbf{s}) = \int_{\mathcal{Y}} \ell^{\boldsymbol{u}}(y, \mathbf{s}) \mathrm{d}\triangle(y) = \int_{\mathcal{Y}} \ell(\boldsymbol{u}(y), \mathbf{s}) \mathrm{d}\triangle(y).$$

Moreover, $\underline{\mathcal{L}^{\boldsymbol{u}}}(P)$ and $\underline{L^{\boldsymbol{u}}}(\triangle)$ refer to the respective optimal risks.

5.1 Regret Bounds for Common Surrogate Losses

We first give a summary of the different bounds obtained in the following of the section for both pointwise losses, Bregman divergences and pairwise losses, and then present the three methods used on the latter families of losses to achieve these bounds.

Given a $(\boldsymbol{u}, \phi)$-GPPM, for these families of surrogate scoring losses, we obtain the same regret bound up to a constant factor c, which intuitively correspond to the rescaling with respect to the surrogate loss.

$$\overline{\mathcal{R}}(P) - \mathcal{R}(P, \mathbf{f}) \leq c C_\phi(2) \sqrt{\mathcal{L}^{\boldsymbol{u}}(P, \mathbf{f}) - \underline{\mathcal{L}^{\boldsymbol{u}}}(P)} \tag{9}$$

with $C_\phi(p) = \left(\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \left(\phi(i) - \phi(n-i+1) \right)^p \right)^{\frac{1}{p}}$, for any positive integer p.

Table 2 details the different examples of Bregman divergences, pointwise losses and pairwise losses satisfying this surrogate regret bound (9) by giving the constant c. The methods for achieving such bounds are detailed in the following of the section: Theorem 5.2 for the pointwise losses, Theorem 5.3 for the Bregman divergences and Theorem 5.4 for pairwise losses. The proofs ensuring that the surrogate losses given in Table 2 satisfy the assumptions of the corresponding later theorems are given in Appendix B.

The differences in the constant factor c come from the fact that it represents a scaling factor between the surrogate loss and the expected utilities. Actually the magnitude of the loss may vary consequently from one to another. Furthermore, the bounds on the pointwise *Square Hinge* and the pointwise *Differentiable Hinge* depends respectively on t and α. Indeed, these parameters control the magnitude of the range within the optimal scores vary, so the scaling between the optimal scores and the expected utilities.

Notice that, $C_\phi(p)$ is generally strictly lower than $\|\phi\|_p$, thus, for the pointwise *Squared Error*, our approach allows us to obtain a slightly better bound than in

Table 2 Summary of surrogate regret bounds. Recalling that the v_i are upper-bounded, η can be chosen as $\eta > \max_i U_i(\triangle)$ and φ_α is a differentiable version of the *Hinge Loss*, where $\alpha \in (0; \frac{\eta}{2})$ is a parameter to choose: $\varphi_\alpha(x) = 0$ if $x \leq 0$, $\varphi_\alpha(x) = \frac{x^2}{2\alpha}$ if $x \in [0, \alpha]$, and $\varphi_\alpha(x) = x - \frac{\alpha}{2}$ otherwise.

Pointwise Losses (3): $\ell(\mathbf{v}, \mathbf{s}) = \sum_{i=1}^n \lambda(v_i, s_i)$		
Name	$\lambda(v_i, s_i)$	$\boldsymbol{c}$
Squared Error	$(v_i - s_i)^2$	$\sqrt{2}$
Logistic	$v_i \log(1 + e^{-s_i}) + (\eta - v_i)\log(1 + e^{s_i})$	$\sqrt{\eta}$
Exponential	$v_i e^{-s_i} + (\eta - v_i)e^{s_i}$	$\sqrt{\eta}$
Square Hinge	$v_i \max(0, t - s_i)^2 + (\eta - v_i)\max(0, s_i)^2$	$\frac{\sqrt{2\eta}}{t}$
Differentiable Hinge	$v_i\varphi_\alpha(1 - s_i) + (\eta - v_i)\varphi_\alpha(s_i)$	$4\sqrt{\frac{\eta}{\alpha}}$
Pairwise Losses (4): $\ell(\mathbf{v}, \mathbf{s}) = \sum_{i<j} \lambda(v_i, v_j, s_i - s_j)$		
Name	$\lambda(v_i, v_j, d_{ij})$	$\boldsymbol{c}$
Squared Error	$(v_i - v_j - d_{ij})^2$	1
Logistic	$v_i \log(1 + e^{-d_{ij}}) + v_j \log(1 + e^{d_{ij}})$	$2\sqrt{\|\boldsymbol{U}(\triangle)\|_\infty}$
Exponential	$v_i e^{-d_{ij}} + v_j e^{d_{ij}}$	$2\sqrt{\|\boldsymbol{U}(\triangle)\|_\infty}$
Bregman Divergence (5): $\ell(\mathbf{v}, \mathbf{s}) = B_\psi\big(\mathbf{v} \| \mathbf{g}(\mathbf{s})\big)$.		
	$\psi(.)$	$\boldsymbol{c}$
	μ-strongly convex (12)	$\frac{2}{\sqrt{\mu}}$

[14, Theorem 2]. The regret bound of the pointwise *Squared Error* is a crucial result since it helps to obtain the regret bounds on Bregman divergences. This explain why, applying the method of [26, Theorem 10] for Bregman divergences in our Theorem 5.3, we also reach a slightly better bound than them. Finally, for pairwise losses, to the best of our knowledge, no bound has already been proposed.

5.2 General Results to Derive Regret Bounds

Now, we aim at describing the methods allowing us to explicit the results of Table 2. The main argument is to combine lower bound of the surrogate regret with the upper bound on the performance regret given by the Lemma 4.1. We always use the same upper bound on the performance regret deduced from Lemma 4.1 so we explicit it here in the following lemma. Then, we will only work on the surrogate regret to obtain the bounds.

Lemma 5.1 *Let r be a $(\boldsymbol{u}, \phi)$-GPPM, $\triangle \in \Delta \subset \mathcal{D}_r$, and $C_{\mathbf{s}} \subset \{1..n\}^2$ given by Lemma 4.1. For $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$ if we denote*

$$C_\phi(p) = \left(\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \big(\phi(i) - \phi(n - i + 1)\big)^p \right)^{\frac{1}{p}} \tag{10}$$

then for any $\mathbf{s} \in \mathbb{R}^n$*, we have*

$$\overline{R}(\triangle) - R(\triangle, \mathbf{s}) \leq C_\phi(p) \left(\sum_{(i,j)\in C_\mathbf{s}} \big(U_i(\triangle) - U_j(\triangle)\big)^q \right)^{\frac{1}{q}}$$

The proof can be found in Appendix B.3.

We first treat the case of pointwise losses, then Bregman divergences and finally pairwise losses.

Specific Order-Preserving Pointwise Losses The case of the pointwise template loss (3) is clearly the easier. Indeed, in a pointwise loss, the dimensions are independent from each other. Lemma 4.1 breaks some dependencies into a set of non-overlapping pairs of items and allows us to link more easily the performance regret and the regret of a pointwise loss. Now we can consider only independent pairs of indexes to study the surrogate regret. First we define the optimal value of a surrogate loss ℓ w.r.t. a pair of indexes (i, j), and the near-optimal value given that the corresponding items are misordered as follows:

$$H_{ij}(\boldsymbol{u}, \triangle) = \underline{\Lambda^{u_i}}(\triangle) + \underline{\Lambda^{u_j}}(\triangle)$$
$$H^-_{ij}(\boldsymbol{u}, \triangle) = \inf_{\substack{\mathbf{s}_i, \mathbf{s}_j \in \mathbb{R} \\ (\mathbf{s}_i - \mathbf{s}_j)(U_i(\triangle) - U_j(\triangle)) \leq 0}} \Lambda^{u_i}(\triangle, s_i) + \Lambda^{u_j}(\triangle, s_j)$$

where $\Lambda^{u_i}(\triangle, s) = \int_\mathcal{Y} \lambda(u_i(\mathbf{y}), s) d\triangle(\mathbf{y})$ with λ defined in (3) and $\underline{\Lambda^{u_i}}(\triangle)$ is its infimum over s, i.e. $\underline{\Lambda^{u_i}}(\triangle) = \inf_{s\in\mathbb{R}} \Lambda^{u_i}(\triangle, s)$. In order to link H^-_{ij} and H_{ij} with the bound of the performance regret, we will use the assumption given by (11) below and verify that this assumption is met for natural instances of λ.

Theorem 5.2 *Let* r *be a* $(\boldsymbol{u}, \phi)$*-GPPM and* ℓ *a pointwise template loss. If there exists* $c > 0$ *and* $q \geq 1$ *such that for any* $\triangle \in \mathcal{D}_{r,\ell}$*,*

$$c^q (H^-_{ij}(\boldsymbol{u}, \triangle) - H_{ij}(\boldsymbol{u}, \triangle)) \geq |U_i(\triangle) - U_j(\triangle)|^q \tag{11}$$

Then, for any distribution P *on* $\mathcal{X} \times \mathcal{Y}$ *of type* $\mathcal{D}_{\ell,r}$ *such that* $\overline{\mathcal{R}}(P) < +\infty$ *and* $\underline{\mathcal{L}}(P) < \infty$*, we have, for any measurable scoring function* $\mathbf{f}$*:*

$$\overline{\mathcal{R}}(P) - \mathcal{R}(P, \mathbf{f}) \leq c\, C_\phi(p) \left(\mathcal{L}^{\boldsymbol{u}}(P, \mathbf{f}) - \underline{\mathcal{L}^{\boldsymbol{u}}}(P)\right)^{\frac{1}{q}}$$

where $p \geq 1$ *satisfies* $\frac{1}{p} + \frac{1}{q} = 1$ *and* $C_\phi(p)$ *is defined in* (10)*.*

Proof We consider $C_\mathbf{s} \subset \{1..n\}^2$ as defined in Lemma 4.1 with $\nu = id$. Indeed, considering the symmetry of the problem, we can consider the expected utilities are already ordered without any loss of generality. Consequently, for any $(i, j) \in C_\mathbf{s}$ we have $i < j$. Lemma 5.1 and (11) give

$$\begin{aligned}
\overline{R}(\triangle) - R(\triangle, \mathbf{s}) &\leq C_\phi(p) \left(\sum_{(i,j)\in C_\mathbf{s}} |U_i(\triangle) - U_j(\triangle)|^q \right)^{\frac{1}{q}} \\
&\leq C_\phi(p) \left(c^q \sum_{(i,j)\in C_\mathbf{s}} H^-_{ij}(\boldsymbol{u}, \triangle) - H_{ij}(\boldsymbol{u}, \triangle) \right)^{\frac{1}{q}}
\end{aligned}$$

Now, we denote $\overline{C_{\mathbf{s}}} = \{i \in \{1..n\} \mid \exists j/(i,j) \in C_{\mathbf{s}} \text{ or } (j,i) \in C_{\mathbf{s}}\}$ and $S_{\boldsymbol{u}}(C_{\mathbf{s}}) = \left\{\mathbf{s}' \in \mathbb{R}^n \mid \forall (i,j) \in C_{\mathbf{s}}, s'_i \leq s'_j\right\}$. Since $\mathbf{s} \in S_{\boldsymbol{u}}(C_{\mathbf{s}})$ then we have

$$\begin{aligned}
L^{\boldsymbol{u}}(\triangle, \mathbf{s}) - \underline{L^{\boldsymbol{u}}}(\triangle) &\geq \inf_{\mathbf{s}' \in S_{\boldsymbol{u}}(C_{\mathbf{s}})} L^{\boldsymbol{u}}(\triangle, \mathbf{s}') - \underline{L^{\boldsymbol{u}}}(\triangle) \\
&= \inf_{\mathbf{s}' \in S_{\boldsymbol{u}}(C_{\mathbf{s}})} \sum_{i=1}^{n} \left(\Lambda^{u_i}(\triangle, s'_i) - \underline{\Lambda^{u_i}}(\triangle)\right) \\
&= \inf_{\mathbf{s}' \in S_{\boldsymbol{u}}(C_{\mathbf{s}})} \left[\sum_{(i,j) \in C_{\mathbf{s}}} \left(\Lambda^{u_i}(\triangle, s'_i) + \Lambda^{u_j}(\triangle, s'_j)\right) + \sum_{k \notin \overline{C_{\mathbf{s}}}} \Lambda^{u_k}(\triangle, s'_k) \right] \\
&\quad - \sum_{i=1}^{n} \underline{\Lambda^{u_i}}(\triangle) \\
&= \sum_{(i,j) \in C_{\mathbf{s}}} \left[\inf_{s'_i \leq s'_j} \left(\Lambda^{u_i}(\triangle, s'_i) + \Lambda^{u_j}(\triangle, s'_j)\right) - \underline{\Lambda^{u_i}}(\triangle) - \underline{\Lambda^{u_j}}(\triangle) \right] \\
&\quad + \sum_{k \notin \overline{C_{\mathbf{s}}}} \underline{\Lambda^{u_k}}(\triangle) - \sum_{k \notin \overline{C_{\mathbf{s}}}} \underline{\Lambda^{u_k}}(\triangle) \\
&= \sum_{(i,j) \in C_{\mathbf{s}}} \left(H_{ij}^{-}(\boldsymbol{u}, \triangle) - H_{ij}(\boldsymbol{u}, \triangle)\right)
\end{aligned}$$

The inversion between the *inf* and the *sum* is possible because of the *independence* of the pairs in $C_{\mathbf{s}}$. Combining both inequalities gives the bound on the inner regret. Then the bound on the regret is deduced from using [28, Theorems 3.2 and 2.13]. □

Bregman Divergence Since we propose a bound on the pointwise Squared Error, we can apply a method similar to [26, Theorem 10] to obtain regret bounds on losses that derive from a Bregman divergence like those of (5). Moreover, pointwise losses *Logistic* and *Exponential* can be rewritten as Bregman divergences. This gives another way to obtain their corresponding bounds. Thus, we propose to use Lemma 5.1 to extend this theorem to the case of $(\boldsymbol{u}, \phi)$-GPPM's using almost the same conditions like strong convexity of the function ψ which generate the Bregman divergence. We say a function f is called μ-strongly convex if and only if for any x, y in the domain and $t \in [0,1]$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\mu}{2} t(1-t)\|x - y\|_2^2 \tag{12}$$

So, if ψ is μ-strongly convex, we have $B_\psi(u\|v) \geq \frac{\mu}{2}\|u - v\|_2^2$.

Theorem 5.3 *Let r be a $(\boldsymbol{u}, \phi)$-GPPM, $\psi : \Gamma_\psi \to \mathbb{R}$ in $\mathcal{C}^1$ a μ-strongly convex function and $g : \mathbb{R}^n \to \mathbb{R}^n$ an invertible map such that for any i, j we have $s_i < s_j \Rightarrow g_i(s) < g_j(s)$ such that $\Gamma_\psi = g(\mathbb{R}^n)$. For a scoring loss ℓ defined as (5), we have*

$$\overline{\mathcal{R}}(P) - \mathcal{R}(P, \mathbf{f}) \leq \frac{2C_\phi(2)}{\sqrt{\mu}} \sqrt{\mathcal{L}^{\boldsymbol{u}}(P, \mathbf{f}) - \underline{\mathcal{L}^{\boldsymbol{u}}}(P)}$$

Proof We consider $C_{\mathbf{s}} \subset \{1..n\}^2$ as defined in Lemma 4.1 with $\nu = id$ without loss of generality. We first start with a first bound on the suboptimality of $\ell^{\boldsymbol{u}}$ from strong convexity. Since $\underline{L^{\boldsymbol{u}}}(\triangle) = 0$, we have:

$$\begin{aligned} L^{\boldsymbol{u}}\left(\triangle, \mathbf{s}'\right) - \underline{L^{\boldsymbol{u}}}(\triangle) &= \int_{\mathcal{Y}} B_{\psi}(\mathbf{y} \| g(\mathbf{s}')) \mathrm{d}\triangle(\mathbf{y}) \\ &\geq \frac{\mu}{2} \int_{\mathcal{Y}} \|\boldsymbol{u}(\mathbf{y}) - g(\mathbf{s}')\|_2^2 d\triangle(\mathbf{y}) \\ &\geq \frac{\mu}{2} \|\boldsymbol{U}(\triangle) - g(\mathbf{s}')\|_2^2 \end{aligned}$$

The first inequality comes from strong convexity of ψ, while the second comes from the convexity of the squared 2-norm. Now, let us denote
$S_{\boldsymbol{u}}(C_{\mathbf{s}}) = \{\mathbf{s}' \in \mathbb{R}^n \,|\, \forall (i,j) \in C_{\mathbf{s}}, s_i \leq s_j\}$. Since $\mathbf{s} \in S_{\boldsymbol{u}}(C_{\mathbf{s}})$ and using the above inequality, we obtain

$$\begin{aligned} L^{\boldsymbol{u}}(\triangle, \mathbf{s}) - \underline{L^{\boldsymbol{u}}}(\triangle) &\geq \inf_{\mathbf{s}' \in S_{\boldsymbol{u}}(C_{\mathbf{s}})} L^{\boldsymbol{u}}\left(\triangle, \mathbf{s}'\right) - \underline{L^{\boldsymbol{u}}}(\triangle) \\ &\geq \frac{\mu}{2} \inf_{\mathbf{s}' \in S_{\boldsymbol{u}}(C_{\mathbf{s}})} \|\boldsymbol{U}(\triangle) - g(\mathbf{s}')\|_2^2 \end{aligned}$$

which is actually equals to the *Squared Error* regret taken in $g(\mathbf{s}')$. Then, combine with the regret bound on the *Squared Error* (see Table 2) to obtain the bound. □

Specific Order-Preserving Pairwise Losses In this section, we study the popular family of pairwise losses (see (4)) through two sub-families. We propose the first one to overcome the non-consistency of the classic pairwise hinge loss cast in light by [18]. The second one is just a mean squared error on the pairs of indexes.

Pairwise surrogate losses integrate complex correlations between the different dimensions of the predicted vector of score when optimizing. This is why it's not immediate to benefit from the independence given by the bound of Lemma 4.1. For pairwise surrogate losses, the main idea of the method is to treat them as pointwise losses on pairs of items with some additional constraints. Then, we compare the optima of the loss with and without the constraints.

The notations Λ^{u_i,u_j} and $\underline{\Lambda^{u_i,u_j}}$ are defined similarly to the ones of pointwise losses. We denote the following set of constraints which impose a solution equivalent to a score for each item to rank.

$$D = \{d \in \mathbb{R}^n \times \mathbb{R}^n / \forall i,j,k \in \{1..n\}, d_{ij} = d_{ik} + d_{kj}\}$$

With this set of constraints D, we can reduce the conditions on the pairwise surrogate loss in the following lemma to a condition on the Bayes risk and a pointwise condition w.r.t. the pairs (i,j) of items.

Theorem 5.4 *Let r be a $(\boldsymbol{u}, \phi)$-GPPM and ℓ a template pairwise scoring loss as described in* (4)*. For any $\triangle \in \mathcal{D}_{r,\ell}$ and $\mathbf{s} \in \mathbb{R}^n$, if $\ell^{\boldsymbol{u}}$ satisfies:*

1. $\underline{L^{\boldsymbol{u}}}(\triangle) = \inf_{\mathbf{d} \in D} \sum_{i<j} \Lambda^{u_i,u_j}(\triangle, d_{ij}) = \inf_{\mathbf{d} \in \mathbb{R}^n \times \mathbb{R}^n} \sum_{i<j} \Lambda^{u_i,u_j}(\triangle, d_{ij})$
2. *There exist $c > 0$ and $q \geq 1$ such that*

$$\inf_{d_{ij} \leq 0} \Lambda^{u_i,u_j}(\triangle, d_{ij}) - \inf_{d_{ij} \in \mathbb{R}} \Lambda^{u_i,u_j}(\triangle, d_{ij}) \geq \frac{1}{c^q} |U_i(\triangle) - U_j(\triangle)|^q$$

Then, for any distribution P on $\mathcal{X} \times \mathcal{Y}$ of type $\mathcal{D}_{\ell,r}$ such that $\overline{\mathcal{R}}(P) < +\infty$ and $\underline{\mathcal{L}}(P) < \infty$, we have, for any scoring function $\mathbf{f}$:

$$\overline{\mathcal{R}}(P) - \mathcal{R}(P, \mathbf{f}) \leq c\, C_\phi(p) \left(\mathcal{L}^{\boldsymbol{u}}(P, \mathbf{f}) - \underline{\mathcal{L}^{\boldsymbol{u}}}(P)\right)^{\frac{1}{q}} \tag{13}$$

where $p \geq 1$ satisfies $\frac{1}{p} + \frac{1}{q} = 1$ and $C_\phi(p)$ is defined in (10).

Proof For $C_{\mathbf{s}}$ defined as in Lemma 4.1, we denote

$$S_{\boldsymbol{u}}(C_{\mathbf{s}}) = \{\mathbf{s}' \in \mathbb{R}^n \,|\, \forall (i,j) \in C_{\mathbf{s}}, s_i \leq s_j\} \ ,$$

and

$$\Gamma(C_{\mathbf{s}}) = \{d \in \mathbb{R}^n \times \mathbb{R}^n | \forall (i,j) \in C_{\mathbf{s}}, d_{ij} \leq 0\} \ .$$

Since $\mathbf{s} \in S_{\boldsymbol{u}}(C_{\mathbf{s}})$ then we have

$$\begin{aligned}
L^{\boldsymbol{u}}(\triangle, \mathbf{s}) - \underline{L^{\boldsymbol{u}}}(\triangle) &\geq \inf_{\mathbf{s}' \in S_{\boldsymbol{u}}(C_{\mathbf{s}})} L^{\boldsymbol{u}}(\triangle, \mathbf{s}') - \underline{L^{\boldsymbol{u}}}(\triangle) \\
&\geq \inf_{\mathbf{d} \in \Gamma(C_{\mathbf{s}})} \sum_{i<j} \Lambda^{u_i,u_j}(\triangle, d_{ij}) - \inf_{\mathbf{d} \in \mathbb{R}^n \times \mathbb{R}^n} \sum_{i<j} \Lambda^{u_i,u_j}(\triangle, d_{ij}) \\
&\geq \inf_{\mathbf{d} \in \Gamma(C_{\mathbf{s}})} \sum_{\substack{i<j \\ (i,j) \in C}} \Lambda^{u_i,u_j}(\triangle, d_{ij}) + \sum_{\substack{i<j \\ (i,j) \notin C_{\mathbf{s}}}} \Lambda^{u_i,u_j}(\triangle, d_{ij}) \\
&\quad - \inf_{\mathbf{d} \in \mathbb{R}^n \times \mathbb{R}^n} \sum_{i<j} \Lambda^{u_i,u_j}(\triangle, d_{ij}) \\
&\geq \sum_{\substack{i<j \\ (i,j) \in C_{\mathbf{s}}}} \inf_{d_{ij} \leq 0} \Lambda^{u_i,u_j}(\triangle, d_{ij}) - \inf_{d_{ij} \in \mathbb{R}} \Lambda^{u_i,u_j}(\triangle, d_{ij}) \\
&\geq \frac{1}{c^q} \sum_{\substack{i<j \\ (i,j) \in C_{\mathbf{s}}}} |U_i(\triangle) - U_j(\triangle)|^q
\end{aligned}$$

Then, just apply Lemma 5.1 in the same way as in the proof of theorem 5.2 to plug this inequality to the performance inner regret to obtain the bound on the inner regret. The bound on the regret is deduced using [28, Theorems 3.2 and 2.13]. □

6 Discussion and Related Work

In this section, we discuss the most closely related works, and then summarize our results and discuss some of their practical implications.

Surrogate Regret Bounds for Learning to Rank The calibration and uniform calibration have been extensively studied in (cost-sensitive) binary classification (see e.g. [2, 35, 28, 27]) and multiclass classification [35, 29]. In the context of learning to rank, the calibration of surrogate losses in learning to rank has been previously studied by [14, 18, 26]. The authors of [14] proved the calibration of some variants of regression losses based on the mean squared error with respect to the DCG, and proved the first surrogate regret bound for ranking. In [26], the authors generalize this work to obtain the calibration of losses based on Bregman divergences (which include the squared error loss) with respect to the (N)DCG, and provide surrogate

regret bounds for this class of surrogate losses. In this paper, we extend the work of [26] in several ways. First, we consider a wider class of ranking performance measures, the GPPMs, essentially by noticing that it is not necessary to restrict the supervision to relevance judgments. Second, we consider a much larger class of surrogate losses (the order-preserving ones), which, in particular, are not constrained to have a unique minimizer. Relaxing these two assumptions, we obtain a new and general result on the existence of surrogate regret bounds for any loss calibrated with respect to a GPPM when the supervision space is finite, through the equivalence of calibration and uniform calibration for GPPMs (Corollary 4.3). Furthermore, our deeper study of the performance measures (Lemma 4.1) allow us to prove both slightly better regret bounds than [35] and [26] for the mean squared regression and for the Bregman divergences, as well as new regret bounds for other forms of surrogate losses such as pairwise losses or pointwise losses that do not have a unique minimizer (Section 5). While all these works studied the DCG, [17] proved regret bounds for pointwise losses for the special case of the AUC metric. The pointwise losses they consider are similar to the one we consider in Section 5 (the difference is that in their work, the value of η in these losses are not constants). While our proof technique could be adapted to their specific loss, the bounds we prove are more general since they apply to a larger variety of losses and different performance measures.

We may note here that surrogate regret bounds have also been studied in another context of learning to rank, namely instance ranking [11, 22]. Instance ranking, from which bipartite ranking is the best-known example (the case with binary relevance judgments) is a framework where the prediction task is to order a single set (the sample space itself), and learning is carried out based on an i.i.d. sample from this set. In contrast, in the task we consider here, the goal is to predict the ordering of a finite set for each instance and learning is carried out using an i.i.d. sample of such instances with a supervision that indicates how to rank the finite set given this instance. The evaluation performances for instance ranking are usually the Area Under the ROC Curve, or more generally linear rank statistics [12], which are similar in nature to what we call GPPMs. However, since the underlying sampling assumptions are different in instance ranking and in the framework we consider here, all the notions of inner risks are different, and the analyses carried out in one framework do not apply to the other framework.

Fitting Utility Values When the supervision takes the form of relevance scores on a discrete scale (as usually in search engine applications), it may be natural to simply try to fit them, for instance using classification or ordinal regression approaches. In the presence of noise however, our results show that one should not try to predict the value of the label, but rather its corresponding utility. More precisely, one should learn to rank according to the expected value of the utility; fitting the expected value of the utilities, for instance by minimizing the squared error, leads to a calibrated formulation, but it is only a special case of what one can do: in general, applying any order-preserving template loss is valid. Considering that many performance measures are GPPMs – for instance the (N)DCG, the precision-at-k, the recall-at-k, the AUC, or Spearman's rank correlation coefficient (see Table 1) – our result allow us to provide template calibrated surrogate losses that can be easily instantiated for each of these measures (Sections 2 & 3).

Another important result we obtain in the paper is the non-calibration of any surrogate loss that tries to reproduce the order given by the relevance judgments for the AP and the ERR (Theorem 3.3). The important result is that the non-calibration holds for *any* utility function that one can associate to these metric. Despite the importance of these measures in search engine evaluations, our result thus proves that many common surrogate losses used in learning to rank algorithm are not AP- or ERR-calibrated.

Consequently, the exact form of the supervision we have for the problem at hand – which may be relevance judgments, a preference relation, or total orders – does not dictate the kind of algorithm we should use. Spearman's correlation coefficient (see Table 1), which considers total orders as supervision, is actually a GPPM, and thus any template order-preserving loss can be calibrated with respect to it. This contrast with the case of the ERR or the AP, with respect to which no order-preserving is calibrated even though these performance measures consider real-valued relevance judgments for their supervision.

Pairwise Losses As already mentioned in Remark 2, a traditional approach to learning to rank is to use pairwise-comparison-based losses, as in Ranking SVMs or RankBoost [21, 19, 6]. To take a concrete example, consider the case when the supervision is a vector of relevance judgments. Then, the idea of pairwise-comparison -based losses is to take a loss of the form $\ell(\mathbf{v}, \mathbf{s}) = \sum_{i,j} \mathbf{1}_{\{v_i > v_j\}} \varphi(s_i - s_j)$ ($\mathbf{v}$ here takes the place of the supervision, or any monotonic transform of it). The motivation of these approaches is that only the relative ordering between any two items does matter for ranking, and thus it is somewhat natural to only consider the relative ordering given by the supervision for learning. However, such losses are not order-preserving when φ is convex (see Remark 2; this result is actually a direct consequence of the non-calibration result of [18]), and they are consequently not calibrated with respect to any GPPMs. This is why in this work we propose an alternative formulation $\ell(\mathbf{v}, \mathbf{s}) = \sum_{i<j} (v_i \varphi(s_i - s_j) + v_j \varphi(s_j - s_i))$, which is convex when the values of v_i are non-negative and φ is convex, and which, as we show in Section 5, is also order-preserving. Consequently, this alternative formulation provides a template loss whose instances are calibrated with respect to any GPPM. Notice that from a computational perspective, the two losses (the initial formulation and the alternative that we propose here) are comparable, and thus we strongly encourage to consider the alternative formulation in practice.

Limitations of Scoring Approaches for Ranking ? The difficulty of designing (convex) surrogate formulation for the score-and-sort approach to ranking has previously been addressed in [18], where the authors show that a number of existing surrogate losses are non calibrated with respect to the *pairwise disagreement*, a performance measure used when the supervision contains arbitrary pairwise preferences and which counts the number of pairs of items for which the predicted ordering does not match the supervision. The authors of [18] also conjecture that no convex loss of the scores can be calibrated with respect to the pairwise disagreement. In this work, we prove additional results concerning the possible limitations of scoring approaches: no order-preserving loss can be calibrated with respect to the AP or the ERR in general (Theorem 3.3). While this suggest that some approaches other than scoring may be useful for these evaluation measures, it also

gives new insights on the intrinsic limitations of scoring approaches (in particular regression approaches) for information retrieval.

7 Conclusion

The calibration, uniform calibration and surrogate regret bounds are crucial tools to assess the quality of surrogate losses. We proposed an in-depth study of the calibration of order-preserving losses with respect to GPPMs.

A large body of work remains to be done in learning to rank. As the authors of [18] pointed out, learning from pairwise preferences is still an open issue without making strong assumptions on the preference relations that we may have as supervision. More closely to our work, designing losses with better regret bounds for GPPMs with a cutoff (i.e. $\phi(i) = 0$ for $i > k$ and $k << n$) as in [14], but without any strong prior knowledge on which items should be ranked first and keeping easy-to-optimize surrogate losses, remains critical in many applications and mostly an open problem.

Acknowledgements This work was partially funded by the French DGA, as well as the French Government and Région île de France through the FUI project OpenWay III. The authors thank the anonymous reviewers for their helpful comments and suggestions.

References

1. Banerjee, A., Guo, X., Wang, H.: On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory* **51**(7), 2664–2669 (2005)
2. Bartlett, P., Jordan, M.: Convexity, classification, and risk bounds. *Journal of the American Statistical Society* **101**(473), 138–156 (2006)
3. Buffoni, D., Calauzènes, C., Gallinari, P., Usunier, N.: Learning scoring functions with order-preserving losses and standardized supervision. *In: Proceedings of the International Conference on Machine Learning*, pp. 825–832 (2011)
4. Burges, C.J.C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.N.: Learning to rank using gradient descent. *In: Proceedings of the International Conference on Machine Learning*, pp 89–96 (2005)
5. Cambazoglu, B.B., Zaragoza, H., Chapelle, O., Chen, J., Liao, C., Zheng, Z., Degenhardt, J.: Early exit optimizations for additive machine learned ranking systems. *In: Proceedings of the ACM International Conference on Web Search and Data Mining*, pp. 411–420 (2010)
6. Cao, Y., Xu, J., Liu, T.Y., Li, H., Huang, Y., Hon, H.W.: Adapting ranking SVM to document retrieval. *In: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 186–193 (2006)
7. Cao, Z., Liu, T.Y.: Learning to rank: From pairwise approach to listwise approach. *In: Proceedings of the International Conference on Machine Learning*, pp. 129–136 (2007)
8. Chakrabarti, S., Khanna, R., Sawant, U., Bhattacharyya, C.: Structured learning for non-smooth ranking losses. *In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 88–96 (2008)

9. Chapelle, O., Chang, Y.: Yahoo! learning to rank challenge overview. *Journal of Machine Learning Research* **14**, 1–24 (2011)
10. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. *In: Proceeding of the ACM Conference on Information and Knowledge Management*, pp. 621–630 (2009)
11. Clemençon, S., Lugosi, G., Vayatis, N.: Ranking and scoring using empirical risk minimization. *In: Proceedings of the Conference on Learning Theory*, pp. 783–800 (2005)
12. Clémençon, S., Vayatis, N.: Ranking the best instances. *Journal of Machine Learning Research* **8**, 2671–2699 (2007)
13. Cohen, W.W., Schapire, R.E., Singer, Y.: Learning to order things. *In: Proceedings of Advances in Neural Information Processing Systems*, pp 243–270, (1997)
14. Cossock, D., Zhang, T.: Statistical analysis of bayes optimal subset ranking. *IEEE Transactions on Information Theory* **54**(11), 5140–5154 (2008)
15. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* **2**, 265–292 (2002)
16. Dekel, O., Manning, C.D., Singer, Y.: Log-linear models for label ranking. *In: Proceedings of Advances in Neural Information Processing Systems* (2003)
17. Dembczynski, K., Kotlowski, W., Huellermeier, E.: Consistent multilabel ranking through univariate losses. *In: Proceedings of the International Conference on Machine Learning*, pp. 1319–1326 (2012)
18. Duchi, J., Mackey, L.W., Jordan, M.I.: On the consistency of ranking algorithms. *In: Proceedings of the International Conference on Machine Learning*, pp. 327–334 (2010)
19. Freund, Y., Iyer, R., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research* **4**, 933–969 (2003)
20. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* **20**, 422–446 (2002)
21. Joachims, T.: Optimizing search engines using clickthrough data. *In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 133–142, (2002)
22. Kotlowski, W., Dembczynski, K., Huellermeier, E.: Bipartite ranking through minimization of univariate loss. *In: Proceedings of the International Conference on Machine Learning*, pp. 1113–1120 (2011)
23. Le, Q.V., Smola, A.J.: Direct optimization of ranking measures. Technical Report, NICTA (2007)
24. Lee, J.: *Introduction to smooth manifolds.* Graduate texts in mathematics. Springer (2003)
25. Liu, T.Y.: Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval* **3**, 225–331 (2009)
26. Ravikumar, P.D., Tewari, A., Yang, E.: On ndcg consistency of listwise ranking methods. *Journal of Machine Learning Research - Proceedings Track* **15**, 618–626 (2011)
27. Scott, C.: Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. *In: Proceedings of the International Conference in Machine Learning*, pp 153–160 (2011)

28. Steinwart, I.: How to compare different loss functions and their risks. *Constructive Approximation* **26**(2), 225–287 (2007)
29. Tewari, A., Bartlett, P.: On the consistency of multiclass classification methods. *Journal of Machine Learning Research* **8**, 1007–1025 (2007)
30. Vapnik, Vladimir N.: *Statistical learning theory.* John Wiley & Sons, (1998)
31. Vembu, S., Gärtner, T.: Label ranking algorithms: A survey. *Preference Learning.* Springer,(Jan 2009) pp. 1530–1537 (2009)
32. Voorhees, E., Harman, D.: *TREC: experiment and evaluation in information retrieval.* Digital libraries and electronic publishing. MIT Press (2005)
33. Weston, J., Watkins, C.: Support vector machines for multi-class pattern recognition. *In: Proceedings of the European Symposium On Artificial Neural Networks*, pp. 219–224 (1999)
34. Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A support vector method for optimizing average precision. *In: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 271–278 (2007)
35. Zhang, T.: Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research* **5**, 1225–1251 (2004)

A Uniform Calibration on Infinite Supervision Space

In this appendix, we extend the results regarding the uniform calibration with respect to GPPMs to more general settings than the one studied in Section 4.

While the case of a finite supervision space has been solved, we can anticipate that the case where $\mathcal{Y}$ is infinite (e.g. $\mathcal{Y} = \mathbb{R}^n$) is more problematic: it is, indeed, much more difficult to define a topology on $\mathcal{D}$, in particular a topology that makes $\boldsymbol{\Delta}$ compact. We may, at this point, believe that the compactness of $\boldsymbol{\Delta}$ is probably too strong an assumption, and that we may have equivalence between calibration and uniform calibration with more general assumptions. However, considering the regret bounds we obtain in Section 5, we will see that assumption 1 of Corollary 4.3 is probably quite difficult to relax in general.

To deal with the case where $\mathcal{Y}$ is infinite, let us fix a $(\boldsymbol{u}, \phi)$-GPPM r and let us restrict our attention to scoring losses such that $g_{i,j}(\triangle) = \inf_{\mathbf{s} \in \Omega_{i,j}} L(\triangle, \mathbf{s}) - \underline{L}(\triangle)$ only depends on the expected value $\boldsymbol{U}(\triangle)$. This can be checked for specific losses, like the squared loss described in (3) with $\lambda(v_i, s_i) = (v_i - s_i)^2$ and more generally for the losses based on Bregman divergences like (5) because the only non-linear term in $\mathbf{v}$ cancels out. To simplify the discussion and directly use the continuity result of Lemma 4.4, we will restrict to the template losses of the following form:

$$\ell(\mathbf{v}, \mathbf{s}) = \sum_{i=k}^{n} v_k \varphi_i(\mathbf{s}) \,. \tag{14}$$

The pairwise losses described in (4) with $\lambda(v_i, v_j, s_i - s_j) = v_i \varphi(s_i - s_j) + v_j \varphi(s_j - s_i)$ have this form (take $\varphi_i(\mathbf{s}) = \sum_{j \neq i} \varphi(s_i - s_j)$). The pointwise losses defined in (3) with $\lambda(v_i, s_i) = v_i \varphi(s_i) + (\eta - v_i)\varphi(-s_i)$ do not exactly follow this form because of $\eta - v_i$, but the arguments we develop here can easily be adapted to them.

Bounded Values of $\boldsymbol{U}$ For losses like (14), let us assume that $\boldsymbol{u}(\mathcal{Y}) = K^n$ for some closed set $K \subset \mathbb{R}_+$ [2]. We can then define a pseudo-metric on $\mathcal{D}$ which depends on $\boldsymbol{U}$ by $\omega(\triangle, \triangle') = \|\boldsymbol{U}(\triangle) - \boldsymbol{U}(\triangle')\|_\infty$ and use the induced topology. The set $\boldsymbol{\Delta} = \{\triangle \in \mathcal{D} : \|\boldsymbol{U}(\triangle)\|_\infty \leq B\}$ is then

[2] There is no loss in generality here. If $\boldsymbol{u}$ does not satisfy this assumption, we can simply increase $\mathcal{Y}$ and extend $\boldsymbol{u}$ so that the assumption is satisfied. We will then show that the equivalence between calibration and uniform calibration holds on a larger space of distributions.

compact[3] for any $B > 0$ and condition 1 of Corollary 4.3 is satisfied. Obviously, $\boldsymbol{U}$ is continuous as well, so we only have to check condition 3. Let us assume that the functions $\varphi_k : \mathbb{R}^n \to \mathbb{R}_+, k = 1..n$ are continuous on $\mathbb{R}^n$. Then the $\boldsymbol{u}$-instance of the loss (14) is clearly continuous, and the corresponding inner risk is continuous as well: we have $\underline{L^{\boldsymbol{u}}}(\triangle) = \underline{L} \circ \boldsymbol{U}(\triangle)$, and we know that $\underline{L}$ is continuous on $\boldsymbol{U}(\Delta)$ by Lemma 4.4 because $\boldsymbol{U}(\Delta)$ is compact as the image of a compact by a continuous function. The same argument applies to $\triangle \mapsto \inf_{\mathbf{s} \in \Omega_{i,j}} L(\triangle, \mathbf{s})$, so condition 3 of Corollary 4.3 is satisfied. We summarize this discussion in the following result (the proof is omitted):

Corollary A.1 *Let r be a $(\boldsymbol{u}, \phi)$-GPPM, and assume $\boldsymbol{u}(\mathcal{Y}) = K^n$ for some closed set $K \subset \mathbb{R}^n$. Let $B > 0$ and define $\Delta \subset \mathcal{D}_r$ as $\Delta = \{\triangle \in \mathcal{D}_r | \|\boldsymbol{U}(\triangle)\|_\infty \leq B\}$. Let ℓ be the template loss defined by (14) with continuous $\varphi_k : \mathbb{R}^n \to \mathbb{R}_+, k = 1..n$. Then, the $\boldsymbol{u}$-instance of ℓ is r-calibrated on Δ if and only if it is uniformly r-calibrated.*

Unbounded Values of $\boldsymbol{U}$ As an ending notice before establishing quantitative regret bounds, we would like to consider the case where we do not force the full data distribution to satisfy $\|\boldsymbol{U}(P(.|x))\|_\infty \leq B$ for all x. The explicit regret bounds we will obtain in the next section suggest that if one does not make such an assumption, then there is no regret bound independent of P for many losses. For example, for pairwise template losses satisfying both (4) and (14) and any $\triangle s.t. \|\boldsymbol{U}(\triangle)\|_\infty \leq B$, we obtain regret bounds of the form:

$$\overline{R}(\triangle) - R(\triangle, \mathbf{s}) \leq c\sqrt{B}\sqrt{L(\triangle, \mathbf{s}) - \underline{L}(\triangle)} \tag{15}$$

for some constant $c > 0$. The next section will give more precise values for $c\sqrt{B}$ depending on the loss, but for many template losses we consider, one cannot have a constant independent of B [4]. In that case, we cannot obtain a regret bound which does not depend on the data distribution. However, we can use approach of [28, Theorem 2.18] when uniform calibration does not hold:

Corollary A.2 *Let r be a $(\boldsymbol{u}, \phi)$-GPPM such that $r(y, \sigma) = \sum_{k=1}^n \phi(k) u_{\sigma(k)}(y)$ for all $y \in \mathcal{Y}$ and $\sigma \in \mathfrak{S}_n$. Let ℓ be a scoring loss satisfying (15) with $B = \|\boldsymbol{U}(\triangle)\|_\infty$ for any $\triangle \in \mathcal{D}_{\ell,r}$. Then, for any distribution P on $\mathcal{X} \times \mathcal{Y}$ of type $\mathcal{D}_{\ell,r}$ such that $\overline{\mathcal{R}}(P) < +\infty$ and $\underline{\mathcal{L}}(P) < \infty$, we have, for any measurable scoring function $\mathbf{f}$:*

$$\overline{\mathcal{R}}(P) - \mathcal{R}(P, \mathbf{f}) \leq c\sqrt{\frac{\overline{\mathcal{R}}(P)}{\phi(1)}}\sqrt{\mathcal{L}(P, \mathbf{f}) - \underline{\mathcal{L}}(P)}$$

where c is the constant defined in (15).

Proof Using [28, Theorems 3.2 and 2.13], we know that we can obtain a bound on the regret using Jensen's inequality after taking the expected value of a bound on the inner regret. We assume for now that $x \mapsto \|\boldsymbol{U}(P(.|x))\|_\infty$ is integrable with respect to P_X. Integrating (15) over P_X with $B = \|\boldsymbol{U}(\triangle)\|_\infty$ and $\triangle = P(.|x)$ and using Cauchy-Schwarz inequality, we obtain:

$$\overline{\mathcal{R}}(P) - \mathcal{R}(P, \mathbf{f}) \leq c\sqrt{\int \|\boldsymbol{U}(P(.|x))\|_\infty dP_X(x)}\sqrt{\mathcal{L}(P, \mathbf{f}) - \underline{\mathcal{L}}(P)}$$

Now, note that for any $P(.|x)$, we have $\overline{R}(P(.|x)) \geq \phi(1) \max_{i=1..n} U_i(P(.|x)) = \phi(1)\|\boldsymbol{U}(P(.|x))\|_\infty$. Since $x \mapsto \overline{R}(P(.|x))$ is integrable w.r.t. P_X by assumption, this proves that $x \mapsto \|\boldsymbol{U}(P(.|x))\|_\infty$ is integrable w.r.t. P_X and the desired result. □

Note that it is not a regret bound as we defined them because the bound depends on the data distribution through $\overline{\mathcal{R}}(P)$. Nonetheless, it can still help to obtain non-asymptotic guarantees for the considered losses in practical cases.

[3] We simply have to identify the set of $\triangle$ who have the same expected value under $\boldsymbol{u}$ and assimilate this set to the real value. The only thing that has to be verified is that $\Delta = \{\triangle \in \mathcal{D} : \|\boldsymbol{U}(\triangle)\|_\infty \leq B\}$ is closed, which is the case with our assumption on $\boldsymbol{u}$ since $\{\boldsymbol{U}(\triangle), \triangle \in \Delta\} = [\min K, \min(\sup K, B)]^n$.

[4] Note that if the regret bound for some loss ℓ does not depend on B, as for squared-loss-based scoring losses, then we can directly have a regret bound using Jensen's inequality.

B Proofs

B.1 The Spearman Rank Correlation Coefficient and the AUC are GPPMs

In this subsection, we give the details on how to write the Spearman Rank Correlation Coefficient and the AUC as GPPMs. The calculations are direct for the other measures of Table 1.

$$\begin{aligned}
\texttt{Spearman}(\mathbf{y},\sigma) &= 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(\sigma^{-1}(i) - y^{-1}(i))^2 \\
&= 1 - \frac{6}{n(n^2-1)}\sum_{k=1}^{n}((n-k) - (n - y^{-1}(\sigma(k))))^2 \\
&= 1 - \frac{6}{n(n^2-1)}\sum_{k=1}^{n}(n-k)^2 + (n - y^{-1}(\sigma(k)))^2 - 2(n-k)(n - y^{-1}(\sigma(k))) \\
&= \sum_{k=1}^{n}\frac{12(n-k)(n - y^{-1}(\sigma(k)))}{n(n^2-1)} + 1 - \frac{2n(n-1)(2n-1)}{n(n^2-1)} \\
&= \sum_{k=1}^{n}\frac{12(n-k)(n - y^{-1}(\sigma(k)))}{n(n^2-1)} - \frac{3(n-1)}{(n+1)}\,.
\end{aligned}$$

$$\begin{aligned}
\texttt{AUC}(\mathbf{y},\sigma) &= \frac{1}{\|\mathbf{y}\|_1(n-\|\mathbf{y}\|_1)}\sum_{i:y_i=1}\sum_{j:y_j=0}\mathbf{1}_{\{\sigma^{-1}(i)<\sigma^{-1}(j)\}} \\
&= \frac{1}{\|\mathbf{y}\|_1(n-\|\mathbf{y}\|_1)}\sum_{i=1}^{n} y_i \sum_{j:y_j=0}\mathbf{1}_{\{\sigma^{-1}(i)<\sigma^{-1}(j)\}} \\
&= \frac{1}{\|\mathbf{y}\|_1(n-\|\mathbf{y}\|_1)}\sum_{k=1}^{n} y_{\sigma(k)} \sum_{k'=1}^{n}(1-y_{\sigma(k')})\mathbf{1}_{\{k<k'\}} \\
&= \frac{1}{\|\mathbf{y}\|_1(n-\|\mathbf{y}\|_1)}\left(\sum_{k=1}^{n} y_{\sigma(k)}\sum_{k'=1}^{n}\mathbf{1}_{\{k<k'\}} - \sum_{k=1}^{n}\sum_{k'=1}^{n} y_{\sigma(k)}y_{\sigma(k')}\mathbf{1}_{\{k<k'\}}\right) \\
&= \sum_{k=1}^{n}\frac{y_{\sigma(k)}(n-k)}{\|\mathbf{y}\|_1(n-\|\mathbf{y}\|_1)} - \frac{\|\mathbf{y}\|_1 - 1}{2(n-\|\mathbf{y}\|_1)}\,.
\end{aligned}$$

B.2 Proof of Theorem 3.3

We first define the notion of standardization function (see [3] for details).

Definition B.1 (Standardization Function) *Let $r : \mathcal{Y}\times\mathbb{R}^n \to \mathbb{R}^+$ be a scoring performance measure. A* standardization function *of $\mathcal{Y}$ for r is a function $\boldsymbol{u} : \mathcal{Y} \to \mathbb{R}^n_+$ which, for any distribution $\triangle \in \mathcal{D}_r$ on $\mathcal{Y}$, satisfies:*

$$\boldsymbol{U}(\triangle) \in \operatorname*{arg\,max}_{\mathbf{s}\in\mathbb{R}^n} R\left(\triangle, \mathbf{s}\right).$$

We prove the following lemma before proving Theorem 3.3.

Lemma B.2 *Fix $n \geq 3$ and $\mathcal{Y} = \{0,1\}^n$. Let r be a scoring performance measure. such that for any $\sigma \in \mathfrak{S}_n$, any two indexes i, j and any $\mathbf{y} \in \mathcal{Y}$:*

1. *$r\left(\mathbf{y},\sigma\right) = r\left(\tau_{ij}(\mathbf{y}), \tau_{ij}\circ\sigma\right)$ (symmetry), where τ_{ij} is the transposition $i\leftrightarrow j$*
2. *$\left(y_i > y_j \text{ and } \sigma^{-1}(i) > \sigma^{-1}(j)\right) \Rightarrow r\left(\mathbf{y},\sigma\right) < r\left(\mathbf{y}, \tau_{ij}\circ\sigma\right)$ (strict monotonicity),*

Then, for any standardization function $\boldsymbol{u}$ for r:

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

Proof 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 $y^i_k = 0$ for $k \neq i$. Moreover, we will denote $\delta_{\mathbf{y}}$ Dirac distribution at point $\mathbf{y}$.

We prove the second point by contradiction. Suppose $\mathbf{y}$ is such that $y_i = y_j$ and $u_i(\mathbf{y}) > u_j(\mathbf{y})$. Define a mixture between the two Dirac distributions, $\triangle^\alpha = \alpha\delta_{\mathbf{y}^j} + (1-\alpha)\delta_{\mathbf{y}}$ for $\alpha > 0$ such that $(1-\alpha)u_i(\mathbf{y}) + \alpha u_i(\mathbf{y}^j) > (1-\alpha)u_j(\mathbf{y}) + \alpha u_j(\mathbf{y}^j)$. 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 α implies $U_j\left(\triangle^\alpha\right) > U_i\left(\triangle^\alpha\right)$, which is impossible considering our choice of α.

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 $u_i(\mathbf{y}) + u_i(\mathbf{y}') > u_j(\mathbf{y}) + u_j(\mathbf{y}')$. 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 $y'_i = 0$). Define a mixture between the three Dirac distributions, $\triangle^\beta = \beta\delta_{\mathbf{y}^j} + \frac{1-\beta}{2}\delta_{\mathbf{y}} + \frac{1-\beta}{2}\delta_{\mathbf{y}'}$, and $\beta > 0$ small enough so that $U_i\left(\triangle^\beta\right) > U_j\left(\triangle^\beta\right)$. Since $\sum_k y_k = \sum_k y'_k$ and using the symmetry of the ranking performance measure, 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 β imposes $U_j\left(\triangle^\beta\right) > U_i\left(\triangle^\beta\right)$ by strict monotonicity. This is impossible considering our choice of β. □

Finally, we prove Theorem 3.3.

Proof 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 distribution $\triangle$ which gives probability $1/2$ to each of them. Aiming at a contradiction, suppose the `ERR` (resp. the `AP`) has a standardization function $\boldsymbol{u}^{\texttt{ERR}}$ (resp. $\boldsymbol{u}^{\texttt{AP}}$). Then, by the third point of Lemma B.2, $u_i^{\texttt{ERR}}(\triangle)$ (resp. $u_i^{\texttt{AP}}(\triangle)$) 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:

$$\texttt{ERR}(\triangle, 1 \succ 3 \succ 2 \succ 4) = \texttt{ERR}(\triangle, 1 \succ 2 \succ 3 \succ 4) + \frac{1}{24}$$
$$\texttt{AP}(\triangle, 1 \succ 3 \succ 2 \succ 4) = \texttt{AP}(\triangle, 1 \succ 2 \succ 3 \succ 4) - \frac{1}{12}$$

Thus, some permutations are suboptimal, which contradicts the existence of a standardization function. By contradiction, if ℓ is `ERR`-calibrated (resp `AP`-calibrated), then $\boldsymbol{u}$ is a standardization function, which contradicts the latter assumption. □

B.3 Proof of Lemma 5.1

Proof With the notations of Lemma 4.1, we consider $\nu \in \arg\operatorname{sort}(\boldsymbol{U}(\triangle))$. Then, for any $\sigma \in \mathfrak{S}_n$, we can apply Hölder's inequality on the result of Lemma 4.1. So, we have:

$$\begin{aligned}\overline{R}\left(\triangle\right)\text{-}R\left(\triangle, \mathbf{s}\right) &\leq \left(\sum_{(i,j)\in C_{\mathbf{s}}} \left(\phi(\nu^{-1}(i)) - \phi(\nu^{-1}(j))\right)^q\right)^{\frac{1}{p}} \left(\sum_{(i,j)\in C_{\mathbf{s}}} \left(U_i\left(\triangle\right) - U_j\left(\triangle\right)\right)^q\right)^{\frac{1}{q}} \\ &\leq \left(\sum_{i=1}^{(n+1)/2} \left(\phi(i) - \phi(n-i+1)\right)^q\right)^{\frac{1}{p}} \left(\sum_{(i,j)\in C_{\mathbf{s}}} \left(U_i\left(\triangle\right) - U_j\left(\triangle\right)\right)^q\right)^{\frac{1}{q}}\end{aligned}$$

□

B.4 Results for Pointwise Losses

Now, the objective is to find q and c such that previous pointwise losses verify the conditions of Theorem 5.2. Practically, since we use arguments of convexity, we obtain surrogate regret bounds with $q = 2$. For the rest of this paragraph, for any $v \in \mathbb{R}$, we denote $\underline{\lambda}(v) = \inf_{s\in\mathbb{R}} \lambda(v, s)$. Indeed, for most of our pointwise losses we have,

$$H_{ij}^{-}(\triangle) - H_{ij}(\triangle) = 2\underline{\lambda}\left(\frac{U_i(\triangle) + U_j(\triangle)}{2}\right) - \underline{\lambda}(U_i(\triangle)) - \underline{\lambda}(U_j(\triangle)) \tag{16}$$

We can treat all proposed pointwise losses directly with Theorem 5.2. However, we want to illustrate also the use of Theorem 5.3, so we propose proofs based on it for the *Logistic* and the *Exponential*. Indeed, Theorem 5.3 can't handle the *Square Hinge* and the *Differentiable Hinge* because their minima are not always unique. So we can't find an invertible map between the utilities and the optimal scores. Thus, for these two ones we use the main method described by Theorem 5.2.

Proof Squared Error
We directly prove the conditions of Theorem 5.2 by saying the value of $H_{ij}^{-}(\boldsymbol{u}, \triangle) - H_{ij}(\boldsymbol{u}, \triangle) = \frac{1}{2}(U_i(\triangle) - U_j(\triangle))^2$. □

Proof Square Hinge
$\underline{\lambda}(x) = \frac{t^2}{\eta}x(\eta - x)$ which is $\frac{2t^2}{\eta}$-strongly concave because it's a second-degree polynomial. Since the Square Hinge satisfy (16), $H_{ij}^{-}(\triangle) - H_{ij}(\triangle) \geq \frac{\mu}{8}(U_i(\triangle) - U_j(\triangle))^2$. □

Proof Hinge Differentiable
$\underline{\lambda}(x) = \begin{cases} \left(1 - \frac{\alpha}{2}\right)(\eta - x) - \frac{\alpha(\eta-x)^2}{2x} & \text{if } x \geq \frac{\eta}{2} \\ \left(1 - \frac{\alpha}{2}\right)x - \frac{\alpha x^2}{2(\eta-x)} & \text{if } x \leq \frac{\eta}{2} \end{cases}$ It satisfy (16), so for $U_i(\triangle), U_j(\triangle) \leq \frac{\eta}{2}$

$$\begin{aligned} H_{ij}^{-}(\triangle) - H_{ij}(\triangle) &= \frac{\alpha}{2}\left(\frac{U_i(\triangle)^2}{\eta - U_i(\triangle)} + \frac{U_j(\triangle)^2}{\eta - U_j(\triangle)} - \frac{(U_i(\triangle) + U_j(\triangle))^2}{2\eta - U_i(\triangle) - U_j(\triangle)}\right) \\ &= \frac{\alpha\eta^2}{2}\left(\frac{(U_i(\triangle) - U_j(\triangle))^2}{(\eta - U_i(\triangle))(\eta - U_j(\triangle))(2\eta - U_i(\triangle) - U_j(\triangle))}\right) \\ &\geq \frac{\alpha}{4\eta}(U_i(\triangle) - U_j(\triangle))^2 \end{aligned}$$

The calculus is completely symmetric when $U_i(\triangle), U_j(\triangle) \geq \frac{\eta}{2}$. Then, for the case $U_i(\triangle) < \frac{U_i(\triangle)+U_j(\triangle)}{2} \leq \frac{\eta}{2} < U_j(\triangle)$ (others are symmetric), we just use

$$\begin{aligned} H_{ij}^{-}(\triangle) - H_{ij}(\triangle) &> \underline{\lambda}\left(\frac{U_i(\triangle) + U_j(\triangle)}{2}\right) - \underline{\lambda}(U_i(\triangle)) \\ &\geq \frac{\alpha}{16\eta}(U_i(\triangle) - U_j(\triangle))^2 \end{aligned}$$

□

The *Logistic* and the *Exponential* can be rewritten as Bregman divergences (5). So we can use Theorem 5.3 to handle these two losses.

Proof Logistic
We denote $\psi(x) = x\log(x) + (1-x)\log(1-x)$ which is 4-strongly convex, $h(u) = \log\left(\frac{u}{\eta-u}\right)$ for $u \in (0;\eta)$ invertible satisfying the conditions of Theorem 5.3. As $\lambda(v_i, s_i) = \eta B_\psi\left(\frac{v_i}{\eta}\middle\|.h^{-1}(s_i)\right) - \eta\psi\left(\frac{v_i}{\eta}\right)$, we can directly apply Theorem 5.3. □

Proof Exponential
We denote $\psi(x) = -2\sqrt{x(\eta - x)}$ which is $\frac{4}{\eta}$-strongly convex, $h(u) = \log\left(\sqrt{\frac{u}{\eta-u}}\right)$ for $u \in (0;\eta)$ invertible satisfying the conditions of Theorem 5.3. As $\lambda(v_i, s_i) = B_\psi\left(v_i\middle\|.h^{-1}(s_i)\right) - \psi(v_i)$, we can directly apply Theorem 5.3. □

B.5 Results for Pairwise Losses

Here we gives the proofs that the proposed pairwise losses verify the assumptions of Theorem 5.4.

Proof Squared Error

1. For $\mathbf{d}^*_{ij} = \arg\min_{d_{ij}\in\mathbb{R}} \Lambda^{u_i,u_j}(\triangle, d_{ij})$, we have $d^*_{ij} = U_i(\triangle) - U_j(\triangle)$, so $\mathbf{d}^* \in D$.
2. $\inf_{d_{ij}\leq 0} \Lambda^{u_i,u_j}(\triangle, d_{ij}) - \inf_{d_{ij}\in\mathbb{R}} \Lambda^{u_i,u_j}(\triangle, d_{ij}) = (U_i(\triangle) - U_j(\triangle))^2$.

□

Proof Logistic

1. Since $\frac{\partial \Lambda^{u_i,u_j}(\triangle,d_{ij})}{\partial d_{ij}} = \frac{-U_i(\triangle)}{1+e^{d_{ij}}} + \frac{U_j(\triangle)}{1+e^{-d_{ij}}}$ thus we obtain $d^*_{ij} = \log\left(\frac{U_i(\triangle)}{U_j(\triangle)}\right)$, so $\mathbf{d}^* \in D$.
2. For $d_{ij} \leq 0$, $\frac{\partial \Lambda^{u_i,u_j}(\triangle,d_{ij})}{\partial d_{ij}} < 0$, so $\inf_{d_{ij}\leq 0} \Lambda^{u_i,u_j}(\triangle, d_{ij})$ is reached for $d_{ij} = 0$. So

$$\begin{aligned}\inf_{d_{ij}\leq 0} \Lambda^{u_i,u_j}(\triangle, d_{ij}) - \inf_{d_{ij}\in\mathbb{R}} \Lambda^{u_i,u_j}(\triangle, d_{ij}) &= U_i(\triangle)\log(U_i(\triangle)) + U_j(\triangle)\log(U_j(\triangle)) \\ &\quad -(U_i(\triangle) + U_j(\triangle))\log\left(\frac{U_i(\triangle) + U_j(\triangle)}{2}\right) \\ &\geq \frac{1}{4\|\boldsymbol{U}(\triangle)\|_\infty}(U_i(\triangle) - U_j(\triangle))^2\end{aligned}$$

The last inequality comes from the strong convexity of $x \mapsto x\log(x)$ on a bounded interval.

□

Proof Exponential

1. Since $\frac{\partial \Lambda^{u_i,u_j}(\triangle,d_{ij})}{\partial d_{ij}} = -U_i(\triangle)\, e^{-d_{ij}} + U_j(\triangle)\, e^{d_{ij}}$ thus we obtain $d^*_{ij} = \log\left(\sqrt{\frac{U_i(\triangle)}{U_j(\triangle)}}\right)$, so $\mathbf{d}^* \in D$.
2. For $d_{ij} \leq 0$, $\frac{\partial \Lambda^{u_i,u_j}(\triangle,d_{ij})}{\partial d_{ij}} < 0$, so $\inf_{d_{ij}\leq 0} \Lambda^{u_i,u_j}(\triangle, d_{ij})$ is reached for $d_{ij} = 0$. So

$$\begin{aligned}&\inf_{d_{ij}\leq 0} \Lambda^{u_i,u_j}(\triangle, d_{ij}) - \inf_{d_{ij}\in\mathbb{R}} \Lambda^{u_i,u_j}(\triangle, d_{ij}) \\ &\qquad = U_i(\triangle)\left(1 - \sqrt{\frac{U_j(\triangle)}{U_i(\triangle)}}\right) + U_j(\triangle)\left(1 - \sqrt{\frac{U_i(\triangle)}{U_j(\triangle)}}\right) \\ &\qquad = \left(\sqrt{U_i(\triangle)} - \sqrt{U_j(\triangle)}\right)^2 \\ &\qquad \geq \frac{1}{4\|\boldsymbol{U}(\triangle)\|_\infty}(U_i(\triangle) - U_j(\triangle))^2\end{aligned}$$

□