Top-k Consistency of Learning to Rank Methods

Fen Xia
Institute of Automation
Chinese Academy of Sciences
fen.xia@ia.ac.cn

Tie-Yan Liu
Microsoft Research Asia
tyliu@microsoft.com

Hang Li
Microsoft Research Asia
hanglig@microsoft.com

Abstract

This paper is concerned with the consistency analysis on listwise ranking methods. Among various ranking methods, the listwise methods have competitive performances on benchmark datasets and are regarded as one of the state-of-the-art approaches. Most listwise ranking methods manage to optimize ranking on the whole list (permutation) of objects, however, in practical applications such as information retrieval, correct ranking at the top k positions is much more important. This paper aims to analyze whether existing listwise ranking methods are statistically consistent in the top-k setting. For this purpose, we define a top-k ranking framework, where the true loss (and thus the risks) are defined on the basis of top-k subgroup of permutations. This framework can include the permutation-level ranking framework proposed in previous work as a special case. Based on the new framework, we derive sufficient conditions for a listwise ranking method to be consistent with the top-k true loss, and show an effective way of modifying the surrogate loss functions in existing methods to satisfy these conditions. Experimental results show that after the modification, the methods can work significantly better than their original versions, indicating the correctness of our theoretical analysis.

1 Introduction

Ranking is the central problem in many applications including information retrieval (IR). In recent years, machine learning technologies have been successfully applied to ranking, and many learning to rank methods have been proposed, including the pointwise [12] [9] [6], pairwise [8] [7] [2], and listwise methods [13] [3] [15]. Empirical results on benchmark datasets have demonstrated that the listwise ranking methods have very competitive ranking performances [10].

To explain the high ranking performances of the listwise ranking methods, a theoretical framework was proposed in [15]. In the framework, existing listwise ranking methods are interpreted as making use of different surrogate loss functions of the permutation-level 0-1 loss. Theoretical analysis shows that these surrogate loss functions are all statistically consistent in the sense that the minimization of the conditional expectation of them will lead to obtaining the Bayes ranker, i.e., the optimal ranked list of the objects.

Here we point out that there is a gap between the analysis in [15] and many real ranking problems, where the correct ranking of the entire permutation is not needed. For example, in IR, users usually care much more about the top ranking results and thus only correct ranking at the top positions is important. In this new situation, it is no longer clear whether existing listwise ranking methods are still statistically consistent. The motivation of this work is to perform formal study on the issue.

For this purpose, we propose a new ranking framework, in which the "true loss" is defined on the top-k subgroup of permutations instead of on the entire permutation. The new true loss only measures errors occurring at the top k positions of a ranked list, therefore we refer to it as the top-k true loss (Note that when k equals the length of the ranked list, the top-k true loss will become exactly the permutation-level 0-1 loss). We prove a new theorem which gives sufficient conditions for a surrogate loss function to be consistent with the top-k true loss. We also investigate the change of the conditions with respect to different k. Our analysis shows that, as k decreases, to guarantee the consistency of a surrogate loss function, the requirement on the probability space becomes weaker while the requirement on the surrogate loss function itself becomes stronger. As a result, a surrogate loss function that is consistent with the permutation-level 0-1 loss might not be consistent with the top-k true loss any more. Therefore, the surrogate loss functions in existing listwise ranking methods, which have been proved to be consistent with the permutation-level 0-1 loss, are not theoretically guaranteed to have good performances in the top-k setting. Modifications to these surrogate loss functions are needed to further make them consistent with the top-k true loss. We show how to make such modifications, and empirically verify that such modifications can lead to significant performance improvement. This validates the correctness of our theoretical analysis.

The rest of the paper is organized as follows. Section 2 reviews the permutation-level ranking framework proposed in [15]. Section 3 presents our proposed top-k ranking framework. The theoretical analysis with the proposed framework and its application to modifying existing loss functions are given in Section 4. Experimental results are reported in Section 5, and conclusions are made in the last section.

2 Permutation-level ranking framework

We review the permutation-level ranking framework proposed in [15].

Let X be the input space whose elements are groups of objects to be ranked, Y be the output space whose elements are permutations of objects, and P_{XY} be an unknown but fixed joint probability distribution of X and Y. Let $h : X \to Y$ be a ranking function, and H be its function space (i.e., $h \in H$). Let $\mathbf{x} \in X$ and $y \in Y$, and let $y(i)$ be the index of the object that is ranked at position i in y. The task of ranking is to learn a function that can minimize the expected risk $R(h)$, defined as,

$$R(h) = \int_{X \times Y} l(h(\mathbf{x}), y) dP(\mathbf{x}, y), \tag{1}$$

where $l(h(\mathbf{x}), y)$ is the true loss such that

$$l(h(\mathbf{x}), y) = \begin{cases} 1, & \text{if } \ h(\mathbf{x}) \neq y \\ 0, & \text{if } \ h(\mathbf{x}) = y. \end{cases} \tag{2}$$

The above true loss indicates that if the permutation of the predicted results is exactly the same as the permutation in the ground truth, then the loss is zero; otherwise the loss is one. For ease of reference, we call it permutation-level 0-1 loss. The optimal ranking function which can minimize the expected true risk $R(h^*) = \inf R(h)$ is referred to as the permutation-level Bayes ranker.

$$h^*(\mathbf{x}) = \arg\max_{y \in Y} P(y|\mathbf{x}). \tag{3}$$

In practice, for efficiency consideration, the ranking function is usually defined as $h(\mathbf{x}) = \text{sort}(g(x_1), \ldots, g(x_n))$, where $g(\cdot)$ denotes the scoring function, and sort$(\cdot)$ denotes the sorting function. Since the risks are non-continuous and non-differentiable with respect to the scoring function g, a continuous and differentiable surrogate loss function $\phi(\mathbf{g}(\mathbf{x}), y)$ is usually used as an approximation of the true loss. In this way, the expected risk becomes

$$R^{\phi}(\mathbf{g}) = \int_{X \times Y} \phi(\mathbf{g}(\mathbf{x}), y) dP(\mathbf{x}, y), \tag{4}$$

where $\mathbf{g}(\mathbf{x}) = (g(x_1), \ldots, g(x_n))$ is a vector-valued function induced by g.

It has been shown in [15] that many existing listwise ranking methods can fall into the above framework, with different surrogate loss functions used. And furthermore, their surrogate loss functions are statistically consistent under certain conditions with respect to the permutation-level 0-1 loss. However, as shown in the next section, the permutation-level 0-1 loss cannot well describe the ranking problem in many real applications.

3 Top-k ranking framework

We first describe the real ranking problem, and then propose a new ranking framework.

3.1 Top-k ranking problem

In real ranking applications like IR, people pay more attention to the top-ranked objects. Therefore correct ranking on the top positions is critically important. For example, modern web search engines only return top $1,000$ results and 10 results in each page. According to a user study[1], 62% of search engine users only click on the results within the first pages, and 90% of users click on the results within the first three pages. It means that two ranked lists of documents will likely provide the same experience to the users (and thus suffer the same loss), if they have the same ranking results for the top positions. This, however, cannot be reflected in the permutation-level 0-1 loss in Eq.(2). This characteristic of ranking problems has also been explored in earlier studies in different settings [4, 5, 14], we refer to it as top-k ranking problem.

3.2 Top-k true loss

To better describe the top-k ranking problem, we propose defining the true loss based on the top k positions in a ranked list, referred to as the top-k true loss.

$$l_k(h(\mathbf{x}), y) = \begin{cases} 0, & \text{if } \hat{y}(i) = y(i) \quad \forall i \in \{1, \ldots, k\}, \text{where } \hat{y} = h(\mathbf{x}), \\ 1, & \text{otherwise} \ . \end{cases} \tag{5}$$

The actual value of parameter k is determined by application. When k equals the length of the entire ranked list, the top-k true loss will become the permutation-level 0-1 loss. In this regard, the top-k true loss is more general than the permutation-level 0-1 loss.

With Eq.(5), the expected risk becomes

$$R_k(h) = \int_{X \times Y} l_k(h(\mathbf{x}), y) dP(\mathbf{x}, y). \tag{6}$$

The optimal ranking function of Eq. (6) (i.e., the top-k Bayes ranker) is given by the following proposition.

Proposition 1. *The optimal ranking function of Eq. (6) is any permutation in the top-k subgroup having the highest probability[2], i.e.,*

$$h_k^*(\mathbf{x}) \in \arg\max_{G_k(j_1, j_2, \ldots, j_k) \in G_k} P(G_k(j_1, j_2, ..., j_k)|\mathbf{x}), \tag{7}$$

where $G_k(j_1, j_2, ..., j_k) = \{y \in Y | y(t) = j_t, \forall t = 1, 2, \ldots\ k\}$ denotes a top-k subgroup in which all the permutations have the same top-k true loss; G_k denotes the collection of all top-k subgroups..

Proof. The expected risk of a ranking function $h(\mathbf{x})$ is:

$$\begin{aligned} R_k(h) &= \int_X \sum_{y \in Y} l_k(h(\mathbf{x}), y) P(y|\mathbf{x}) dP(\mathbf{x}) \qquad (8) \\ &= \int_X [1 - \sum_{y \in \{y \in Y | y(t) = h(t), \forall t = 1, 2, \ldots\ k\}} P(y|\mathbf{x})] dP(\mathbf{x}). \end{aligned}$$

Then $h_k^*(\mathbf{x}) \in \arg\max_{G_k(j_1, j_2, \ldots, j_k) \in G_k} P(G_k(j_1, j_2, ..., j_k)|\mathbf{x})$ minimizes (8) for every $\mathbf{x}$. □

With the above setting, we will analyze the consistency of the surrogate loss functions in existing ranking methods with the top-k true loss in the next section.

[1] iProspect Search Engine User Behavior Study, April 2006, http://www.iprospect.com/

[2] Note that the probability of a top-k subgroup is defined as the sum of the probabilities of the permutations in the subgroup (cf., Definitions 6 and 7 in [3]).

4 Theoretical analysis

In this section, we first give the sufficient conditions of consistency for the top-k ranking problem. Second, we show how these conditions change with respect to k. Last, we discuss whether surrogate loss functions in existing methods are consistent, and how to make them consistent if not.

4.1 Statistical consistency

We investigate what kinds of surrogate loss functions $\phi(\mathbf{g}(\mathbf{x}), y)$ are statistically consistent with the top-k true loss. For this purpose, we study whether the ranking function that minimizes the conditional expectation of the surrogate loss function defined as follows coincides with the top-k Bayes ranker as defined in Eq.(7).

$$Q(P(y|\mathbf{x}), \mathbf{g}(\mathbf{x})) = \sum_{y \in Y} P(y|\mathbf{x})\phi(\mathbf{g}(\mathbf{x}), y). \tag{9}$$

According to [1], the above condition is the weakest one providing theoretical guarantee that optimizing a surrogate loss function will lead to obtaining a model achieving the Bayes risk (the expected true risk of the top-k Bayes ranker in our case), when the training sample size approaches infinity .

For ease of explanation, we denote $Q(P(y|\mathbf{x}), \mathbf{g}(\mathbf{x}))$ as $Q(\mathbf{p}, \mathbf{g})$, $\mathbf{g}(\mathbf{x})$ as $\mathbf{g}$ and $P(y|\mathbf{x})$ as p_y. Hence, $Q(\mathbf{p}, \mathbf{g})$ is the pointwise loss of $\mathbf{g}$ at $\mathbf{x}$ with respect to the conditional probability distribution p_y. The key idea is to decompose the optimal sorting of $\mathbf{g}$ into pairwise relationship between scores of objects. For this purpose, we denote $Y_{i,j}$ as a permutation set in which each permutation ranks object i before object j, i.e., $Y_{i,j} \triangleq \{y \in Y : y^{-1}(i) < y^{-1}(j)\}$, and introduce the following definitions.

Definition 2. *We define Λ_{G_k} as a top-k subgroup probability space, i.e., $\Lambda_{G_k} \triangleq \{\mathbf{p} \in R^{|G_k|} : \sum_{G_k(j_1,j_2,...,j_k) \in G_k} p_{G_k(j_1,j_2,...,j_k)} = 1, p_{G_k(j_1,j_2,...,j_k)} \geq 0\}$.*

Definition 3. *A top-k subgroup probability space Λ_{G_k} is order preserving with respect to objects i and j, if $\forall y \in Y_{i,j}$ and $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, we have $p_{G_k(y(1),y(2),...,y(k))} > p_{G_k(\sigma_{i,j}^{-1}y(1),\sigma_{i,j}^{-1}y(2),...,\sigma_{i,j}^{-1}y(k))}$. Here $\sigma_{i,j}^{-1}y$ denotes the permutation in which the positions of objects i and j are exchanged while those of the other objects remain the same as in y.*

Definition 4. *A surrogate loss function ϕ is top-k subgroup order sensitive on a set $\Omega \subset R^n$, if ϕ is a non-negative differentiable function and the following three conditions hold for $\forall$ objects i and j: (1) $\phi(\mathbf{g}, y) = \phi(\sigma_{i,j}^{-1}\mathbf{g}, \sigma_{i,j}^{-1}y)$; (2)Assume $g_i < g_j$, $\forall y \in Y_{i,j}$. If $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, then $\phi(\mathbf{g}, y) \geq \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$ and for at least one y, the strict inequality holds; Otherwise, $\phi(\mathbf{g}, y) = \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$. (3) Assume $g_i = g_j$. $\exists y \in Y_{i,j}$ with $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, we have $\frac{\partial \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)}{\partial g_i} > \frac{\partial \phi(\mathbf{g}, y)}{\partial g_i}$.*

The *order preserving* property of a top-k subgroup probability space (see Definition 3) indicates that if the top-k subgroup probability on a permutation $y \in Y_{i,j}$ is larger than that on permutation $\sigma_{i,j}^{-1}y$, then the relation holds for any other permutation y' in $Y_{i,j}$ if its associated top-k subgroup is different from that of $\sigma_{i,j}^{-1}y'$. The *order sensitive* property of a surrogate loss function (see Definition 4) indicates that (i) $\phi(\mathbf{g}, y)$ should exhibit the natural symmetry in the sense that if the ground truth and the predicted scores are simultaneously exchanged with respect to objects i and j, the surrogate loss does not change; (ii) when a permutation is transformed to another permutation by exchanging the positions of two objects, if the two permutations do not belong to the same top-k subgroup, the loss on the permutation that ranks the two objects in the decreasing order of their scores will be no greater than the loss on its counterpart. (iii) There exists a permutation, for which the speed of change in loss with respect to the score of an object will become faster if exchanging its position with another object with the same score that is ranked after it. A top-k subgroup order sensitive surrogate loss function has several nice properties as shown below.

Proposition 5. *Let $\phi(\mathbf{g}, y)$ be a top-k subgroup order sensitive loss function. $\forall y, \forall \pi \in G_k(y(1), y(2), \ldots, y(k))$, we have $\phi(\mathbf{g}, \pi) = \phi(\mathbf{g}, y)$.*

Proof. 1) We start with the case of transposition where $\pi = \sigma_{i,j}^{-1} y$, i.e., π is the permutation in which the positions of object i and object j are exchanged while those of the other objects remain the same as in y. If $g_i = g_j$, it is obvious that $\phi(\mathbf{g}, \pi) = \phi(\mathbf{g}, y)$. If $g_i \neq g_j$, without loss of generality, we assume $y^{-1}(i) < y^{-1}(j)$ and $g_i < g_j$. Since $\phi(\mathbf{g}, y)$ is a top-k order sensitive loss function and $\pi \in G_k(y(1), y(2), \ldots, y(k))$, we have $\phi(\mathbf{g}, \pi) = \phi(\mathbf{g}, y)$. Thus, a transposition cannot change the value of $\phi(\mathbf{g}, y)$.

In the case of $\pi \neq \sigma_{i,j}^{-1} y$, since $\pi \in G_k(y(1), y(2), \ldots, y(k))$, there exists a permutation σ^{-1} such that $\pi = \sigma^{-1} y$ and σ^{-1} can be represented as a product of finite transpositions within the top-k subgroup. Since a transposition cannot change the value of $\phi(\mathbf{g}, y)$, $\pi = \sigma^{-1} y$ implies $\phi(\mathbf{g}, \pi) = \phi(\mathbf{g}, y)$.

□

Proposition 6. *Let $\phi(\mathbf{g}, y)$ be a top-k subgroup order sensitive surrogate loss function. $\forall$ objects i and j with $g_i = g_j$, $\forall y \in Y_{i,j}$, if $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1} y(1), \sigma_{i,j}^{-1} y(2), ..., \sigma_{i,j}^{-1} y(k))$, then $\frac{\partial \phi(\mathbf{g}, \sigma_{i,j}^{-1} y)}{\partial g_i} \geq \frac{\partial \phi(\mathbf{g}, y)}{\partial g_i}$. Otherwise, $\frac{\partial \phi(\mathbf{g}, \sigma_{i,j}^{-1} y)}{\partial g_i} = \frac{\partial \phi(\mathbf{g}, y)}{\partial g_i}$.*

Proof. $\forall \mathbf{g}$ with $g_i = g_j$, $\forall y \in Y_{i,j}$, it is easy to see $\phi(\mathbf{g}, y) = \phi(\mathbf{g}, \sigma_{i,j}^{-1} y)$. Now, consider a small change of g_i in $\mathbf{g}$. Denote the new score vector as $\mathbf{g}_\epsilon$.

In the case of $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1} y(1), \sigma_{i,j}^{-1} y(2), ..., \sigma_{i,j}^{-1} y(k))$, if $\epsilon > 0$, then $\phi(\mathbf{g}_\epsilon, y) \leq \phi(\mathbf{g}_\epsilon, \sigma_{i,j}^{-1} y)$; If $\epsilon < 0$, then $\phi(\mathbf{g}_\epsilon, y) \geq \phi(\mathbf{g}_\epsilon, \sigma_{i,j}^{-1} y)$. Under both conditions, $\frac{\phi(\mathbf{g}_\epsilon, y) - \phi(\mathbf{g}, y)}{\epsilon} \leq \frac{\phi(\mathbf{g}_\epsilon, \sigma_{i,j}^{-1} y) - \phi(\mathbf{g}, \sigma_{i,j}^{-1} y)}{\epsilon}$. When ϵ approaches zero, we have $\frac{\partial \phi(\mathbf{g}, \sigma_{i,j}^{-1} y)}{\partial g_i} \geq \frac{\partial \phi(\mathbf{g}, y)}{\partial g_i}$.

In the case of $G_k(y(1), y(2), ..., y(k)) = G_k(\sigma_{i,j}^{-1} y(1), \sigma_{i,j}^{-1} y(2), ..., \sigma_{i,j}^{-1} y(k))$, according to Proposition 4, $\phi(\mathbf{g}_\epsilon, y) = \phi(\mathbf{g}_\epsilon, \sigma_{i,j}^{-1} y)$. Thus, $\frac{\partial \phi(\mathbf{g}, \sigma_{i,j}^{-1} y)}{\partial g_i} = \frac{\partial \phi(\mathbf{g}, y)}{\partial g_i}$. □

Proposition 5 shows that all permutations in the same top-k subgroup share the same loss $\phi(\mathbf{g}, y)$ and thus share the same partial difference with respect to the score of a given object. Proposition 6 indicates that the partial difference of $\phi(\mathbf{g}, y)$ also has a similar property to $\phi(\mathbf{g}, y)$ (see the second condition in Definition 4).

Before giving the main theorem, we give the following theorem, which specifies the score relation between two objects for the minimizer of $Q(\mathbf{p}, \mathbf{g})$.

Theorem 7. *Let $\phi(\mathbf{g}, y)$ be a top-k subgroup order sensitive loss function. $\forall i$ and j, if the top-k subgroup probability space is order preserving with resect to them, and $\mathbf{g}$ is a vector which minimizes the point-wise surrogate loss $Q(\mathbf{p}, \mathbf{g})$ in Eq.(9), then $g_i > g_j$.*

Proof. Without loss of generality, we assume $i = 1$, $j = 2$, $g'_1 = g_2$, $g'_2 = g_1$, and $g'_k = g_k (k > 2)$.

First, first prove $g_1 \geq g_2$ by contradiction. Assume $g_1 < g_2$, we have

$$Q(\mathbf{p}, \mathbf{g}') - Q(\mathbf{p}, \mathbf{g}) = \sum_{y \in Y} (p_{\sigma_{1,2}^{-1} y} - p_y) \phi(\mathbf{g}, y) = \sum_{y \in Y_{1,2}} (p_{\sigma_{1,2}^{-1} y} - p_y)(\phi(\mathbf{g}, y) - \phi(\mathbf{g}, \sigma_{1,2}^{-1} y)).$$

The first equation uses the fact $\mathbf{g}' = \sigma_{1,2}^{-1} \mathbf{g}$, and the second equation uses the fact $\sigma_{1,2}^{-1} \sigma_{1,2}^{-1} y = y$. After some algebra, by using Proposition 5, we have,

$$Q(\mathbf{p}, \mathbf{g}') - Q(\mathbf{p}, \mathbf{g}) = \sum_{G_k(y) \in \{G_k : G_k(y) \neq G_k(\sigma_{1,2}^{-1} y)\} : y \in Y_{1,2}} (p_{G_k(\sigma_{1,2}^{-1} y)} - p_{G_k(y)})(\phi(\mathbf{g}, y) - \phi(\mathbf{g}, \sigma_{1,2}^{-1} y)),$$

where $G_k(y)$ denotes the subgroup that y belongs to.

Since $g_1 < g_2$, we have $\phi(\mathbf{g}, y) \geq \phi(\mathbf{g}, \sigma_{1,2}^{-1} y)$. Meanwhile, $p_{G_k(\sigma_{1,2}^{-1} y)} < p_{G_k(y)}$ due to the order preserving of the top-k subgroup probability space. Thus each component in the sum is non-positive and at least one of them is negative, which means $Q(\mathbf{p}, \mathbf{g}') < Q(\mathbf{p}, \mathbf{g})$. This is a contradiction to the optimality of $\mathbf{g}$. Therefore, we must have $g_1 \geq g_2$.

Second, we prove $g_1 \neq g_2$ also by contradiction. Assume $g_1 = g_2$. By setting the derivative of $Q(\mathbf{p}, \mathbf{g})$ with respect to g_1 and g_2 to zero and compare them[3], we have,

$$\sum_{y \in Y_{1,2}} (p_y - p_{\sigma_{1,2}^{-1} y})(\frac{\partial \phi(\mathbf{g}, y)}{\partial g_1} - \frac{\partial \phi(\mathbf{g}, \sigma_{1,2}^{-1} y)}{\partial g_1}) = 0.$$

After some algebra, we obtain,

$$\sum_{G_k(y) \in \{G_k : G_k(y) \neq G_k(\sigma_{1,2}^{-1} y)\} : y \in Y_{1,2}} (p_{G_k(y)} - p_{G_k(\sigma_{1,2}^{-1} y)})(\frac{\partial \phi(\mathbf{g}, y)}{\partial g_1} - \frac{\partial \phi(\mathbf{g}, \sigma_{1,2}^{-1} y)}{\partial g_1}) = 0.$$

According to Proposition 6, we have $\frac{\partial \phi(\mathbf{g}, y)}{\partial g_1} \leq \frac{\partial \phi(\mathbf{g}, \sigma_{1,2}^{-1} y)}{\partial g_1}$. Meanwhile, $p_{G_k(\sigma_{1,2}^{-1} y)} < p_{G_k(y)}$ due to the order preserving of the top-k subgroup probability space. Thus, the above equation cannot hold since every component in the sum is non-positive and at least one of them is negative according to Definition 4. □

Based on Theorem 7, we give the main theorem (Theorem 8), which states the conditions for a surrogate loss function to be consistent with the top-k true loss.

Theorem 8. *Let ϕ be a top-k subgroup order sensitive loss function on $\Omega \subset R^n$. For $\forall n$ objects, if its top-k subgroup probability space is order preserving with respect to $n-1$ object pairs $\{(j_i, j_{i+1})\}_{i=1}^{k}$ and $\{(j_{k+s_i}, j_{k+i} : 0 \leq s_i < i)\}_{i=2}^{n-k}$, then the loss $\phi(\mathbf{g}, y)$ is consistent with the top-k true loss as defined in Eq.(5).*

Proof. We first prove that the top-k subgroup with the maximum conditional probability is $G_k(j_1, j_2, \cdots, j_k)$. We then show the ranked list derived by sorting the minimizer of the pointwise surrogate loss of $Q(\mathbf{p}, \mathbf{g})$ belongs to $G_k(j_1, j_2, \cdots, j_k)$, which implies the consistency.

$\forall y$ with $G_k(y(1), y(2), \cdots, y(k)) \neq G_k(j_1, j_2, \cdots, j_k)$, we start from the first position of y, i.e., $i = 1$. If $y(1) = j_l$, $l > k+1$, then we can use finite times of transposition of object pairs in $\{(j_{k+s_i}, j_{k+i} : 0 \leq s_i < i)\}_{i=2}^{n-k}$ to achieve $y(1) = j_k$. If $y(1) = j_l$, $l \leq k+1$, then we can use $l-1$ times of transposition of object pairs in $\{(j_i, j_{i+1})\}_{i=1}^{k}$ to achieve $y(1) = j_1$. If $y(1) = j_1$, then a similar process can be done for $i = 2$. At last, we terminate the transposition process at $i = k$ and obtain a permutation whose associated top-k subgroup is $G_k(j_1, j_2, \cdots, j_k)$. Note that the probability of the associated top-k subgroup continuously increases throughout the whole process due to the order preserving property. This implies $\forall y$ with $G_k(y(1), y(2), \cdots, y(k)) \neq G_k(j_1, j_2, \cdots, j_k)$, $G_k(y(1), y(2), \cdots, y(k)) < G_k(j_1, j_2, \cdots, j_k)$.

According to Theorem 7, since the top-k subgroup probability space is order preserving with respect to $n-1$ object pairs $\{(j_i, j_{i+1})\}_{i=1}^{k}$ and $\{(j_{k+s_i}, j_{k+i} : 0 \leq s_i < i)\}_{i=2}^{n-k}$ and ϕ is a top-k subgroup order sensitive loss function, the minimizer of pointwise surrogate loss should have $g_{j_i} > g_{j_{i+1}}$, $i = 1, \cdots, k$ and $g_{j_{k+s_i}} > g_{j_{k+i}}$, $i = 2, \cdots, n-k$ and $0 \leq s_i < i$. This indicates that the ranked list derived by sorting the minimizer $\mathbf{g}$ of $Q(\mathbf{p}, \mathbf{g})$ belongs to $G_k(j_1, j_2, \cdots, j_k)$. □

4.2 Consistency with respect to k

In this subsection, we discuss the consistency conditions with respect to various k values.

First, we have the following proposition for the top-k subgroup probability space.

Proposition 9. *If the top-k subgroup probability space is order preserving with respect to objects i and j, then the top-$(k-1)$ subgroup probability space is order preserving with respect to i and j.*

Proof. $\forall y \in Y_{i,j}$ with $G_{k-1}(y(1), y(2), ..., y(k-1)) \neq G_{k-1}(\sigma_{i,j}^{-1} y(1), \sigma_{i,j}^{-1} y(2), ..., \sigma_{i,j}^{-1} y(k-1))$. It is easy to see $i \in \{y(1), y(2), \cdots, y(k-1)\}$. The y falls into one of two cases, i.e., $j \in \{y(1), y(2), \cdots, y(k-1)\}$ or $j \notin \{y(1), y(2), \cdots, y(k-1)\}$.

[3] By trivial modification, one can handle the case when g_1 or g_2 is infinite (cf. [16]).

For $j \in \{y(1), y(2), \cdots, y(k-1)\}$, denote indexes of objects that are not ranked within the top $k-1$ positions of y as $(j_1, j_2, \cdots, j_{n-k+1})$, we have

$$\begin{aligned} & p_{G_{k-1}(y(1),...,y(k-1))} - p_{G_{k-1}(\sigma_{i,j}^{-1}y(1),...,\sigma_{i,j}^{-1}y(k-1))} \\ = & \sum_{i=1}^{n-k+1} \left(p_{G_k(y(1),...,y(k-1),j_i)} - p_{G_k(\sigma_{i,j}^{-1}y(1),...,\sigma_{i,j}^{-1}y(k-1),j_i)}\right) \end{aligned} \tag{10}$$

Since the top-k subgroup probability space is order preserving with respect to objects i and j, every component in the sum is positive and thus $p_{G_{k-1}(y(1),y(2),...,y(k-1))} > p_{G_{k-1}(\sigma_{i,j}^{-1}y(1),\sigma_{i,j}^{-1}y(2),...,\sigma_{i,j}^{-1}y(k-1))}$.

For $j \notin \{y(1), y(2), \cdots, y(k-1)\}$, we can prove $p_{G_{k-1}(y(1),y(2),...,y(k-1))} > p_{G_{k-1}(\sigma_{i,j}^{-1}y(1),\sigma_{i,j}^{-1}y(2),...,\sigma_{i,j}^{-1}y(k-1))}$ using the similar method. The only difference is that, in the sum, when the top-k subgroup is $G_k(y(1), ..., y(k-1), j)$ for $G_{k-1}(y(1), y(2), ..., y(k-1))$, the corresponding top-$k$ subgroup is $G_k(y(1), ..., y(k-1), i)$ for $G_{k-1}(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k-1))$. □

To better understand the theorem, we give an example here. Suppose there are three objects $\{1, 2, 3\}$ to be ranked. If the top-2 subgroup probability space is order preserving with respect to objects 1 and 2, then we have $p_{G_2(1,2)} > p_{G_2(2,1)}$, $p_{G_2(1,3)} > p_{G_2(2,3)}$ and $p_{G_2(3,1)} > p_{G_2(3,2)}$. On the other hand, for top-1, we have $p_{G_1(1)} > p_{G_1(2)}$. Note that $p_{G_1(1)} = p_{G_2(1,2)} + p_{G_2(1,3)}$ and $p_{G_1(2)} = p_{G_2(2,1)} + p_{G_2(2,3)}$. Thus, it is easy to verify that Proposition 9 holds for this case while the opposite does not.

Second, we obtain the following proposition for the surrogate loss function ϕ.

Proposition 10. *If the surrogate loss function ϕ is top-k subgroup order sensitive on a set $\Omega \subset R^n$, then it is also top-(k+1) subgroup order sensitive on the same set.*

Proof. It is easy to see that only the second property in Definition 3 needs to be proved for the property of top-(k+1) subgroup order sensitive. $\forall$ object pair i and j with $g_i < g_j$, $\forall y \in Y_{i,j}$. If $G_{k+1}(y(1), y(2), ..., y(k+1)) \neq G_{k+1}(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k+1))$, then we have either $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$ or $G_k(y(1), y(2), ..., y(k)) = G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$. Since ϕ is top-k subgroup order sensitive, for $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, we have $\phi(\mathbf{g}, y) \geq \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$ and for at least one y, the strict inequality holds. For $G_k(y(1), y(2), ..., y(k)) = G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, $\phi(\mathbf{g}, y) = \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$ according to Proposition 5. Combining both cases, we have $\phi(\mathbf{g}, y) \geq \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$ and for at least one y, the strict inequality holds.

If $G_{k+1}(y(1), y(2), ..., y(k+1)) = G_{k+1}(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k+1))$, it is easy to see that $G_k(y(1), y(2), ..., y(k)) = G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$. Thus, $\phi(\mathbf{g}, y) = \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$. □

Again, we provide an example here. Consider the same setting in the previous example. Assume $g_1 < g_2$. If ϕ is top-1 subgroup order sensitive, then we have $\phi(\mathbf{g}, (1,2,3)) \geq \phi(\mathbf{g}, (2,1,3))$, $\phi(\mathbf{g}, (1,3,2)) \geq \phi(\mathbf{g}, (2,3,1))$, and $\phi(\mathbf{g}, (3,1,2)) = \phi(\mathbf{g}, (3,2,1))$. Considering Proposition 5, the two inequalities become strict. On the other hand, if ϕ is top-2 subgroup order sensitive, the following inequalities hold with at least one of them becoming strict: $\phi(\mathbf{g}, (1,2,3)) \geq \phi(\mathbf{g}, (2,1,3))$, $\phi(\mathbf{g}, (1,3,2)) \geq \phi(\mathbf{g}, (2,3,1))$, and $\phi(\mathbf{g}, (3,1,2)) \geq \phi(\mathbf{g}, (3,2,1))$. Therefore top-1 subgroup order sensitive is a special case of top-2 subgroup order sensitive.

From the above propositions, we have the following conclusions.

- For the consistency with the top-k true loss, when k becomes smaller, the requirement on the probability space becomes weaker but the requirement on the surrogate loss function becomes stronger. Since we never know the real property of the (unknown) probability space, it is more likely the requirement on the probability space for the consistency with the top-k true loss can be satisfied than that for the top-l ($l > k$) true loss. Specifically, it is the riskiest to assume the requirements for the permutation-level 0-1 loss to hold.

- If we fix the true loss to be top-k and the probability space to be top-k subgroup order preserving, then the surrogate loss function should be at most top-l ($l \leq k$) subgroup order sensitive in order to meet the consistency conditions. It is not guaranteed that a top-l ($l > k$) subgroup order sensitive surrogate loss function can be consistent with the top-k true loss. Specifically, a top-1 subgroup order sensitive surrogate loss function may be consistent with any top-k true loss, but a permutation-level order sensitive surrogate loss function will be inconsistent with any top-k true loss, if k is smaller than the length of the ranked list.

To ease the understanding of the above discussions, let us see an example shown in the following proposition. It basically says that given a probability space that is top-1 subgroup order preserving, a top-3 subgroup order sensitive surrogate loss function may not be consistent with the top-1 true loss.

Proposition 11. *Suppose there are three objects to be ranked. ϕ is a top-3 subgroup order sensitive loss function and the strict inequality $\phi(\mathbf{g},(3,1,2)) < \phi(\mathbf{g},(3,2,1))$ holds when $g_1 > g_2$. The probabilities of permutations are $p_{123} = p_1$, $p_{132} = 0$, $p_{213} = p_2$, $p_{231} = 0$, $p_{312} = 0$, $p_{321} = p_2$ respectively, where $p_1 > p_2$. Then ϕ is not consistent with respect to the top-1 true loss.*

Proof. Since the sum of probabilities equals to one and $p_1 > p_2$, we get $p_2 = 1/2(1 - p_1)$, and $1/3 < p_1 \leq 1$. Consider $\mathbf{g} = [g_1, g_2, g_3]$ with $g_1 > g_2$. We shall prove that with some p_1 values $\mathbf{g}$ cannot be the optimal solution. Let $\mathbf{g}' = [g'_1, g'_2, g'_3]$ and $g'_1 = g_2, g'_2 = g_1, g'_3 = g_3$. It is easy to see $\mathbf{g}' = \sigma^{-1}_{1,2}(\mathbf{g})$

$$Q(\mathbf{p},\mathbf{g}) = p_{123}\phi(\mathbf{g},(1,2,3)) + p_{213}\phi(\mathbf{g},(2,1,3)) + p_{321}\phi(\mathbf{g},(3,2,1))$$

$$\begin{aligned} Q(\mathbf{p},\mathbf{g}') &= p_{123}\phi(\mathbf{g}',(1,2,3)) + p_{213}\phi(\mathbf{g}',(2,1,3)) + p_{321}\phi(\mathbf{g}',(3,2,1)) \\ &= p_{123}\phi(\mathbf{g},(2,1,3)) + p_{213}\phi(\mathbf{g},(1,2,3)) + p_{321}\phi(\mathbf{g},(3,1,2)) \end{aligned}$$

Since $\phi(\mathbf{g}, y)$ is top-3 subgroup order sensitive and $g_2 < g_1$, $\phi(\mathbf{g},(1,2,3)) \leq \phi(\mathbf{g},(2,1,3))$ and $\phi(\mathbf{g},(3,1,2)) < \phi(\mathbf{g},(3,2,1))$.

After some algebra, the term $Q(\mathbf{p},\mathbf{g}') < Q(\mathbf{p},\mathbf{g})$ is equivalent to

$$\frac{\phi(\mathbf{g},(2,1,3)) - \phi(\mathbf{g},(1,2,3))}{\phi(\mathbf{g},(3,2,1)) - \phi(\mathbf{g},(3,1,2))} < \frac{1-p_1}{3p_1 - 1} \tag{11}$$

The right-hand side of formula (11) is a decreasing function of p_1. Since $1/3 < p_1 \leq 1$, $\frac{1-p_1}{3p_1-1} \in [0, +\infty)$. Note that $\forall\mathbf{g}$, the left-hand side of formula (11) is non negative and bounded. Therefore, we can make p_1 approach $1/3$ enough so that the inequality (11) holds. This indicates that $\mathbf{g}$ with $g_1 > g_2$ is not optimal with some p_1 values, which implies ϕ is not consistent. □

The above discussions imply that although the surrogate loss functions in existing listwise ranking methods are consistent with the permutation-level 0-1 loss (under a rigid condition), they may not be consistent with the top-k true loss any more (under a milder condition). Therefore, it is necessary to modify these surrogate loss functions. We will make discussions on this in the next subsection.

4.3 Consistent surrogate loss functions

In [15], the surrogate loss functions in ListNet, RankCosine, and ListMLE have been proved to be permutation-level order sensitive. According to the discussion in the previous subsection, however, they may not be top-k subgroup order sensitive any more, and therefore not consistent with the top-k true loss. Even for the consistency with the permutation-level 0-1 loss, in order to guarantee these surrogate loss functions to be consistent, the requirement on the probability space may be too strong in some real scenarios. To tackle the challenge, it is desirable to modify these surrogate loss functions to make them top-k subgroup order sensitive. Actually this is doable, and the modifications to the aforementioned surrogate loss functions are given as follows.

4.3.1 Likelihood loss

The likelihood loss is the loss function used in ListMLE [15], which is defined as below,

$$\phi(\mathbf{g}(\mathbf{x}), y) = -\log P(y|\mathbf{x};\mathbf{g}), \qquad \text{where } P(y|\mathbf{x};\mathbf{g}) = \prod_{i=1}^{n} \frac{\exp(g(x_{y(i)}))}{\sum_{t=i}^{n}\exp(g(x_{y(t)}))}. \tag{12}$$

We propose replacing the permutation probability with the top-k subgroup probability (which is also defined with the Luce model [11]) in the above definition:

$$P(y|\mathbf{x};\mathbf{g}) = \prod_{i=1}^{k} \frac{\exp(g(x_{y(i)}))}{\sum_{t=i}^{n}\exp(g(x_{y(t)}))}. \tag{13}$$

It can be proved that the modified likelihood loss function is top-k subgroup order sensitive, as given in the following proposition.

Proposition 12. *The modified likelihood loss function of Eq.(13) is top-k subgroup order sensitive.*

Proof. It is easy to verify that the first and third properties of Definition 3 hold for the modified likelihood loss function.

For simplicity, we omit $\mathbf{x}$ in the modified likelihood loss function and prove the second property of Definition 3.

For $\forall$ object pair i and j with $g_i < g_j$, $\forall y \in Y_{i,j}$, we have:

$$\phi(\mathbf{g}, y) - \phi(\mathbf{g}, \sigma_{i,j}^{-1} y) \;=\; \log P(\sigma_{i,j}^{-1} y|\mathbf{g}) - \log P(y|\mathbf{g}).$$

If $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, similar to Theorem 3 and Theorem 8 in [3], it is easy to show $P(\sigma_{i,j}^{-1} y|\mathbf{g}) > \log P(y|\mathbf{g})$. Thus, $\phi(\mathbf{g}, y) > \phi(\mathbf{g}, \sigma_{i,j}^{-1} y)$.

If $G_k(y(1), y(2), ..., y(k)) = G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, it is easy to verify that $P(\sigma_{i,j}^{-1} y|\mathbf{g}) = \log P(y|\mathbf{g})$, and hence $\phi(\mathbf{g}, y) = \phi(\mathbf{g}, \sigma_{i,j}^{-1} y)$. □

4.3.2 Cosine loss

The cosine loss is the loss function used in RankCosine [13], which is defined as follows,

$$\phi(\mathbf{g}(\mathbf{x}), y) = \frac{1}{2}\left(1 - \frac{\psi_y(\mathbf{x})^T \mathbf{g}(\mathbf{x})}{\|\psi_y(\mathbf{x})\|\|\mathbf{g}(\mathbf{x})\|}\right), \tag{14}$$

where the score vector of the ground truth is produced by a mapping function $\psi_y(\cdot) : R^d \rightarrow R$, which retains the order in a permutation, i.e., $\psi_y(x_{y(1)}) > \cdots > \psi_y(x_{y(n)})$.

We propose changing the mapping function as follows. Let the mapping function retain the order for the top k positions in a permutation and map all the remaining positions to a small value (which is smaller than the score of any object ranked at the top-k positions), i.e., $\psi_y(x_{y(1)}) > \cdots > \psi_y(x_{y(k)}) > \psi_y(x_{y(k+1)}) = \cdots = \psi_y(x_{y(n)}) = \epsilon$. It can be proved that after the modification, the cosine loss becomes top-k subgroup order sensitive.

Proposition 13. *The modified cosine loss function of Eq.(14) is top-k subgroup order sensitive.*

Proof. Similar to the proof of Proposition 12, we only prove the second property of Definition 3 and omit $\mathbf{x}$ in the modified cosine loss function for simplicity.

For $\forall$ object pair i and j with $g_i < g_j$, $\forall y \in Y_{i,j}$, we have:

$$\begin{aligned} \phi(\mathbf{g}, y) - \phi(\mathbf{g}, \sigma_{i,j}^{-1} y) &= \frac{1}{2}\left(\frac{\psi_{\sigma_{i,j}^{-1}y}^T \mathbf{g}}{\|\psi_{\sigma_{i,j}^{-1}y}\|\|\mathbf{g}\|} - \frac{\psi_y^T \mathbf{g}}{\|\psi_y\|\|\mathbf{g}\|}\right) \\ &= \frac{(\psi_{y_{y^{-1}(j)}} - \psi_{y_{y^{-1}(i)}})(g_i - g_j)}{2\|\psi_y\|\|\mathbf{g}\|} \end{aligned}$$

The second equality is based on the fact $\|\psi_{\sigma_{i,j}^{-1}y}\| = \|\psi_y\|$.

If $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, according to the property of the mapping function, since $y^{-1}(i) < y^{-1}(j)$, $\psi_{y_{y^{-1}(j)}} < \psi_{y_{y^{-1}(i)}}$. Using the condition $g_i < g_j$, it is easy to show $\phi(\mathbf{g}, y) > \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$.

If $G_k(y(1), y(2), ..., y(k)) = G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, since $\psi_{y_{y^{-1}(j)}} = \psi_{y_{y^{-1}(i)}} = \epsilon$, and hence $\phi(\mathbf{g}, y) = \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$. □

4.3.3 Cross entropy loss

The cross entropy loss is the loss function used in ListNet [3], defined as follows,

$$\phi(\mathbf{g}(\mathbf{x}), y) = D(P(\pi|\mathbf{x}; \psi_y)||P(\pi|\mathbf{x}; \mathbf{g})), \tag{15}$$

where ψ is a mapping function which definition is similar to that in RankCosine, and $P(\pi|\mathbf{x}; \psi_y)$ and $P(\pi|\mathbf{x}; \mathbf{g})$ are Luce model based permutation probabilities.

We propose using a similar mapping function to that in the modified cosine loss for the modification of the cross entropy loss[4]. It can be proved that such a modification can make it top-k subgroup order sensitive.

Proposition 14. *The modified cross entropy loss function of Eq.(15) is top-k subgroup order sensitive.*

Proof. Similar to the proof of Proposition 12, we only prove the second property of Definition 3 and omit $\mathbf{x}$ in the modified cross entropy loss function for simplicity.

$\forall$ object pair i and j with $g_i < g_j$, $\forall y \in Y_{i,j}$, we have:

$$\begin{aligned}\phi(\mathbf{g}, y) - \phi(\mathbf{g}, \sigma_{i,j}^{-1}y) &= \sum_{\pi \in Y}(P(\pi|\psi_{\sigma_{i,j}^{-1}y}) - P(\pi|\psi_y)) \log P(\pi|\mathbf{g}) \\ &= \sum_{\pi \in Y_{i,j}}(P(\sigma_{i,j}^{-1}\pi|\psi_y) - P(\pi|\psi_y))(\log P(\pi|\mathbf{g}) - \log P(\sigma_{i,j}^{-1}\pi|\mathbf{g})).\end{aligned}$$

The fact of symmetry, i.e., $P(\pi|\psi_y) = P(\sigma_{i,j}^{-1}\pi|\psi_{\sigma_{i,j}^{-1}y})$, is used in the second equation.

If $G_k(y(1), y(2), ..., y(k)) \neq G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, according to the property of the mapping function, since $y^{-1}(i) < y^{-1}(j)$, $\psi_{y_{y^{-1}(j)}} < \psi_{y_{y^{-1}(i)}}$. Hence, according to Theorem 3 in [3], for $\pi \in Y_{i,j}$ $P(\sigma_{i,j}^{-1}\pi|\psi_y) < P(\pi|\psi_y)$. Similarly, since $g_i < g_j$, $\log P(\pi|\mathbf{g}) < \log P(\sigma_{i,j}^{-1}\pi|\mathbf{g})$. Thus every components in the sum is positive, which means $\phi(\mathbf{g}, y) > \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$.

If $G_k(y(1), y(2), ..., y(k)) = G_k(\sigma_{i,j}^{-1}y(1), \sigma_{i,j}^{-1}y(2), ..., \sigma_{i,j}^{-1}y(k))$, since $\psi_{y_{y^{-1}(j)}} = \psi_{y_{y^{-1}(i)}} = \epsilon$, and hence $P(\sigma_{i,j}^{-1}\pi|\psi_y) = P(\pi|\psi_y)$. This implies $\phi(\mathbf{g}, y) = \phi(\mathbf{g}, \sigma_{i,j}^{-1}y)$. □

5 Experimental results

In order to validate the theoretical analysis in this work, we conducted some empirical study. Specifically, we used OHSUMED, TD2003, and TD2004 in the LETOR benchmark dataset [10] for the experiments. As evaluation measure, we adopted Normalized Discounted Cumulative Gain (N) at positions 1, 3, and 10, and Precision (P) at positions 1, 3, and 10.[5] It is obvious that these measures are top-k related and are suitable to evaluate the ranking performance for top-k ranking problems.

[4]Note that in [3], a top-k cross entropy loss was also proposed, by using the top-k Luce model. However, it can be validated that the so-defined top-k cross entropy loss is still permutation-level order sensitive, but not top-k subgroup order sensitive. In other words, the modification to the mapping function is still needed to make it top-k subgroup order sensitive, and thus consistent with the top-k true loss.

[5]On datasets with only two ratings such as TD2003 and TD2004, N@1 equals P@1.

We chose ListMLE as an example to perform the experimental study since the likelihood loss has nice properties such as convexity, soundness and linear computational complexity [15]. We refer to the new method we obtained by applying the modifications mentioned in Section 4.3 as top-k ListMLE. We tried different values of k (i.e., k=1, 3, 10, and the exact length of the ranked list). Obviously the last case corresponds to the original likelihood loss in ListMLE.

Since the training data in LETOR is given in the form of multi-level ratings, we adopted the methods proposed in [15] to produce the ground truth ranked list in the experiments. We then used stochastic gradient descent as the algorithm for optimization. As for ranking model, we chose linear Neural Network, since the same model has been widely used in other work [3, 13, 15].

The experimental results are summarized in Tables 1-3.

Table 1: Ranking accuracies on OHSUMED

Methods	N@1	N@3	N@10	P@1	P@3	P@10
ListMLE	0.548	0.473	0.446	0.642	0.582	0.495
Top-1 ListMLE	0.529	0.482	0.447	0.652	0.595	0.499
Top-3 ListMLE	0.535	0.484	0.445	0.671	0.608	0.504
Top-10 ListMLE	0.558	0.473	0.444	0.672	0.601	0.509

Table 2: Ranking accuracies on TD2003

Methods	N/P@1	N@3	N@10	P@3	P@10
ListMLE	0.24	0.253	0.261	0.22	0.146
Top-1 ListMLE	0.4	0.329	0.314	0.3	0.176
Top-3 ListMLE	0.44	0.382	0.343	0.34	0.204
Top-10 ListMLE	0.5	0.410	0.378	0.38	0.22

Table 3: Ranking accuracies on TD2004

Methods	N/P@1	N@3	N@10	P@3	P@10
ListMLE	0.4	0.351	0.356	0.284	0.188
Top-1 ListMLE	0.52	0.469	0.451	0.413	0.248
Top-3 ListMLE	0.506	0.456	0.458	0.417	0.261
Top-10 ListMLE	0.52	0.469	0.472	0.413	0.269

From the tables, we can see that with the modifications the ranking accuracies of ListMLE can be significantly boosted, in terms of all measures, on both TD2003 and TD2004. This clearly validates our theoretical analysis. On OHSUMED, all the loss functions achieve comparable performances. The possible explanation is that the probability space in OHSUMED is well formed such that it is order preserving for many different k values.

Next, we take Top-10 ListMLE as an example to make comparison with some other baseline methods such as Ranking SVM [8], RankBoost [7], ListNet [3], and RankCosine [13]. The results are listed in Tables 4-6. We can see from the tables, Top-10 ListMLE achieves the best performance among all the methods on the TD2003 and TD2004 datasets in terms of almost all the measures. On the OHSUMED dataset, it also performs fairly well as compared to the other methods. Especially for N@1 and P@1, it significantly outperforms all the other methods on all the datasets.

From the above experimental results, we can come to the conclusion that for real ranking applications like IR (where top-k evaluation measures are widely used), it is better to use the top-k true loss than the permutation-level 0-1 loss, and is better to use the modified surrogate loss functions than the original surrogate loss functions.

6 Conclusion

In this paper we have proposed a new ranking framework, which can better describe real ranking applications like information retrieval. In the new framework, the true loss is defined on the top-k

Table 4: Ranking accuracies on OHSUMED

Methods	N@1	N@3	N@10	P@1	P@3	P@10
RankBoost	0.497	0.472	0.435	0.604	0.586	0.495
Ranking SVM	0.495	0.464	0.441	0.633	0.592	0.507
ListNet	0.523	**0.477**	**0.448**	0.642	**0.602**	**0.509**
RankCosine	0.523	0.475	0.437	0.642	0.589	0.493
Top-10 ListMLE	**0.558**	0.473	0.444	**0.672**	0.601	**0.509**

Table 5: Ranking accuracies on TD2003

Methods	N/P@1	N@3	N@10	P@3	P@10
RankBoost	0.26	0.270	0.285	0.24	0.178
Ranking SVM	0.42	0.378	0.341	0.34	0.206
ListNet	0.46	0.408	0.374	0.36	**0.222**
RankCosine	0.36	0.346	0.322	0.3	0.182
Top-10 ListMLE	**0.5**	**0.410**	**0.378**	**0.38**	0.22

subgroup of permutations. We have formulated sufficient conditions for a surrogate loss function to be statistically consistent with the top-k true loss. We have also discussed how to modify the loss functions in existing listwise ranking methods to make them consistent with the top-k true loss. Our experiments have shown that with the proposed modifications, ListMLE can significantly outperform its original version, and also many other ranking methods.

As for future work, we plan to investigate the following issues. (1) we will empirically study the modified ListNet and RankCosine, to see whether their performances can also be significantly boosted in the top-k setting. (2) We will study the consistency of the pointwise and pairwise loss functions with the top-k true loss.

References

[1] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proc. of ICML'05*, pages 89–96, 2005.

[3] Z. Cao, T. Qin, T. Y. Liu, M. F. Tsai, and H. Li. Learning to rank: From pairwise approach to listwise approach. In *Proc. of ICML'07*, pages 129–136, 2007.

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

[5] D. Cossock and T. Zhang. Subset ranking using regression. In *Proc. of COLT*, pages 605–619, 2006.

[6] D. Cossock and T. Zhang. Statistical analysis of bayes optimal subset ranking. *Information Theory*, 54:5140–5154, 2008.

[7] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. In *Proc. of ICML'98*, pages 170–178, 1998.

Table 6: Ranking accuracies on TD2004

Methods	N/P@1	N@3	N@10	P@3	P@10
RankBoost	0.48	0.463	0.471	0.404	0.253
Ranking SVM	0.44	0.409	0.420	0.351	0.225
ListNet	0.439	0.437	0.457	0.399	0.257
RankCosine	0.439	0.397	0.405	0.328	0.209
Top-10 ListMLE	**0.52**	**0.469**	**0.472**	**0.413**	**0.269**

[8] R. Herbrich, T. Graepel, and K. Obermayer. Support vector vector learning for ordinal regression. In *Proc. of ICANN'99*, pages 97–102, 1999.

[9] P. Li, C. Burges, and Q. Wu. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in Neural Information Processing Systems 20(NIPS 07)*, pages 897–904, Cambridge, MA, 2008. MIT Press.

[10] T. Y. Liu, T. Qin, J. Xu, W. Y. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *LR4IR 2007, in conjunction with SIGIR 2007*, 2007.

[11] J. I. Marden, editor. *Analyzing and Modeling Rank Data*. Chapman and Hall, London, 1995.

[12] R. Nallapati. Discriminative models for information retrieval. In *Proc. of SIGIR'04*, pages 64–71, 2004.

[13] T. Qin, X.-D. Zhang, M.-F. Tsai, D.-S. Wang, T.-Y. Liu, and H. Li. Query-level loss functions for information retrieval. *Information processing and management*, 44:838–855, 2008.

[14] C. Rudin. Ranking with a p-norm push. In *Proc. of COLT*, pages 589–604, 2006.

[15] F. Xia, T. Y. Liu, J. Wang, W. S. Zhang, and H. Li. Listwise approach to learning to rank - theory and algorithm. In *Proc. of ICML'08*, pages 1192–1199, 2008.

[16] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.