Is Top-k Sufficient for Ranking?

Yanyan Lan
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R.China
lanyanyan@ict.ac.cn

Shuzi Niu
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R.China
niushuzi@software.ict.ac.cn

Jiafeng Guo
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R.China
guojiafeng@ict.ac.cn

Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R.China
cxq@ict.ac.cn

ABSTRACT

Recently, 'top-k learning to rank' has attracted much attention in the community of information retrieval. The motivation comes from the difficulty in obtaining a full-order ranking list for training, when employing reliable pairwise preference judgment. Inspired by the observation that users mainly care about top ranked search result, top-k learning to rank proposes to utilize top-k ground-truth for training, where only the total order of top k items are provided, instead of a full-order ranking list. However, it is not clear whether the underlying assumption holds, i.e. top-k ground-truth is sufficient for training. In this paper, we propose to study this problem from both empirical and theoretical aspects. Empirically, our experimental results on benchmark datasets LETOR4.0 show that the test performances of both pairwise and listwise ranking algorithms will quickly increase to a stable value, with the growth of k in the top-k ground-truth. Theoretically, we prove that the losses of these typical ranking algorithms in top-k setting are tighter upper bounds of (1–NDCG@k), compared with that in full-order setting. Therefore, our studies reveal that learning on top-k ground-truth is surely sufficient for ranking, which lay a foundation for the new learning to rank framework.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Performance, Experimentation, Theory

Keywords

Learning to Rank, Top-k, Full-Order, Sufficient

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505685.

1. INTRODUCTION

Learning to rank has become an important means to tackle the ranking problem in many applications, such as information retrieval, collaborative filtering and natural language processing. Taking Web search as an example, the process of learning to rank is as follows. In training, a number of queries are given, and each query associates with a number of documents and labels representing their rankings (usually in terms of multi-level ratings). Then a ranking function is constructed by minimizing a certain loss function on the training data. In testing, given a new query and associated documents, the ranking function is applied to produce a ranking list and the performance of the ranking algorithm is evaluated by IR measures such as MAP [3], NDCG [11]and ERR [7].

Recently, a new learning to rank framework named 'top-k learning to rank' has emerged and gain much attention. The motivation comes from the difficulty in obtaining reliable training data for applying learning to rank to real systems: (1) conventional multi-level ratings based training data are not reliable [16, 19, 20]; (2) when employing more reliable pairwise preference judgment, it would be prohibitively expensive to obtain a full-order ranking list [5, 6, 19]. Based on the fact that users mainly care more about top ranked search result, top-k learning to rank proposes to utilize top-k ground-truth for training (k is usually small), where only the full ordering of top k items are provided, instead of a full-order ranking list.

The underlying assumption of top-k learning to rank is that top-k ground-truth is sufficient for ranking, i.e. training on top-k ground-truth is as good as that on a full-order ranking list. On this basis, top-k learning to rank describes how to conduct labeling, ranking and evaluation process. However, it is unclear whether the assumption holds. In this paper, we propose to study this problem from both empirical and theoretical aspects.

Empirically, we proposed to conduct experiments to study how the test performances of ranking algorithms change with respect to k in the training data of top-k learning to rank. Intuitively, with the increase of k, more information is conveyed by the training data, and the test performances of ranking algorithms will increase. If k reaches the maximum (i.e. full-order ground-truth), we obtain the best test performance as all the ranking information is involved. Therefore,

if the test performances quickly increase to a stable value, we say that top-k ground-truth is sufficient for ranking.

We conduct extensive experiments based on benchmark data sets LETOR4.0, and consider three state-of-the-art pairwise algorithms, Ranking SVM [10], RankBoost [9] and RankNet [1], and a popular listwise algorithm, ListMLE [22]. We plot the test performance curves of these algorithms on two data sets in LETOR4.0, MQ2007-list and MQ2008-list. Our experimental results indeed show that the test performances of all the four algorithms increase quickly to a stable value with the increase of k. As a consequence, we haven proven empirically that top-k ground-truth is sufficient for ranking.

At a first glance, the theoretical analysis of 'whether top-k is sufficient for ranking' is to study the relationship between the loss functions of these algorithms in top-k setting and that in full-order setting. It is obvious that the former ones are lower bounds of the latter ones, which means that the minimization of these loss functions in full-order setting will lead to the minimization of them in top-k setting. However, what we really care about is the opposite side of the coin, i.e. whether the minimization of these loss functions in top-k setting will lead to the minimization of them in full-order setting. Seemingly the answer is negative.

By revisiting the problem of 'whether training on top-k ground-truth is as good as that on a full-order ranking list', we find that the theoretical analysis need to further take IR evaluation measures into consideration, due to the fact that the performances of ranking algorithms are usually evaluated by them. In this paper, we take NDCG as an example to conduct the theoretical analysis.

To reveal the relationships among the three, we define a loss function named Weighted Kendall's Tau (WKT for short). First, it can be proved that WKT is an upper bound of (1–NDCG@k). Second, it can be proved that the pairwise losses in Ranking SVM, RankBoost and RankNet, and the listwise loss in ListMLE are all upper bounds of WKT, in top-k setting. As a consequence, we come to the conclusion that the loss functions used in these ranking algorithms in top-k setting can bound (1–NDCG@k). Further considering the relationship between loss functions in full-order setting and that in top-k setting, we can see that loss functions in top-k setting are tighter bounds of (1–NDCG@k), as compared with those in full-order setting. Therefore, we have proven theoretically that top-k ground-truth is not only sufficient, but even better than full-order ground-truth for ranking.

According to the above empirical and theoretical study, we come to the conclusion that top-k ground-truth is sufficient for ranking, which lay a foundation for the new top-k learning to rank.

The reminder of the paper is organized as follows. In Section 2, we introduce some background on conventional learning to rank and top-k learning to rank. In Section 3, we describe some ranking algorithms both in top-k and full-order settings, including Ranking SVM, RankBoost, RankNet and ListMLE. Section 4 and Section 5 presents our experimental and theoretical analysis on whether top-k is sufficient for ranking, respectively. Section 6 concludes the paper.

2. BACKGROUNDS

In this section, we will introduce some backgrounds on conventional learning to rank and top-k learning to rank, respectively.

2.1 Conventional Learning to Rank

Taking Web search as an example, we describe the framework of conventional learning to rank as follows.

In the training process, a number of queries $q_1, q_2, \cdots, q_N$ are given. For each query q_i, we are given a set of associated documents $\mathbf{x}_i = (x_1^{(i)}, \cdots, x_{n_i}^{(i)})$ and their ground-truth labels $\mathbf{y}_i = (y_1^{(i)}, y_2^{(i)}, \cdots, y_{n_i}^{(i)})$, which are usually represented in the form of multi-level ratings, such as 3-level ratings (highly relevant:2, relevant:1, irrelevant:0).

With the training data, different ranking algorithms are proposed to conduct the learning process. According to the different objects considered in the loss functions, they are mainly divided into three categories: pointwise, pairwise and listwise approach. In pointwise approach [13], single items are viewed as the objects and ranking is transformed to regression on items to represent the absolute label on each item. In pairwise approach [1, 9, 10], item pairs are recognized as the objects and ranking is transformed to a pairwise classification problem to represent the preference between these two items. In listwise approach [4, 18, 22, 23], instances as document lists are taken as objects and ranking is transformed to a permutation level prediction problem.

In the testing process, for a query q_t, we are given its associated documents $\mathbf{x}_t = (x_1^{(t)}, x_2^{(t)}, \cdots, x_{n_t}^{(t)})$ and the ground-truth labels $\mathbf{y}_t = (y_1^{(t)}, y_2^{(t)}, \cdots, y_{n_t}^{(t)})$. State-of-the-art IR measures such as MAP [3], NDCG [11], ERR [7] are usually adopted to evaluate the performance of the learned ranking function.

2.2 Top-k Learning to Rank

Although conventional learning to rank techniques have been widely applied to many real applications, such as information retrieval and collaborative filtering, and made a great success, they are mainly criticized for depending on unreliable training data [2, 19, 20]. To address this problem, many researchers have proposed to utilize more reliable pairwise preference judgment as an alternative [6, 19, 20]. However, the complexity would be $O(n \log n)$ to construct a full-order ranking list with size n under the pairwise preference judgments [16, 19].

Based on the assumption that top-k ground-truth is sufficient for ranking, i.e. *training on top-k ground-truth is as good as that in full-order setting*, a new top-k learning to rank framework [16, 21] is proposed to utilize top-k ground-truth for training, instead of a full-order ranking list. Specifically, the top-k ground-truth is represented as a mixture of the total order of the top k items, and the relative preferences between the set of top k items and the set of the rest $n-k$ items. With the top-k ground-truth, new top-k ranking algorithms are proposed to facilitate the learning process. For example, Xia et.al [21] proposed to modify traditional listwise ranking algorithms such as ListMLE[22], ListNet [4] and RankCosine [18] to fit the top-k setting. Niu et.al [16] introduced a mixed model named FocusedRank, in which pairwise losses and listwise losses are employed to model the relative preference relationship and the total order relationship, respectively. In [17], a new probabilistic model based on the sequential generation process was proposed for the top-k ranking problem.

Since top-k learning to rank introduces a novel reliable training data construction method and the performances of top-k ranking methods have been shown more effective, it

has gained great attention recently[1]. However, it is not clear whether the underlying assumption that top-k ground-truth is sufficient for ranking is correct. In [21], the authors prove that listwise ranking algorithms such as ListMLE, ListNet and RankCosine are consistent with respect to a permutation level 0-1 loss in top-k setting. However, permutation level 0-1 loss is not appropriate for evaluation since it omits the position information, which is quite an important factor for the ranking problem [8, 14]. Therefore, the correctness of the underlying assumption remains an open question. In this paper, we propose to study the problem from both empirical and theoretical aspects.

3. RANKING ALGORITHMS

As mentioned above, the underlying assumption of top-k learning to rank is that training on top-k ground-truth is as good as that on a full-order ranking list. Therefore, to investigate the correctness of the assumption, we propose to conduct both empirical and theoretical analysis on ranking algorithms in both full-order and top-k settings. Before the analysis, we first introduce the precise forms of ranking algorithms in both full-order and top-k settings in this section. Specifically, four state-of-the-art ranking algorithms are utilized in this paper, including pairwise ranking algorithms such as Ranking SVM [10], RankBoost [9] and RankNet [1], and a listwise ranking algorithm ListMLE [22].

3.1 Algorithms in Full-Order Setting

In full-order setting, a full-order ranking list is utilized as the ground-truth. Therefore, we formulate the training data as $\{(q_i, \mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where q_i stands for a query, $\mathbf{x}_i = (x_1^{(i)}, ..., x_{n_i}^{(i)})$ stands for the n_i associated documents, and $\mathbf{y}_i$=$(y_1^{(i)}, ..., y_{n_i}^{(i)})$ stands for a full-order ranking list with $y_j^{(i)}$ denotes the index of the item ranked in the j-th position of $\mathbf{y}_i$. Please note that $x_{y_j^{(i)}}^{(i)}$ is simply denoted as $x_{y_j}^{(i)}$ hereafter.

3.1.1 Pairwise Algorithms in Full-Order Setting

Pairwise ranking algorithms utilize the losses on all the pairs as the training objective, therefore the training loss in the full-order setting can be formulated as follows.

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i-1} \sum_{l=j+1}^{n_i} L^p(f; x_{y_j}^{(i)}, x_{y_l}^{(i)}), \quad (1)$$

where L^p stands for a pairwise loss, such as the hinge loss used in Ranking SVM, the exponential loss used in RankBoost and the logistic loss used in RankNet.

(1) Ranking SVM in Full-Order Setting

Ranking SVM [10] utilizes the following hinge loss as the loss function, and applies the SVM technology to optimize the number of misclassified pairs respectively.

$$L_{hig}^p(f; x_{y_j}^{(i)}, x_{y_l}^{(i)}) = \max\{0, 1 - w^T(x_{y_j}^{(i)} - x_{y_l}^{(i)})\}. \quad (2)$$

Based on the total loss represented by Eq.(1), we formulated Ranking SVM in full-order setting as the following optimization probelm:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\sum_{j=1}^{n_i-1}\sum_{l=j+1}^{n_i}\xi_{j,l}^{(i)}$$
$$s.t.: w^T x_j^{(i)} - w^T x_l^{(i)} \geq 1 - \xi_{j,l}^{(i)},$$
$$\xi_{j,l}^{(i)} \geq 0, i = 1, ..., m; \forall j = 1, ..., n_i - 1; l = j+1, ..., n_i,$$

where $\frac{1}{2}\|w\|^2$ controls the complexity of the model w, and C is a trade-off parameter between the model complexity and hinge loss relaxations.

(2) RankBoost in Full-Order Setting

RankBoost [9] adopts the boosting technology to output a ranking model by combining the week rankers, where the combination coefficients are determined by the probability distribution on document pairs.

Based on the total loss represented by Eq.(1) and the exponential loss presented as follows, we give the detailed algorithm for RankBoost in full-order setting, as shown in Alg.1.

$$L_{exp}^p(f; x_{y_j}^{(i)}, x_{y_l}^{(i)}) = \exp\left(-(w^T(x_{y_j}^{(i)} - x_{y_l}^{(i)}))\right), j < l \quad (3)$$

(3) RankNet in Full-Order Setting

RankNet aims to optimize a cross entropy between the target probability and the modeled probability, where the probability is defined based on the exponential function of difference between the scores of any two documents in all document pairs given by the scoring function f. The loss function in full-order setting is presented as follows.

$$L_{log}^p(f; x_{y_j}^{(i)}, x_{y_l}^{(i)}) = -\bar{P}_{jl}\log P_{jl}(f) - (1-\bar{P}_{jl})\log(1-P_{jl}(f)). \quad (4)$$

where, $\bar{P}_{jl} = 1$, if $j < l$, and $\bar{P}_{jl} = 0$, otherwise.

$$P_{jl}(f) = \frac{\exp(w^T x_{y_j}^{(i)} - w^T x_{y_l}^{(i)})}{1 + \exp(w^T x_{y_j}^{(i)} - w^T x_{y_l}^{(v)})},$$

3.1.2 Listwise Algorithms in Full-Order Setting

In this paper, we use ListMLE [22] as an example of listwise ranking algorithms due to its nice empirical and theoretical properties [8]. ListMLE models the generation of a ranking list according to Plackett-Luce Model [15], and utilizes the following likelihood loss for training.

$$L^l(f; \mathbf{x}, \mathbf{y}) = -\log P(\mathbf{y}|\mathbf{x}, f), \quad (5)$$

where $P(\mathbf{y}|\mathbf{x})$ is defined as:

$$P(\mathbf{y}|\mathbf{x}, f) = \prod_{j=1}^{n-1} \frac{\exp\{f(x_{y_j})\}}{\sum_{l=j}^{n}\exp\{f(x_{y_l})\}}.$$

Alg.1: Learning Algorithm for RankBoost in Full-Order Setting

Input: training data in terms of full-order ground-truth.
Given: initial distribution D_i on all the pairs of q_i, i=1,$\cdots$,N.
For $t = 1, \cdots, T$
  train weak ranker f_t to minimize:
    $r_t = \sum_{i=1}^{N}\sum_{j=1}^{n_i-1}\sum_{l=j+1}^{n_i} D_t(x_{y_j}^{(i)}, x_{y_l}^{(i)}) L_{exp}^p(f; x_{y_j}^{(i)}, x_{y_l}^{(i)})$.
  choose $\alpha_t = \frac{1}{2}\log(\frac{1+r_t}{1-r_t})$.
  update
    $D_{t+1}(x_{y_j}^{(i)}, x_{y_l}^{(i)}) = \frac{1}{Z_t} D_t(x_{y_j}^{(i)}, x_{y_l}^{(i)}) \exp(\alpha_t(w^T x_{y_j}^{(i)} - w^T x_{y_l}^{(i)}))$,
    where,
    $Z_t = \sum_{i=1}^{N}\sum_{j=1}^{n_i-1}\sum_{l=j+1}^{n_i} D_t(x_{y_j}^{(i)}, x_{y_l}^{(i)}) \exp(\alpha_t(w^T x_{y_j}^{(i)} - w^T x_{y_l}^{(i)}))$.
Output: $f(x) = \sum_t \alpha_t f_t(x)$.

[1]Please note that [16] has won the best student paper award of SIGIR2012.

On this basis, the total loss on the training data $\{(q_i, \mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ is represented as follows.

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i-1} \{-f(x_{y_j}^{(i)}) + \log(\sum_{l=j}^{n_i} \exp\{f(x_{y_l^{(i)}})\})\}. \quad (6)$$

3.2 Algorithms in Top-k Setting

According to the definition of top-k ground-truth that only the total order of top k items are given, we can formulate the training data in top-k setting as $(q_i, \mathbf{x}_i, Y_k^{(i)}), i = 1, \cdots, N$, where q_i stands for a query, $\mathbf{x}_i$ stands for the n_i associated documents, $Y_k^{(i)}$ stands for a set of full-order ranking lists with size n_i, such that the total orders of the top k items in these ranking lists are the same while the remaining $n_i - k$ items form different permutations.

3.2.1 Pairwise Algorithms in Top-k Setting

Pairwise algorithms are proposed to utilize all the pairs constructed from training data as the total loss for optimization. Therefore, with top-k ground-truth, pairs constructed between the last $n-k$ items will no longer exist in the total loss, different from the setting of full ordering ground-truth. Furthermore, according to the definition of top-k ground-truth, for any $\mathbf{y} \in Y_k$, the following loss is the same.

$$\sum_{i=1}^{N} \sum_{j=1}^{k} \sum_{l=j+1}^{n_i} L^p(f; x_{y_j}^{(i)}, x_{y_l}^{(i)})$$

Therefore, the total loss of a pairwise ranking algorithm in top-k setting can be represented as follows.

$$\sum_{i=1}^{N} \min_{\mathbf{y} \in Y_k^{(i)}} \sum_{j=1}^{k} \sum_{l=j+1}^{n_i} L^p(f; x_{y_j}^{(i)}, x_{y_l}^{(i)}). \quad (7)$$

Incorporating different loss functions L^p as described in Eq.(2), Eq.(3) and Eq.(4) into the above equation, we can obtain the total loss of Ranking SVM, RankBoost and RankNet in top-k setting. Since the optimization processes are the same as that in full-order setting, we omit them here for clear representation.

3.2.2 Listwise Algorithms in Top-k Setting

As described in [21], the top-k ground-truth with respect to a full-order ranking list $\mathbf{y}$ is generated according the probability as follows.

$$\prod_{j=1}^{k} \frac{\exp\{f(x_{y_j})\}}{\sum_{l=j}^{n_i} \exp\{f(x_{y_l})\}}.$$

Similarly to the above analysis for pairwise algorithms, for each $\mathbf{y} \in Y_k$, the above probability is the same. Therefore, the probability of top-k ground-truth can be represented as follows.

$$P(Y_k|\mathbf{x}, f) = \min_{\mathbf{y} \in Y_k} \prod_{j=1}^{k} \frac{\exp\{f(x_{y_j})\}}{\sum_{l=j}^{n_i} \exp\{f(x_{y_l})\}}.$$

As a consequence, the total loss in top-k setting can be formulated as the following form.

$$\sum_{i=1}^{N} \sum_{j=1}^{k} \min_{\mathbf{y}_i \in Y_k^{(i)}} \{-f(x_{y_j}^{(i)}) + \log(\sum_{l=j}^{k} \exp\{f(x_{y_l}^{(i)})\})\}. \quad (8)$$

4. EMPIRICAL ANALYSIS

In this section, we propose to empirically study whether the assumption that top-k ground-truth is sufficient for ranking holds. As described in Section 2, the assumption can be formulated as training in top-k setting is as good as that in full-order setting. Therefore, the empirical study can be conducted by comparing the test performances of ranking algorithms in top-k setting and that in full-order setting.

4.1 Experimental Settings

We conduct extensive experiments on the benchmark datasets LETOR4.0[2]. In LETOR4.0, the ground-truth for a query is a full-order ranking list in MQ2007-list and MQ2008-list. Therefore, it is easy to construct top-k ground-truth by just preserving the total order of top k items. As a result, these two datasets are suitable for our study in both top-k and full-order setting. The four learning to rank algorithms as mentioned above are all included in our experiments, including pairwise ranking algorithms such as Ranking SVM, RankBoost and RankNet and a listwise ranking algorithm ListMLE. The training set, validation set and test set have already been divided in LETOR4.0 and we follow the default setting of LETOR 4.0 in our experiments.

In order to compare the test performances of ranking algorithms in top-k setting and that in full-order setting, we conduct training process for each ranking algorithm with top-k ground-truth. It is obvious to see that, when k equals to the total number of documents for each query, the top-k setting becomes the full-order setting. For each k, parameters are selected through the validation set. For example, in RankSVM, the tradeoff parameter C is tuned from $\{10^{-5}, 10^{-4}, \cdots, 10^{-1}, 0.2, \cdots, 1, 10, 100, 1000\}$. In RankBoost, the relative loss variation between two iterations are chosen from 0.1 to 10^{-6} to control the stop condition, and the maximal number of iterations is set to 500. For gradient descent procedures as in RankNet and ListMLE, the learning rate is selected from $\{10^{-5}, 10^{-4}, \cdots, 10^{-1}\}$ with the maximal number of iterations 500.

Finally, the performances on test set with selected parameters is evaluated with the full-order ground-truth, using NDCG as the evaluation measure.

4.2 Experimental Results

We plot the performance curves of different ranking algorithms with the increase of k in terms of NDCG@5 and NDCG@10, as shown in Figure 1 and Figure 2. For each ranking algorithm, the top sub-figure stands for the overall test performance curves, with k varies from 1 to 1000, and the value in full-order setting is plotted as the rightmost point in the figure. The bottom sub-figure stands for the test performance curves with k varies from 1 to 100.

From the results in Figure 1 and 2, we can see that:

(1) Overall, the test performances of ranking algorithms in top-k setting increase to a stable value with the growth of k. This can be clearly illustrated by the experimental results on MQ2008-list as shown in Figure 2. In general, the experimental results on MQ2007-list also agree with the claim. However, it shows in the figure that when k keeps increasing, the performances will decrease. For example, the performances of Ranking SVM, RankNet and ListMLE

[2]http://research.microsoft.com/en-us/um/beijing/projects/letor//

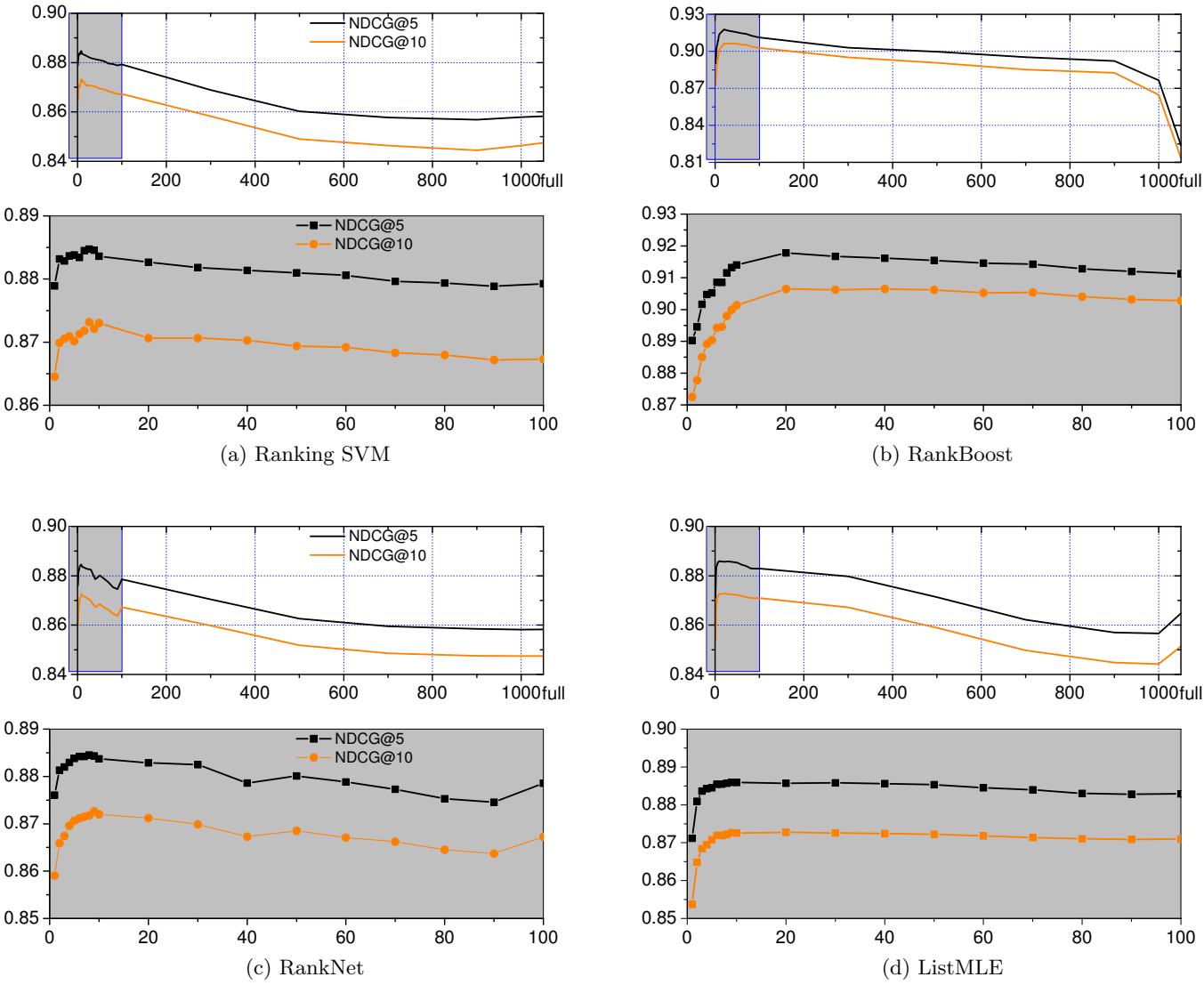


Figure 1: Performance variations of different ranking algorithms in top-k setting on MQ2007-list with the increase of k

will drop when k exceeds 100. More surprisingly, the performances of these algorithms in full-order setting are even not comparable with that in top-1 setting. At a first glance, this experimental results seem deviate from our intuition that the performance should be better with larger k since more information is conveyed in the training data. However, the results can be explained since there are usually many noises in real training data, especially for labels on the tail documents in a full-order setting. For example, in MQ2007-list, there are 1700 queries and 700 documents per query on average. Therefore, it is difficult to obtain a reliable full-order ground-truth with respect to such a large data set, especially to obtain reliable orders among the tail documents in a list. That is why ranking algorithms in full-order setting perform so badly.

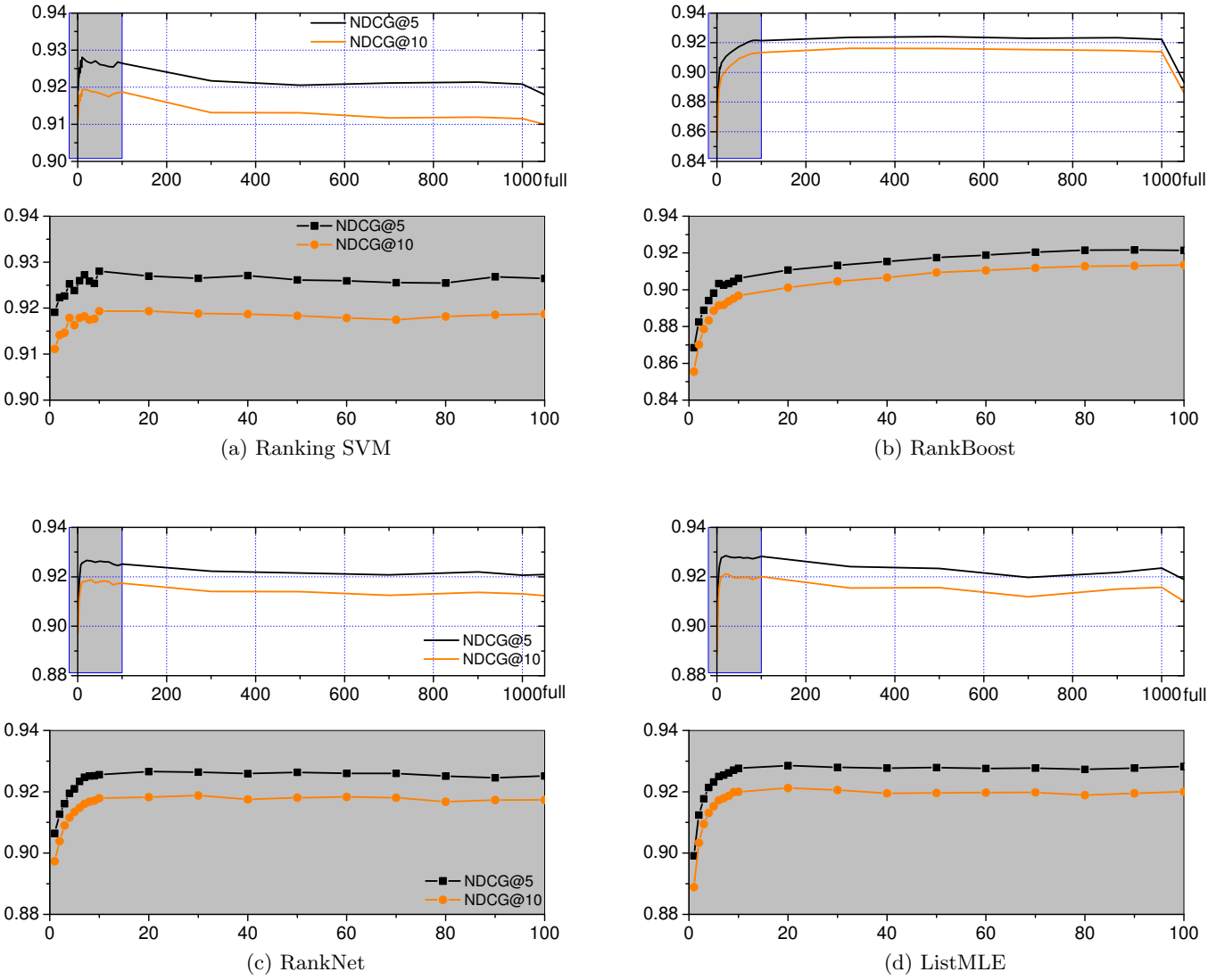
(2) By carefully looking into the variance of curves with k varying from 1 to 100 as shown in the bottom sub-figure of each ranking algorithm, we can see that the test performances of all the four ranking algorithms increase quickly to a stable value with the increase of k. For example, when k exceeds 10 in MQ2007-list and MQ2008-list, the performances of Ranking SVM, RankNet and ListMLE keep stable. when k exceeds 20, the performance of RankBoost also keeps stable.

In summary, our experimental results on benchmark datasets LETOR4.0 show that the test performances quickly increase to a stable value with the growth of k in the training data. Therefore, we have empirically proven that top-k ground-truth is sufficient for ranking.

5. THEORETICAL ANALYSIS

In this section, we propose to study whether the assumption of top-k is sufficient for ranking holds from theory aspect.

Firstly, we formalize the problem as finding the relationships among losses in top-k setting, losses in full-order set-



(a) Ranking SVM (b) RankBoost

(c) RankNet (d) ListMLE

Figure 2: Performance variation of different ranking algorithms in top-k setting on MQ2008-list with the increase of k

ting and IR evaluation measures, theoretically. Inspired by the technique used in [8], we propose a new loss function named Weighted Kendall's Tau (WKT for short), and then deduce the concerned relationships by finding the relationship between WKT and losses in top-k setting, and the relationship between WKT and IR evaluation meeasures, respectively.

5.1 Problem Formalization

As described in Section 2, the assumption can be formulated as training in top-k setting is as good as that in full-order setting. At a first glance, it seems natural to formalize the problem (i.e. whether the assumption holds) as finding the relationship between losses in top-k setting and that in full-order setting. Based on the formulations of total losses in top-k setting and that in full-order setting, we can obtain the following relationships.

(1) The pairwise loss functions in full-order setting are upper bounds of that in top-k setting, described as follows.

$$\sum_{i=1}^{N} \min_{\mathbf{y}_i \in Y_k^{(i)}} \sum_{j=1}^{k} \sum_{l=j+1}^{n_i} L^p(f; x_{y_j}^{(i)}, x_{y_l}^{(i)}) \leq \sum_{i=1}^{N} \sum_{j=1}^{n_i-1} \sum_{l=j+1}^{n_i} L^p(f; x_{y_j}^{(i)}, x_{y_l}^{(i)}),$$

(2) The listwise loss function of ListMLE in full-order setting are upper bounds of that in top-k setting, described as follows.

$$\sum_{i=1}^{N} \min_{\mathbf{y}_i \in Y_k^{(i)}} \sum_{j=1}^{k} \{-f(x_{y_j}^{(i)}) + \log(\sum_{l=j}^{n_i} \exp\{f(x_{y_l^{(i)}})\})\} \leq \sum_{i=1}^{N} \sum_{j=1}^{n_i-1} \{-f(x_{y_j}^{(i)}) + \log(\sum_{l=j}^{n_i} \exp\{f(x_{y_l^{(i)}})\})\}.$$

These relationships mean that the minimization of the former (i.e. loss functions in full-order setting) will lead to the minimization of the latter (i.e. loss functions in top-k setting). However, what we really care about is the opposite side of the coin, i.e. whether the minimization of these loss functions in top-k setting will lead to the minimization of them in full-order setting. Seemingly the answer is negative.

Inspired by the contradiction between the negative answer and the claim of top-k learning to rank, we need to revisit the assumption: training on top-k ground-truth is as good as that on a full-order ranking list. We can see that the term 'as good as' is actually related to the evaluation procedure in ranking. According to the fact that a ranking algorithm is usually evaluated by IR evaluation measures in learning to rank, we find that the theoretical analysis on whether the assumption holds need to further take IR evaluation measures into consideration.

Therefore, a more appropriate formalization of the theoretical problem is to study *the relationships among losses in top-k setting, losses in full-order setting and IR evaluation measures*. In this paper, we take NDCG as an example to conduct the theoretical analysis, and leave the analysis on other measures such as MAP, ERR in the future work.

5.2 Theoretical Results

As described above, a reasonable theoretical formalization is to study the relationships between the losses in full-order setting, losses in top-k setting and NDCG. Since the relationships between losses in full-order setting and that in top-k setting has been revealed in the last subsection, we focus on the relationship between losses in top-k setting and NDCG here.

Firstly, we give the precise definition of NDCG as follows.

$$NDCG@k(f,\mathbf{x},\mathbf{y}) = \frac{1}{N_k}\sum_{j=1}^{k} g(l(y_j))D(r_j), \tag{9}$$

where $\mathbf{x}$ and $\mathbf{y}$ stands for the set of documents and the corresponding full-order ground-truth, respectively. r_j stands for the rank of x_j in the ranking list obtained by f. In the evaluation[16], the position information is transformed to labels for the computation of NDCG as $l(y_j) = n - y_j$. $g(l(y_j))$ is the gain function with $g(l(y_j)) = 2^{l(y_j)} - 1$, $D(r_j)$ is the discount function with $D(r_j) = \frac{1}{log(1+r_j)}$, and N_k stands for the maximum of $\sum_{j=1}^{K} g(y_j)D(r_j)$.

Here we also list the pairwise losses and listwise loss in top-k setting to make theoretical analysis easier.

$$L^p(f;\mathbf{x},Y_k)=\min_{\mathbf{y}\in Y_k}\sum_{j=1}^{k}\sum_{l=j+1}^{n} L^p(f;x_{y_j},x_{y_l}). \tag{10}$$

$$L^l(f;\mathbf{x},Y_k)=\min_{\mathbf{y}\in Y_k}\sum_{j=1}^{k}\{-f(x_{y_j})+\log(\sum_{l=j}^{n}\exp\{f(x_{y_l})\})\}. \tag{11}$$

5.2.1 Weighted Kendall's Tau

Inspired by [8] and [12], we propose a new loss named Weighted Kendall's Tau (WKT) to facilitate the theoretical analysis, which is defined as follows.

$$L_\alpha(f;\mathbf{x},Y_k)=\min_{\mathbf{y}\in Y_k}\sum_{j=1}^{k}\alpha(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j})-f(x_{y_l})<0\}}, \tag{12}$$

where $\alpha(\cdot)$ is a decreasing function to represent the importance of position information, and $I_{\{\cdot\}}$ is the indicator function with $I_A = 1$ if A is true, otherwise $I_A = 0$.

Weighted Kendall's Tau has a nice property that for any full-order ranking list in which the top-k items are consistent with that in top-k ground-truth, the loss is the same, described as the following lemma. The property makes WKT easier to relate NDCG to losses in top-k ground-truth.

LEMMA 1. *For any set of items $\mathbf{x}$, given the full-order ground-truth $\mathbf{y}$ and the top-k ground-truth Y_k, for any ranking function f, the following equalities hold,*

$$L_\alpha(f;\mathbf{x},Y_k)=\sum_{j=1}^{k}\alpha(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_i})-f(x_{y_j})<0\}}, \tag{13}$$

PROOF. Firstly, from the definition of the top-k ground-truth as a set of full-order ranking lists where the total orders of the top k items are the same, we can see that $\mathbf{y} \in Y_k$.

Secondly, we prove that for $\forall \mathbf{y}_1, \mathbf{y}_2 \in Y_k$, the following equality holds.

$$\sum_{j=1}^{k}\alpha(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j^{(1)}})-f(x_{y_l^{(1)}})<0\}} = \sum_{j=1}^{k}\alpha(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j^{(2)}})-f(x_{y_l^{(2)}})<0\}}.$$

According to the definition of top-k ground-truth Y_k, the orders of top k items in $\mathbf{y}_1$ and $\mathbf{y}_2$ are the same. Therefore, the following statement holds.

$$y_j^{(1)} = y_j^{(2)}, j = 1, \cdots, k.$$

As a consequence, we have

$$\begin{aligned}&\sum_{j=1}^{k}\alpha(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j^{(1)}})-f(x_{y_j^{(1)}})<0\}}\\ =&\sum_{j=1}^{k}\alpha(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j^{(2)}})-f(x_{y_j^{(1)}})<0\}}\\ =&\sum_{j=1}^{k}\alpha(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j^{(2)}})-f(x_{y_j^{(2)}})<0\}}.\end{aligned}$$

Combining the above results, we can obtain that:

$$\begin{aligned}L_\alpha(f;\mathbf{x},Y_k) &= \min_{\mathbf{y}\in Y_k}\sum_{j=1}^{k}\alpha(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j})-f(x_{y_l})<0\}}\\ &= \sum_{j=1}^{k}\alpha(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_i})-f(x_{y_j})<0\}}.\end{aligned}$$

Therefore, we have proven the results in the lemma. □

5.2.2 WKT: Upper Bound of Measure Based Error

First, we study the relationship between Weighted Kendall's Tau and (1−NDCG). It can be proven that WKT is an upper bound of (1–NDCG@k), as shown in the following theorem.

THEOREM 1. *For any set of items $\mathbf{x}$, given the full ordering ground-truth $\mathbf{y}$ and the top-k ground-truth Y_k, for any ranking function f, the following inequality holds,*

$$1 - NDCG@k(f;\mathbf{x},\mathbf{y}) \le \frac{1}{N_k}L_\alpha(f;\mathbf{x},Y_k), \tag{14}$$

where $\alpha(j) = G(l(y_j))D(j)$.

PROOF. First, we formulate $NDCG$ as follows,

$$NDCG@k(f;\mathbf{x},\mathbf{y}) = \frac{1}{N_k}\sum_{j=1}^{n} G(l(y_j))D(r_j).$$

According to the definition of N_k, we have,

$$N_k = \sum_{j=1}^{k} G(l(y_j))D(j).$$

Therefore, we have,

$$\begin{aligned} &1-NDCG@k(f;\mathbf{x},\mathbf{y}) \\ &= \frac{1}{N_k}\sum_{j=1}^{k} G(l(y_j))(D(j)-D(r_j)). \end{aligned} \tag{15}$$

Second, we consider the Weighted Kendall's Tau case by case. Let $\alpha(j) = G(l(\mathbf{y}_j))D(j)$, note that

$$L_\alpha(f;\mathbf{x},Y_k) = \sum_{j=1}^{k} G(l(y_j))D(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j})-f(x_{y_l})<0\}}.$$

1. If $\sum_{l=j+1}^{n} I_{\{f(x_{y_j})-f(x_{y_l})<0\}} = 0$, we have

$$I_{\{f(x_{y_j})-f(x_{y_l})<0\}} = 0, \forall l \in [j+1, n].$$

That is, $\forall l \in [j+1, n], f(x_{y_j}) > f(x_{y_l})$. Immediately, we have $r_j \le j$. Since $D(\cdot)$ is a decreasing function, we have $D(j) \le D(r_j)$. We immediately obtain that,

$$D(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j})-f(x_{y_l})<0\}} \ge D(j) - D(r_j).$$

2. Otherwise, at least $\exists l_0 \in [j+1, n]$, such that

$$I_{\{f(x_{y_j})-f(x_{y_l})<0\}} = 1,$$

Therefore we can obtain that,

$$D(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j})-f(x_{y_l})<0\}} \ge D(j) - D(r_j).$$

Combining the above results, we can obtain that,

$$D(j)\sum_{l=j+1}^{n} I_{\{f(x_{y_j})-f(x_{y_l})<0\}} \ge D(j) - D(r_j).$$

Further considering Eq.(15), we have

$$1 - NDCG@k(f;\mathbf{x},\mathbf{y}) \le \frac{1}{N_k} L_\alpha(f;\mathbf{x},\mathbf{y}). \tag{16}$$

Considering the result of Lemma 1, we have proven that the inequality in the theorem holds. □

5.2.3 WKT: Lower Bound of Loss Functions

Here we study the relationship between Weighted Kendall's Tau and loss functions of ranking algorithms in top-k setting. It can be proven that WKT is a lower bound of loss functions of four ranking algorithms, including pairwise loss in Ranking SVM, RankBoost and RankNet, and a listwise loss in ListMLE, as shown in the following theorem.

THEOREM 2. *For any set of items* $\mathbf{x}$*, given the top-k ground-truth* Y_k*, for any ranking function* f*,*

(1) Weighted Kendall's Tau L_α *is the lower bound of the pairwise loss function* $L^p(f;\mathbf{x},Y_k)$*, as shown in the following inequality:*

$$L_\alpha(f;\mathbf{x},Y_k) \le (\max_{1\le i\le k}\alpha(i))L^p(f;\mathbf{x},Y_k); \tag{17}$$

(2) Weighted Kendall's Tau L_α *is the lower bound of the listwise loss function* $L^l(f;\mathbf{x},Y_k)$*, as shown in the following inequality:*

$$L_\alpha(f;\mathbf{x},Y_k) \le \frac{1}{\ln 2}(\max_{1\le i\le k}\alpha(i))L^l(f;\mathbf{x},Y_k); \tag{18}$$

PROOF. We first prove Eq.(17).

From the definition of hinge loss, exponential loss and logistic loss in Eq.(2), Eq.(3) and Eq.(4), the pairwise loss function can all be represented as a function $\phi(\cdot)$ with $\phi(0) = 1$, as shown below.

$$L^p(f;x_{y_j},x_{y_l}) = \phi(f(x_{y_j}) - f(x_{y_l})).$$

Therefore, the following equation holds.

$$I_{\{f(x_{y_j})-f(x_{y_l})<0\}} \le \phi(f(x_{y_j}) - f(x_{y_l})).$$

Therefore, it is obvious that Eq.(17) holds.

Now we prove Eq.(18). Note that

$$L^l(f;\mathbf{x},Y_k) = \sum_{i=1}^{k}(-f(x_{y_j}) + \ln(\sum_{j=i}^{n}(\exp(f(x_{y_l}))))).$$

1. If $\sum_{l=j+1}^{n} I_{\{f(x_{y_j})-f(x_{y_l})<0\}} = 0$, we have

$$\begin{aligned} L_\alpha(f;\mathbf{x},Y_k) &\le (\max_{1\le j\le k}\alpha(j))L^l(f;\mathbf{x},Y_k) \\ &\le \frac{1}{\ln 2}(\max_{1\le j\le k}\alpha(j))L^l(f;\mathbf{x},Y_k). \end{aligned}$$

2. Otherwise, at least $\exists l_0 \in [j+1, n]$, such that

$$I_{\{f(x_{y_j})-f(x_{y_l})<0\}} = 1,$$

that is, $f(x_{y_j}) < f(x_{y_l})$. Then it is obvious that,

$$\sum_{l=i}^{n}\exp(f(x_{\mathbf{y}(l)})) \ge 2\exp(f(x_{\mathbf{y}(j)})),$$

then the following inequality holds:

$$-f(x_{\mathbf{y}(j)}) + \ln(\sum_{l=j}^{n}(\exp(f(x_{y_l})))) \ge \ln 2.$$

Therefore,

$$L_\alpha(f;\mathbf{x},Y_k) \le \frac{1}{\ln 2}(\max_{1\le j\le k}\alpha(j))L^l(f;\mathbf{x},Y_k).$$

Combining the above results, we can obtain that,

$$L_\alpha(f;\mathbf{x},Y_k) \le \frac{1}{\ln 2}(\max_{1\le j\le k}\alpha(j))L^l(f;\mathbf{x},Y_k).$$

Therefore, we have proven that Eq.(18) holds. □

5.3 Results Analysis

Based on the above theorems, the theoretical results can be summarized as follows:

(1) Weighted Kendall's Tau is an upper bound of $(1-\text{NDCG@}k)$, as described below with $\alpha(j) = G(l(y_j))D(j)$.

$$1 - NDCG@k(f; \mathbf{x}, \mathbf{y}) \leq \frac{1}{N_k} L_\alpha(f; \mathbf{x}, Y_k),$$

(2) Weighted Kendall's Tau is a lower bound of loss functions in top-k setting, as described below.

$$L_\alpha(f; \mathbf{x}, Y_k) \leq (\max_{1 \leq i \leq k} \alpha(i)) L^p(f; \mathbf{x}, Y_k);$$

$$L_\alpha(f; \mathbf{x}, Y_k) \leq \frac{1}{\ln 2} (\max_{1 \leq i \leq k} \alpha(i)) L^l(f; \mathbf{x}, Y_k);$$

Based on these results, we immediately obtain that loss functions in top-k setting are upper bounds of (1-NDCG@k), described as below with $\alpha(j) = G(l(y_j))D(j)$.

$$1 - NDCG@k(f; \mathbf{x}, \mathbf{y}) \leq \frac{1}{N_k} (\max_{1 \leq i \leq k} \alpha(i)) L^p(f; \mathbf{x}, Y_k),$$

$$1 - NDCG@k(f; \mathbf{x}, \mathbf{y}) \leq \frac{1}{N_k} \frac{1}{\ln 2} (\max_{1 \leq i \leq k} \alpha(i)) L^l(f; \mathbf{x}, Y_k),$$

Further considering the relationship between loss functions in top-k setting and that in full-order setting, we can obtain the relationships of the three: *loss functions in top-k setting are tighter lower bounds of* (1−*NDCG@k*)*, as compared with loss functions in full-order setting.*

From this theoretical result, we can see that, if a ranking algorithm such as Ranking SVM, RankBoost, RankNet and ListMLE in full-order setting performs good, then the same ranking algorithm in top-k setting will definitely perform better. Therefore, we have proven theoretically that top-k ground-truth is sufficient for ranking. This is also in accordance with our experimental finding that the test performances of algorithms in top-k setting (when k reaches certain value) are better than that in full-order setting.

6. CONCLUSION

This paper addresses the problem of whether the underlying assumption of top-k learning to rank holds from both empirical and theoretical aspects.

(1) Empirically, we propose to check the variance of test performance curves of ranking algorithms with respect to k in top-k ground-truth to study the problem. For this purpose, we conduct extensive experiments on benchmark datasets LETOR4.0 with pairwise ranking algorithms Ranking SVM, RankBoost and RankNet, and a listwise ranking algorithm ListMLE. The results show that the test performances of all the four algorithms quickly increase to a stable value with the growth of k. As a consequence, we have proven empirically that top-k ground-truth is sufficient for ranking.

(2) Theoretically, we formulate the problem as the study of the relationships among loss functions in full-order setting, loss functions in top-k setting and IR evaluation measures such as NDCG. Firstly, it is obvious that loss functions in top-k setting are lower bound of that in full-order setting. Secondly, through a newly defined loss function named Weighted Kendall's Tau, we prove that $(1-\text{NDCG@}k)$ is a lower bound of losses in top-k setting. Therefore, loss functions in top-k settings are tighter lower bounds of $(1-\text{NDCG@}k)$, as compared to that in full-order setting. In other words, we have proven theoretically that top-k ground-truth is sufficient for ranking.

In summary, our analysis have proven the correctness of the assumption of the top-k learning to rank and lay a foundation for the new learning to rank framework.

There are still many issues need further investigation. For example, in this paper, we conduct theoretical analysis based on the relationship between different objectives. It would also makes sense to conduct statistical consistency analysis between algorithms in top-k setting and that in full-order setting.

Acknowledgments

This research work was funded by the National Natural Science Foundation of China under Grant No. 61003166 , No. 61203298 , No. 61232010, 973 Program of China under Grants No. 2012CB316303, No. 2013CB329602, and National Key Technology R&D Program under Grants No. 2012BAH46B04, No. 2012BAH39B02.

7. REFERENCES

[1] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 89–96, New York, NY, USA, 2005. ACM.

[2] R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing and Management*, 28(5):619–627, July 1992.

[3] C. Burkley and E. M. Voorhees. *Retrieval System Evaluation*. MIT Press, 2005.

[4] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 129–136, New York, NY, USA, 2007. ACM.

[5] B. Carterette and P. N. Bennett. Evaluation measures for preference judgments. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 685–686, New York, NY, USA, 2008. ACM.

[6] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: preference judgments for relevance. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 16–27, Berlin, Heidelberg, 2008. Springer-Verlag.

[7] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[8] W. Chen, T.-Y. Liu, Y. Lan, Z.-M. Ma, and H. Li. Ranking measures and loss functions in learning to rank. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances*

in Neural Information Processing Systems 22, pages 315–323. 2009.

[9] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, Dec. 2003.

[10] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 115–132, Cambridge, MA, 2000. MIT Press.

[11] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, Oct. 2002.

[12] Y. Lan, J. Guo, X. Cheng, and T.-Y. Liu. Statistical consistency of ranking methods in a rank-differentiable probability space. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1241–1249. 2012.

[13] P. Li, C. Burges, and Q. Wu. *McRank: Learning to Rank Using Multiple Classification and Gradient Boosting*. MIT Press, Cambridge, MA, 2008.

[14] T.-Y. Liu. Learning to rank for information retrieval. *Foundation and Trends on Information Retrieval*, 3:225–331, 2009.

[15] R. D. Luce. *Individual Choice Behavior*. Wiley, New York, 1959.

[16] S. Niu, J. Guo, Y. Lan, and X. Cheng. Top-k learning to rank: labeling, ranking and evaluation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 751–760, New York, NY, USA, 2012. ACM.

[17] S. Niu, Y. Lan, J. Guo, and X. Cheng. A new probabilistic model for top-k ranking problem. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2519–2522, New York, NY, USA, 2012. ACM.

[18] 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(2), 2008.

[19] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 239–248, New York, NY, USA, 2005. ACM.

[20] R. Song, Q. Guo, R. Zhang, and X. Guo. Select-the-best-ones: A new way to judge relative relevance. *Information Processing and Management*, 47:37–52, 2011.

[21] F. Xia, T.-Y. Liu, and H. Li. Statistical consistency of top-k ranking. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2098–2106. 2009.

[22] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 1192–1199, New York, NY, USA, 2008. ACM.

[23] J. Xu and H. Li. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 391–398, New York, NY, USA, 2007. ACM.