JMLR: Workshop and Conference Proceedings vol (2013) 1–30

A Theoretical Analysis of NDCG Ranking Measures

Yining Wang ANTONIOWYN@GMAIL.COM
The Institute for Theoretical Computer Science (ITCS), Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China

Liwei Wang WANGLW@CIS.PKU.EDU.CN
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing, China

Yuanzhi Li INVINCIBLEC.LEE@GMAIL.COM
The Institute for Theoretical Computer Science (ITCS), Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China

Di He WOLFYINK@GMAIL.COM
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing, China

Wei Chen WCHE@MICROSOFT.COM
Microsoft Research Asia, Beijing, China

Tie-Yan Liu TYLIU@MICROSOFT.COM
Microsoft Research Asia, Beijing, China

Abstract

A central problem in ranking is to design a measure for evaluation of ranking functions. In this paper we study, from a theoretical perspective, the Normalized Discounted Cumulative Gain (NDCG) which is a family of ranking measures widely used in practice. Although there are extensive empirical studies of the NDCG family, little is known about its theoretical properties. We first show that, whatever the ranking function is, the standard NDCG which adopts a logarithmic discount, converges to 1 as the number of items to rank goes to infinity. On the first sight, this result seems to imply that the standard NDCG cannot differentiate good and bad ranking functions on large datasets, contradicting to its empirical success in many applications. In order to have a deeper understanding of the general NDCG ranking measures, we propose a notion referred to as *consistent distinguishability*. This notion captures the intuition that a ranking measure should have such a property: For every pair of substantially different ranking functions, the ranking measure can decide which one is better in a *consistent* manner on almost all datasets. We show that standard NDCG has consistent distinguishability although it converges to the same limit for all ranking functions. We next characterize the set of all feasible discount functions for NDCG according to the concept of consistent distinguishability. Specifically we show that whether an NDCG measure has consistent distinguishability depends on how fast the discount decays; and r^{-1} is a critical point. We then turn to the cut-off version of NDCG, i.e., NDCG@k. We analyze the distinguishability of NDCG@k for various choices of k and the discount functions. Experimental results on real Web search datasets agree well with the theory.

Keywords: Ranking; Ranking measures; NDCG; Consistent Distinguishability

1. Introduction

Ranking has been extensively studied in information retrieval, machine learning and statistics. It plays a central role in various applications such as search engine, recommendation system, expert finding, to name a few. In many situations one wants to have, by learning, a good ranking function (Crammer and Singer, 2002; Freund et al., 2003; Joachims, 2002). Thus a fundamental problem is how to design a ranking measure to evaluate the performance of a ranking function.

Unlike classification and regression for which there are simple and natural performance measures, evaluating ranking functions has proved to be more difficult. Suppose there are n objects to rank. A ranking evaluation measure must induce a total order on the $n!$ possible ranking results. There seem to be many ways to define ranking measures and several evaluation measures have been proposed (Chapelle et al., 2009; Turpin and Scholer, 2006; Baeza-Yates and Ribeiro-Neto, 1999; Agarwal et al., 2004; Rudin, 2009). In fact, as pointed out by some authors, there is no single optimal ranking measure that works for any application (Croft et al., 2010).

The focus of this work is Normalized Discounted Cumulative Gain (NDCG) which is a family of ranking measures widely used in applications (Järvelin and Kekäläinen, 2000, 2002). NDCG has two advantages compared to many other measures. First, NDCG allows each retrieved document has graded relevance while most traditional ranking measures only allow binary relevance. That is, each document is viewed as either relevant or not relevant by previous ranking measures, while there can be degrees of relevancy for documents in NDCG. Second, NDCG involves a discount function over the rank while many other measures uniformly weight all positions. This feature is particularly important for search engines as users care top ranked documents much more than others.

NDCG is a normalization of the Discounted Cumulative Gain (DCG) measure. (For formal definition of both DCG and NDCG, please see Section 2.) DCG is a weighted sum of the degree of relevancy of the ranked items. The weight is a decreasing function of the rank (position) of the object, and therefore called discount. The original reason for introducing the discount is that the probability that a user views a document decreases with respect to its rank. NDCG normalizes DCG by the Ideal DCG (IDCG), which is simply the DCG measure of the best ranking result. Thus NDCG measure is always a number in $[0, 1]$. Strictly speaking, NDCG is a family of ranking measures, since there is flexibility in choosing the discount function. The logarithmic discount $\frac{1}{\log(1+r)}$, where r is the rank, dominated the literature and applications. We will refer to NDCG with logarithmic discount as the *standard* NDCG. Another discount function appeared in literature is r^{-1}, which is called Zipfian in Information Retrieval (Kanoulas and Aslam, 2009). Search engine systems also use a cut-off top-k version of NDCG. That is, the discount is set to be zero for ranks larger than k. Such NDCG measure is usually referred to as NDCG@k.

Given the importance and popularity of NDCG, there have been extensive studies on this measure, mainly in the field of Information Retrieval (Al-Maskari et al., 2007; Kanoulas and Aslam, 2009; Aslam et al., 2005; Voorhees, 2001; Sakai, 2006). All these research are conducted from an empirical perspective by doing experiments on benchmark datasets. There are also works that considered learning discount functions for NDCG

(Zhou et al.). Although these works gained insights about NDCG, there are still important issues unaddressed. We list a few questions that naturally arise.

- As pointed out in (Croft et al., 2010), there has not been any theoretically sound justification for using a logarithmic ($\frac{1}{\log(1+r)}$) discount other than the fact that it is a smooth decay.
- Is it possible to characterize the class of discount functions that are feasible for NDCG?
- For the standard NDCG@k, the discount is a combination of a very slow logarithmic decay and a hard cut-off. Why don't simply use a smooth discount that decays fast?

In this paper, we study the NDCG type ranking measures and address the above questions from a theoretical perspective. The goal of our study is twofold. First, we aim to provide a better understanding and theoretical justification of NDCG as an evaluation measure. Second, we hope that our results would shed light and be useful for further research on learning to rank based on NDCG. Specifically we analyze the behavior of NDCG as the number of objects to rank getting large. Asymptotics, including convergence and asymptotic normality, of many traditional ranking measures have been studied in depth in statistics, especially for Linear Rank Statistics and measures that are U-statistics (Hájek et al., 1967; Kendall, 1938). Clémençon and Vayatis (2009) observed that ranking measures such as Area under the ROC Curve (AUC), P-Norm Push and DCG can be viewed as Conditional Linear Rank Statistics. That is, conditioned on the relevance degrees of the items, these measures are Linear Rank Statistics (Hájek et al., 1967). They show uniform convergence based on an orthogonal decomposition of the measure. The convergence relies on the fact that the measure can be represented as a (conditional) average of a fixed score-generating function. Part of our work consider the convergence of NDCG and are closely related to (Clémençon and Vayatis, 2009). However, their results do not apply to our problem, because the score-generating function for NDCG is not fixed, it changes with the number of objects.

1.1. Our Results

Our study starts from an analysis of the standard NDCG (i.e., the one using logarithmic discount). The first discovery is that for *every* ranking function, the NDCG measure converges to 1 as the number of items to rank goes to infinity. This result is surprising. On the first sight it seems to mean that the widely used standard NDCG cannot differentiate good and bad ranking systems when the data is of large size. This problem may be serious because huge dataset is common in applications such as Web search.

To have a deeper understanding of NDCG, we first study what are the desired properties a good ranking measure should have. In this paper we propose a notion referred to as *consistent distinguishability*, which we believe that every ranking measure needs to have. Before describing the definition of consistent distinguishability, let us see a motivating example. Suppose we want to select, from two ranking functions f_1, f_2, a better one on ranking "sea" images (that is, if an image contains sea, we hope it is ranked near the top). Since there are billions of sea images on the web, a commonly used method is to randomly draw, say, a million data and evaluate the two functions on them. A crucial assumption underlying this approach is that the evaluation result will be "stable" on large datasets. That is, if on this randomly drawn dataset f_1 is better than f_2 according to the ranking measure, then with high probability over the random draw of another large dataset, f_1 should still be better than f_2. In other words, f_1 is *consistently* better than f_2 according to the ranking measure.

Our definition of consistent distinguishability captures the above intuition. It requires that for two substantially different ranking functions, the ranking measure can decide which one is better consistently on almost all datasets. (See Definition 5 for formal description.) Consistent distinguishability is a desired property to all ranking measures. However, it is not a priori clear whether NDCG type ranking measures have consistent distinguishability.

Our next main result shows that although the standard NDCG always converges to 1, it can consistently distinguishes every pair of substantially different ranking functions. Therefore, if one ignores the numerical scaling problem, standard NDCG is a good ranking measure.

We then study NDCG with other possible discount. We characterize the class of discount functions that are feasible for NDCG. It turns out that the Zipfian r^{-1} is a critical point. If a discount function decays slower than r^{-1}, the resulting NDCG measure has strong power of consistent distinguishability. If a discount decays substantially faster than r^{-1}, then it does not have this desired property. Even more, such ranking measures do not converge as the number of objects to rank goes to infinity.

Interestingly, this characterization result also provides a better understanding of the cut-off version NDCG@k. In particular, it gives a theoretical explanation to the previous question that why popular NDCG@k uses a combination of slow logarithmic decay and a hard cut-off as its discount rather than a smooth discount which decays fast.

Finally we consider how to choose the cut-off threshold for NDCG@k from the distinguishability point of view. We analyze the behavior of the measure for various choices of k as well as the discount. We suggest that choosing k as certain function of the size of the dataset may be appropriate.

1.2. Other Related Works

The importance of NDCG as well as other ranking measures in modern search engines is not limited as evaluation metrics. Currently ranking measures are also used as guidance for design of ranking functions due to works from the learning to rank area. Although early results of learning to rank often reduce ranking problem to classification or regression (Crammer and Singer, 2002; Freund et al., 2003; Joachims, 2002; Nallapati, 2004; Balcan et al., 2008), recently there is evidence that learning a ranking function by optimizing a ranking measure such as NDCG is a promising approach (Valizadegan et al., 2009; Yue et al., 2007). However, using the ranking measure as objective function to optimize is computationally intractable. Inspired by approaches in classification, some state of the art algorithms optimize a surrogate loss instead (Burges et al., 2007; Xia et al., 2008).

In the past a few years, there is rapidly growing interest in studying consistency of learning to rank algorithms that optimize surrogate losses. Such studies are motivated by the research of consistency of surrogate losses for classification (Zhang, 2004a; Bartlett et al., 2006; Zhang, 2004b; Tewari and Bartlett, 2007), which is a well-established theory in machine learning. Consistency of ranking is more complicated than classification as there are more than one possible ranking measures. One needs to study consistency with respect to a specific ranking measure. That is, whether the minimization of the surrogate leads to optimal predictions according to the risk defined by the given evaluation measure.

The research of consistency for ranking was initiated in (Cossock and Zhang, 2008; Duchi et al., 2010). In fact, Duchi et al. (2010) showed that no convex surrogate loss can be consistent with the Pairwise Disagreement (PD) measure. This result was further generalized in (Buffoni et al., 2011; Calauzènes et al., 2012), where non-existence of convex surrogate loss with Average Precision and Expected Reciprocal Rank were proved.

In contrast to the above negative results, Ravikumar et al. (2011) showed that there do exist NDCG consistent surrogates. Furthermore, by using a slightly stronger notion of NDCG consistency they showed that any NDCG consistent surrogate must be a Bregman distance. In a sense, these results mean that NDCG is a good ranking measure from a learning-to-rank point of view.

1.3. Organization

The rest of this paper is organized as follows. Section 2 provides basic notions and definitions. Section 3 contains the main theorems and key lemmas for the distinguishability theorem. Proofs and experimental results are all given in Appendix A-F because of the space.

2. Preliminaries

Let $\mathcal{X}$ be the instance space, and let $x_1, \ldots, x_n$ ($x_i \in \mathcal{X}$) be n objects to rank. Let $\mathcal{Y}$ be a finite set of degrees of relevancy. The simplest case is $\mathcal{Y} = \{0, 1\}$, where 0 corresponds to "irrelevant" and 1 corresponds to "relevant". Generally $\mathcal{Y}$ may contain more numbers; and for $y \in \mathcal{Y}$, the larger y is, the more relevant it represents. Let f be a ranking function[1]. We assume that f is a mapping from $\mathcal{X}$ to $\mathbb{R}$. For each object $x \in \mathcal{X}$, f gives it a score $f(x)$. For n objects $x_1, \ldots, x_n$, f ranks them according to their scores $f(x_1), \ldots, f(x_n)$. The resulting ranking list, denoted by $x^f_{(1)}, \ldots, x^f_{(n)}$, satisfies $f\left(x^f_{(1)}\right) \geq \ldots \geq f\left(x^f_{(n)}\right)$.

Let $y_1, \ldots, y_n$ ($y_i \in \mathcal{Y}$) be the degree of relevancy associated with $x_1, \ldots, x_n$. We will denote by $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ the set of data to rank. As in existing literature (Freund et al., 2003; Clemençon et al., 2008), we assume that $(x_1, y_1), \ldots, (x_n, y_n)$ are i.i.d. sample drawn from an underlying distribution P_{XY} over $\mathcal{X} \times \mathcal{Y}$. Also let $y^f_{(1)}, \ldots, y^f_{(n)}$ be the corresponding relevancy of $x^f_{(1)}, \ldots, x^f_{(n)}$.

The following is the formal definition of NDCG. Here we give a slightly simplified version tailored to our problem.

Definition 1 *Let $D(r)$ ($r \geq 1$) be a discount function. Let f be a ranking function, and S_n be a dataset. The Discounted Cumulative Gain (DCG) of f on S_n with discount D is*

1. In this paper we assume that the ranking function is a scoring function, i.e., the function first outputs a score for each object and then ranks them according to their scores. Indeed, a ranking function can be defined in a more general manner: For a fixed n, it is any permutation on $[n]$. Scoring functions are used in most applications. We also point out that Theorem 4 and Theorem 10 hold for any general ranking functions.

defined as[2]

$$\mathrm{DCG}_D(f, S_n) = \sum_{r=1}^{n} y_{(r)}^{f} D(r). \tag{1}$$

Let the Ideal DCG defined as $\mathrm{IDCG}_D(S_n) = \max_{f'} \sum_{r=1}^{n} y_{(r)}^{f'} D(r)$ *be the DCG value of the best ranking function on* S_n*.*

The NDCG of f *on* S_n *with discount* D *is defined as*

$$\mathrm{NDCG}_D(f, S_n) = \frac{\mathrm{DCG}_D(f, S_n)}{\mathrm{IDCG}_D(S_n)}. \tag{2}$$

We call NDCG *standard*, if its associated discount function is the inverse logarithm decay $D(r) = \frac{1}{\log(1+r)}$. Note that the base of the logarithm does not matter for NDCG, since constant scaling will cancel out due to normalization. We will assume it is the natural logarithm throughout this paper.

An important property of eq.(2) is that if a ranking function f' preserves the order of the ranking function f, then $\mathrm{NDCG}_D(f', S_n) = \mathrm{NDCG}_D(f, S_n)$ for all S_n. Here by preserving order we mean that for $\forall x, x' \in \mathcal{X}$, $f(x) > f(x')$ implies $f'(x) > f'(x')$, and vice versa. Thus the ranking measure NDCG is not just defined on a single function f, but indeed defined on an equivalent class of ranking functions which preserve order of each other.

Below we will frequently use a special ranking function $\tilde{f}$ that preserves the order of f.

Definition 2 *Let* f *be a ranking function. We call* $\tilde{f}$ *the canonical version of* f*, which is defined as*

$$\tilde{f}(x) = \Pr_{X \sim P_X}[f(X) \leq f(x)].$$

The canonical $\tilde{f}$ has the following properties, which can be easily proved by the definition.

Lemma 3 *For every ranking function* f*, its canonical version* $\tilde{f}$ *preserves the order of* f*. In addition,* $\tilde{f}(X)$ *has uniform distribution on* $[0, 1]$*.*

Finally, we point out that although originally the discount $D(r)$ is defined on positive integers r, below we will often treat $D(r)$ as a function of a real variable. That is, we view r take nonnegative real values. We will also consider derivative and integral of $D(r)$, denoted by $D'(r)$ and $\int D(r)\mathrm{d}r$ respectively.

3. Main Results

In this section, we give the main results of the paper. In Section 3.1 we study the standard NDCG, i.e., NDCG with logarithmic discount. In Section 3.2 we consider feasible discount other than the standard logarithmic one. We analyze the top-k cut-off version NDCG@k in Section 3.3. For clarity reasons, some of the results in Section 3.1, 3.2, and 3.3 are given for the simplest case that the relevance score is binary. Section 3.4 provides complete results for the general case.

2. Usually DCG is defined as $\mathrm{DCG}_D(f, S_n) = \sum_{r=1}^{n} G(y_{(r)}^{f}) D(r)$, where G is a monotone increasing function (e.g., $G(y) = 2^y - 1$). Here we omit G for notational simplicity. This does not lose any generality as we can assume that $\mathcal{Y}$ changes to $G(\mathcal{Y})$.

3.1. Standard NDCG

To study the behavior of the standard NDCG, we first consider the limit of this measure when the number of objects to rank goes to infinity. As stated in Section 2, we assume the data are i.i.d. drawn from some fixed underlying distribution. Surprisingly, it is easy to show that for every ranking function, standard NDCG converges to 1 almost surely.

Theorem 4 *Let $D(r) = \frac{1}{\log(1+r)}$. Then for every ranking function f,*

$$\mathrm{NDCG}_D(f, S_n) \to 1, \qquad a.s.$$

The proof is given in Appendix E.

At the first glance, the above result is quite negative for standard NDCG. It seems to say that in the limiting case, standard NDCG cannot differentiate ranking functions. However, Theorem 4 only considers the limits. To have a better understanding of NDCG, we need to make a deeper analysis of its power of distinguishability. In particular, Theorem 4 does not rule out the possibility that the standard NDCG can consistently distinguish substantially different ranking functions. Below we give the formal definition that two ranking functions are consistently distinguishable by a ranking measure $\mathcal{M}$.

Definition 5 *Let $(x_1, y_1), (x_2, y_2), \ldots$ be i.i.d. instance-label pairs drawn from the underlying distribution P_{XY} over $\mathcal{X} \times \mathcal{Y}$. Let $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. A pair of ranking functions f_0, f_1 is said to be* consistently distinguishable *by a ranking measure $\mathcal{M}$, if there exists a negligible function*[3] $\mathrm{neg}(N)$ *and $b \in \{0, 1\}$ such that for every sufficiently large N, with probability $1 - \mathrm{neg}(N)$,*

$$\mathcal{M}(f_b, S_n) > \mathcal{M}(f_{1-b}, S_n),$$

holds for all $n \geq N$ simultaneously.

Consistent distinguishability is appealing. One would like a ranking measure $\mathcal{M}$ to have the property that every two substantially different ranking functions are consistently distinguishable by $\mathcal{M}$. The next theorem shows that standard NDCG does have such a desired property. For clarity, here we state the theorem for the simple binary relevance case, i.e., $\mathcal{Y} = \{0, 1\}$. It is easy to extend the result to the general case that $\mathcal{Y}$ is any finite set.

Theorem 6 *For every pair of ranking functions f_0, f_1, let $\overline{y}^{f_i}(s) = \Pr[Y = 1 | \tilde{f}_i(X) = s]$, $i = 0, 1$. Assume $\overline{y}^{f_0}(s)$ and $\overline{y}^{f_1}(s)$ are Hölder continuous in s. Then, unless $\overline{y}^{f_0}(s) = \overline{y}^{f_1}(s)$ almost everywhere on $[0, 1]$, f_0 and f_1 are consistently distinguishable by standard NDCG.*

The proof is given in Appendix B.

Theorem 6 provides theoretical justification for using standard NDCG as a ranking measure, and answers the first question raised in Introduction. Although standard NDCG converges to the same limit for all ranking functions, it is still a good ranking measure with strong consistent distinguishability (if we ignore the numerical scaling issue).

3. A negligible function $\mathrm{neg}(N)$ means that for $\forall c, \mathrm{neg}(N) < N^{-c}$ for sufficiently large N.

3.2. Characterization of Feasible Discount Functions

In the previous section we demonstrate that standard NDCG can consistently distinguish good and bad ranking functions. In both literatures and real applications, standard NDCG is dominant. However, there is no known theoretical evidence that the logarithmic function is the only feasible discount, or it is the optimal one. In this subsection, we will investigate other discount functions. We study the asymptotic behavior and distinguishability of the induced NDCG measures and compare to the standard NDCG. Finally, we will characterize the class of discount functions which we think are feasible for NDCG. For the sake of clarity, the results in this subsection are given for the simplest case that $\mathcal{Y} = \{0, 1\}$. Complete results will be given in Section 3.4.

Standard NDCG utilizes the logarithmic discount which decays slowly. In the following we first consider a discount that decays a little faster. Specifically we consider $D(r) = r^{-\beta}$ $(0 < \beta < 1)$. Let us first investigate the limit of the ranking measure as the number of objects goes to infinity.

Theorem 7 *Assume $D(r) = r^{-\beta}$ where $\beta \in (0, 1)$. Assume also $p = \Pr[Y = 1] > 0$ and $\overline{y}^f(s) = \Pr[Y = 1 | \tilde{f}(X) = s]$ is a continuous function. Then*

$$\mathrm{NDCG}_D(f, S_n) \xrightarrow{p} \frac{(1-\beta)\int_0^1 \overline{y}^f(s) \cdot (1-s)^{-\beta} \mathrm{d}s}{p^{1-\beta}}. \tag{3}$$

The proof will be given in Appendix E.

For $D(r) = r^{-\beta}$ $(\beta \in (0, 1))$, NDCG no longer converges to the same limit for all ranking functions. The limit is actually a correlation between $\overline{y}^f(s)$ and $(1 - s)^{-\beta}$. For a good ranking function f, $\overline{y}^f(s) = \Pr[Y = 1|\tilde{f}(X) = s]$ is likely to be an increasing function of s, and thus has positive correlation with $(1 - s)^{-\beta}$. Therefore, the limit of the ranking measure already differentiate good and bad ranking functions to some extent.

We next study whether NDCG with polynomial discount has power of distinguishability as strong as the standard NDCG. That is, we will see if Theorem 6 holds for NDCG with $r^{-\beta}$ $(\beta \in (0, 1))$.

Theorem 8 *Let $D(r) = r^{-\beta}$, $\beta \in (0, 1)$. Assume $p = \Pr[Y = 1] > 0$. For every pair of ranking functions f_0, f_1, denote $\overline{y}^{f_i}(s) = \Pr[Y = 1|\tilde{f}_i(X) = s]$, $i = 0, 1$, and $\Delta y(s) = \overline{y}^{f_0}(s) - \overline{y}^{f_1}(s)$. Suppose at least one of the following two conditions hold: 1) $\int_0^1 \Delta y(s)(1 - s)^{-\beta}\mathrm{d}s \neq 0$; 2) $\overline{y}^{f_0}(s)$, $\overline{y}^{f_1}(s)$ are Hölder continuous with Hölder continuity constant α satisfying $\alpha > 3(1 - \beta)$, and $\Delta y(1) \neq 0$. Then f_0 and f_1 are strictly distinguishable with high probability by NDCG with discount $D(r)$.*

The proof will be given in Appendix F.

Theorem 8 involves two conditions. Satisfying either of them leads to strictly distinguishable with high probability. The first condition simply means that $\mathrm{NDCG}_D(f_0, S_n)$ and $\mathrm{NDCG}_D(f_1, S_n)$ converge to different limits and therefore the two functions are consistently distinguishable in the strongest sense. The second condition deals with the case that

$\mathrm{NDCG}_D(f_0, S_n)$ and $\mathrm{NDCG}_D(f_1, S_n)$ converge to the same limit. Comparing the distinguishability of NDCG with $r^{(-\beta)}$ discount with the standard NDCG, in most cases $r^{(-\beta)}$ discount has stronger distinguishability than standard NDCG (i.e., when the measures of two ranking functions converge to different limits). On the other hand, if we consider the worst case, standard NDCG is better, because it requires less conditions for consistent distinguishability.

We next study the Zipfian discount $D(r) = r^{-1}$. The following theorem describes the limit of the ranking measure.

Theorem 9 *Assume $D(r) = r^{-1}$. Assume also $p = \Pr[Y = 1] > 0$ and $\overline{y}^f(s) = \Pr[Y = 1|\tilde{f}(X) = s]$ is a continuous function. Then*

$$\mathrm{NDCG}_D(f, S_n) \xrightarrow{p} \Pr[Y = 1|\tilde{f}(X) = 1]. \tag{4}$$

The proof of Theorem 9 will be given in Appendix E.

The limit of NDCG with Zipfian discount depends only on the performance of the ranking function for the top ranks. The relevancy of lower ranked items does not affect the limit.

The next logical step would be analyzing the power of distinguishability of NDCG with Zipfian discount. However we are not able to prove that consistent distinguishability holds for this ranking measure. The techniques developed for distinguishability theorems given above does not apply to the Zipfian discount. Although we cannot disprove it distinguishability, we suspect that Zipfian does not have strong consistent distinguishability power.

Finally, we consider discount functions that decay substantially faster than r^{-1}. We will show that with these discount, NDCG does not converge as the number of objects tends to infinity. More importantly, such NDCG does not have the desired consistent distinguishability property.

Theorem 10 *Let $\mathcal{X}$ be instance space. For any $x \in \mathcal{X}$, let $y_x^* = \mathrm{argmax}_{y \in \mathcal{Y}} \Pr(Y = y|X = x)$. Assume that there is an absolute constant $\delta > 0$ such that for every $x \in \mathcal{X}$, $\Pr(Y = y|X = x) \geq \delta \cdot \Pr(Y = y_x^*|X = x)$ for all $y \in \mathcal{Y}$. If $\sum_{r=1}^{\infty} D(r) \leq B$ for some constant $B > 0$, then $\mathrm{NDCG}_D(f, S_n)$ does not converge in probability for any ranking function f. In particular, if $D(r) \leq r^{-(1+\epsilon)}$ for some $\epsilon > 0$, $\mathrm{NDCG}_D(f, S_n)$ does not converge. Moreover, every pair of ranking functions are not consistently distinguishable by NDCG with such discount.*

The proof is given in Appendix E.

Now we are able to *characterize* the feasible discounts for NDCG according to the results given so far. The logarithmic $\frac{1}{\log(1+r)}$ and polynomial $r^{-\beta}$ ($\beta \in (0, 1)$) are feasible discount functions for NDCG. For different ranking functions, standard NDCG converges to the same limit while the $r^{-\beta}$ ($\beta \in (0, 1)$) one converges to different limits in most cases. However, if we ignore the numerical scaling issue, both logarithmic and $r^{-\beta}$ ($\beta \in (0, 1)$) discount have consistent distinguishability. The Zipfian r^{-1} discount is on the borderline. It is not clear whether it has strong power of distinguishability. Discount that decays faster than $r^{-(1+\epsilon)}$ for some $\epsilon > 0$ is not appropriate for NDCG when the data size is large.

3.3. Cut-off Versions of NDCG

In this section we study the top-k version of NDCG, i.e., NDCG@k. For NDCG@k, the discount function is set as $D(r) = 0$ for all $r > k$. The motivation of using NDCG@k is to pay more attention to the top-ranked results. Logarithmic discount is also dominant for NDCG@k. We will call this measure standard NDCG@k. As already stated in Introduction, a natural question of standard NDCG@k is why use a combination of a very low logarithmic decay and a hard cut-off as the discount function. Why not simply use a smooth discount with fast decay, which seems more natural. In fact, this question has already been answered by Theorem 10. NDCG with such discount does not have strong power of distinguishability.

We next address the issue that how to choose the cut-off threshold k. It is obvious that setting k as a constant independent of n is not appropriate, because the partial sum of the discount is bounded and according to Theorem 10 the ranking measure does not converge. So k must grow unboundedly as n goes to infinity. Below we investigate the convergence and distinguishability of NDCG@k for various choices of k and the discount function. For clarity reason we assume here $\mathcal{Y} = \{0, 1\}$, and general results will be given in Section 3.4. The proofs of all theorems in this section will be given in Appendix E. We fist consider the case $k = o(n)$.

Theorem 11 *Let $\mathcal{Y} = \{0, 1\}$. Assume $D(r)$ is a discount function and $\sum_{r=1}^{\infty} D(r)$ is unbounded. Suppose $k = o(n)$ and $k \to \infty$ as $n \to \infty$. Let $\tilde{D}(r) = D(r)$ for all $r \leq k$ and $\tilde{D}(r) = 0$ for all $r > k$. Assume also that $p = \Pr[Y = 1] > 0$ and $\overline{y}^f(s) = \Pr[Y = 1|\tilde{f}(X) = s]$ is a continuous function. Then*

$$\mathrm{NDCG}_{\tilde{D}}(f, S_n) \xrightarrow{p} \Pr[Y = 1|\tilde{f}(X) = 1]. \tag{5}$$

The limit of NDCG@k where $k = o(n)$ is exactly the same as NDCG with Zipfian discount. Also like the Zipfian, the distinguishability power of this NDCG@k measure is not clear.

We next consider the case $k = cn$ for some constant $c \in (0, 1)$. We study the standard logarithmic and the polynomial discount respectively in the following two theorems.

Theorem 12 *Assume $D(r) = \frac{1}{\log(1+r)}$ and $\mathcal{Y} = \{0, 1\}$. Let $k = cn$ for some constant $c \in (0, 1)$. Define the cut-off discount function $\tilde{D}$ as $\tilde{D}(r) = D(r)$ if $r \leq k$ and $\tilde{D}(r) = 0$ otherwise. Assume also $p = \Pr[Y = 1] > 0$ and $\overline{y}^f(s) = \Pr[Y = 1|\tilde{f}(X) = s]$ is a continuous function. Then*

$$\mathrm{NDCG}_{\tilde{D}}(f, S_n) \xrightarrow{p} \frac{c}{\min\{c, p\}} \cdot \Pr[Y = 1|\tilde{f}(X) \geq 1 - c]. \tag{6}$$

Theorem 13 *Assume $D(r) = r^{-\beta}$ and $\mathcal{Y} = \{0, 1\}$, where $\beta \in (0, 1)$. Let $k = cn$ for some constant $c \in (0, 1)$. Define the cut-off discount function $\tilde{D}(r) = D(r)$ if $r \leq k$ and $\tilde{D}(r) = 0$*

otherwise. Assume also $p = \Pr[Y = 1] > 0$ *and* $\overline{y}^f(s) = \Pr[Y = 1|r(X) = s]$ *is a continuous function. Then*

$$\mathrm{NDCG}_{\tilde{D}}(f, S_n) \xrightarrow{p} \frac{1-\beta}{(\min\{c,p\})^{1-\beta}} \cdot \int_{1-c}^{1} \overline{y}^f(s) \cdot (1-s)^{-\beta} \mathrm{d}s. \tag{7}$$

The consistent distinguishability of the two measures considered in Theorem 12 and Theorem 13 are similar to their corresponding full NDCG respectively. To be precise, for NDCG@k ($k = cn$) with logarithmic discount and NDCG@k with $r^{-\beta}$ ($\beta \in (0,1)$) discount, consistent distinguishability holds under the condition given in Theorem 6 and Theorem 8 respectively. Hence these two cut-off versions NDCG are feasible ranking measures.

3.4. Results for General $\mathcal{Y}$

Some theorems given so far assume $\mathcal{Y} = \{0, 1\}$. Here we give complete results for the general case that $|\mathcal{Y}| \geq 2$, and $\mathcal{Y} = \{\mathfrak{y}_1, \ldots, \mathfrak{y}_{|\mathcal{Y}|}\}$. We only state the theorems and omit the proofs, which are straightforward modifications of the special case $\mathcal{Y} = \{0, 1\}$. The case $D(r) = \frac{1}{\log(1+r)}$ has already been included in Theorem 4. It always converges to 1 whatever the ranking function is. We next consider $r^{-\beta}$ decay.

Theorem 14 *Assume* $D(r) = r^{-\beta}$ *with* $\beta \in (0, 1)$*. Suppose that* $\mathcal{Y} = \{\mathfrak{y}_1, \ldots, \mathfrak{y}_{|\mathcal{Y}|}\}$*, where* $\mathfrak{y}_1 > \ldots > \mathfrak{y}_{|\mathcal{Y}|}$*. Assume* $f(X) \in [a, b]$*;* $f(X)$ *has a probability density function such that* $\mathbb{P}(f(X) = s) > 0$ *for all* $s \in [a, b]$*;* $\Pr(Y = \mathfrak{y}_j) > 0$ *and* $\Pr(Y = \mathfrak{y}_j|\tilde{f}(X) = s)$ *is a continuous function of* s *for all* j*. Then*

$$\mathrm{NDCG}_D(f, S_n) \xrightarrow{p} \frac{(1-\beta)\int_0^1 \mathbb{E}[Y|\tilde{f}(X) = s](1-s)^{-\beta} ds}{\sum_{j=1}^{|\mathcal{Y}|} \mathfrak{y}_j (R_j^{1-\beta} - R_{j-1}^{1-\beta})}$$

where $R_0 = 0$*;* $R_j = \Pr(Y \geq \mathfrak{y}_j)$*.*

The next theorem is for top-k type NDCG measures, where $k = o(n)$.

Theorem 15 *Suppose that* $\mathcal{Y} = \{\mathfrak{y}_1, \ldots, \mathfrak{y}_{|\mathcal{Y}|}\}$*, where* $\mathfrak{y}_1 > \ldots > \mathfrak{y}_{|\mathcal{Y}|}$*. Assume* $D(r)$ *and* k *grow unboundedly and* $k/n = o(1)$*. For any* n*, let* $\tilde{D}(r) = D(r)$ *if* $r \leq k$ *and* $\tilde{D}(r) = 0$ *otherwise. Assume* $f(X) \in [a, b]$*;* $f(X)$ *has a probability density function such that* $\mathbb{P}(f(X) = s) > 0$ *for all* $s \in [a, b]$*;* $\Pr(Y = \mathfrak{y}_j) > 0$ *and* $\Pr(Y = \mathfrak{y}_j|f(X) = s)$ *is a continuous function of* s *for all* j*. Then*

$$\mathrm{NDCG}_{\tilde{D}}(f, S_n) \xrightarrow{p} \frac{1}{\mathfrak{y}_1} \cdot \mathbb{E}[Y|\tilde{f}(X) = 1].$$

The last two theorems are for top-k, where $k/n = c$. We consider both logarithm discount and polynomial discount separately.

Theorem 16 *Suppose that* $\mathcal{Y} = \{\mathfrak{y}_1, \ldots, \mathfrak{y}_{|\mathcal{Y}|}\}$*, where* $\mathfrak{y}_1 > \ldots > \mathfrak{y}_{|\mathcal{Y}|}$*. Let* $k/n = c$ *for some constant* $c > 0$*. Let* $D(r) = \frac{1}{\log(1+r)}$*. For any* n*, let* $\tilde{D}(r) = D(r)$ *if* $r \le k$ *and* $\tilde{D}(r) = 0$ *otherwise. Assume* $f(X) \in [a, b]$*;* $f(X)$ *has a probability density function such that* $\mathbb{P}(f(X) = s) > 0$ *for all* $s \in [a, b]$*;* $\Pr(Y = \mathfrak{y}_j) > 0$ *and* $\Pr(Y = \mathfrak{y}_j | f(X) = s)$ *is a continuous function of* s *for all* j*. Then*

$$\mathrm{NDCG}_{\tilde{D}}(f, S_n) \xrightarrow{p} \frac{c \cdot \mathbb{E}[Y | \tilde{f}(X) \ge 1 - c]}{\sum_{j=1}^{t} \mathfrak{y}_j (R_j - R_{j-1}) + \mathfrak{y}_{t+1}(c - R_t)}.$$

where $R_0 = 0$*;* $R_j = \mathbb{P}(Y \ge \mathfrak{y}_j)$*;* t *is defined by* $R_t < c \le R_{t+1}$*.*

Theorem 17 *Let* $D(r) = r^{-\beta}$ *with* $\beta \in (0, 1)$*, and* $\tilde{D}(r) = D(r)$ *if* $r \le k$ *and* $\tilde{D}(r) = 0$ *otherwise. Using the same notions and under the same conditions as in Theorem 16*

$$\mathrm{NDCG}_{\tilde{D}}(f, S_n) \xrightarrow{p} \frac{(1 - \beta) \int_{1-c}^{1} \mathbb{E}[Y | \tilde{f}(X) = s](1 - s)^{-\beta} ds}{\sum_{j=1}^{t} \mathfrak{y}_j (R_j^{1-\beta} - R_{j-1}^{1-\beta}) + \mathfrak{y}_{t+1}(c^{1-\beta} - R_t^{1-\beta})}.$$

Acknowledgement

We thank Kai Fan and Ziteng Wang for helpful discussions. This work was supported by National Science Foundation of China (61222307, 61075003, 61033001, 61061130540); National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301; NCET-12-0015; and a grant from MOE-Microsoft Laboratory of Statistics and Information Technology of Peking University.

Appendix A. Experimental Results

All theoretical results in this paper are proved under the assumption that the objects to rank are i.i.d. data. Often in real applications the data are not strictly i.i.d or even not random. Here we conduct experiments on a real dataset — Web search data. The aim is to see to what extent the behavior of the ranking measures on real datasets agree with our theory obtained under the i.i.d. assumption.

The dataset we use contains click-through log data of a search engine. We collected the clicked documents for 40 popular queries as test set, which are regarded as 40 independent ranking tasks. In each task, there are 5000 Web documents with clicks. To avoid heavy work of human labeling, we simply label each document by its click number according to the following rule. We assign relevancy $y = 2$ to documents with more than 1000 clicks, 1 to those with 100 to 1000 clicks, and 0 to the rest. In each task, we extracted 40 features for each item representing its relevance to the given query. A detail is how to construct S_n. In our theoretical analysis we assume S_n contains i.i.d. data. Since the goal of the experiments is to see how our theory works for real applications, we construct S_n as follows. For each query, there are totally 5000 documents which we denote by $x_1, \ldots, x_{5000}$. Assume each document has a generating time. Without loss of generality we assume x_1 was generated earliest and x_{5000} latest. We set $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ for each $1 \le n \le 5000$. Such a construction simulates that in reality there may be increasing number of documents needed

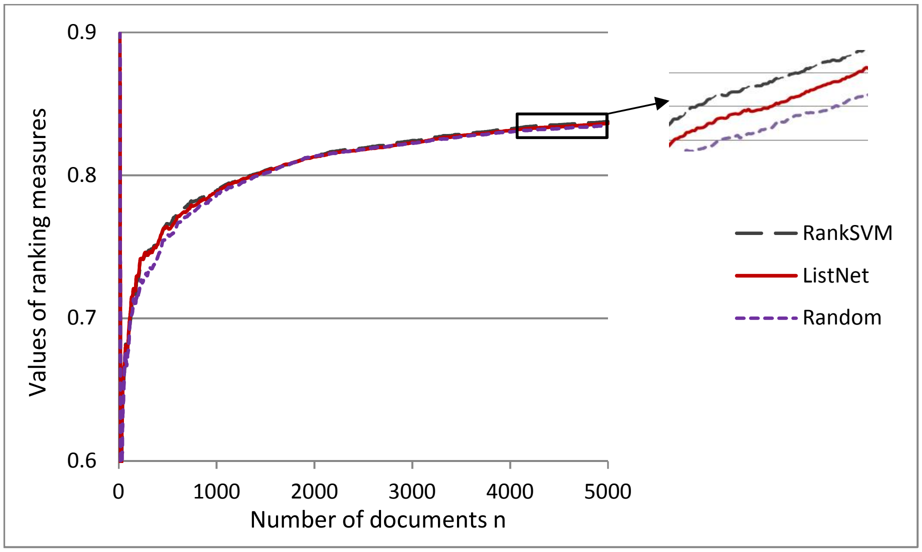

Figure 1: Standard NDCG: Converges to the same limit but distinguishes well the ranking functions.

to rank by a search engine over time. We use three ranking functions in the experiments: a trained RankSVM model (Herbrich et al., 1999), a trained ListNet model (Cao et al., 2007), and a function chosen randomly. To be concrete, the random function is constructed as follows. For each $x \in \mathcal{X}$, we set $f(x)$ by choosing a number uniformly random from $[-1, 1]$. For the trained models (i.e., listNet and RankSVM), parameters are learned from a separate large training set construct in the same manner as the test set. Clearly, ListNet and RankSVM are relatively good ranking functions and the random function is bad.

We analyze the following typical NDCG type ranking measures by experiments:

- Standard NDCG: $D(r) = \frac{1}{\log(1+r)}$. See Figure 1, Theorem 4 and Theorem 6.
- NDCG with a feasible discount function: $D(r) = r^{-1/2}$. See Figure 2, Theorem 7 and Theorem 8.
- NDCG with too fast decay: $D(r) = 2^{-r}$. See Figure 3 and Theorem 10.
- NDCG@k: $k = n/5$; $D(r) = \frac{1}{\log(1+r)}$. See Figure 4 and Theorem 12.

Figure 1 agrees well with Theorem 4 and Theorem 6. On the one hand, the NDCG measures of the three ranking functions are very close and seem to converge to the same limit. On the other hand, one can see from the enlarged part (we enlarge and stretch the vertical axis) in the figure that in fact the measures distinguish well the ranking functions.

Figure 2 demonstrates the result of NDCG with the feasible discount $r^{-1/2}$. In this experiment, it seems that the ranking measures of the three ranking functions converge to different limits and therefore distinguish them very well. In our experimental setting, it is not easy to find two ranking functions whose NDCG measures converge to the same limit.

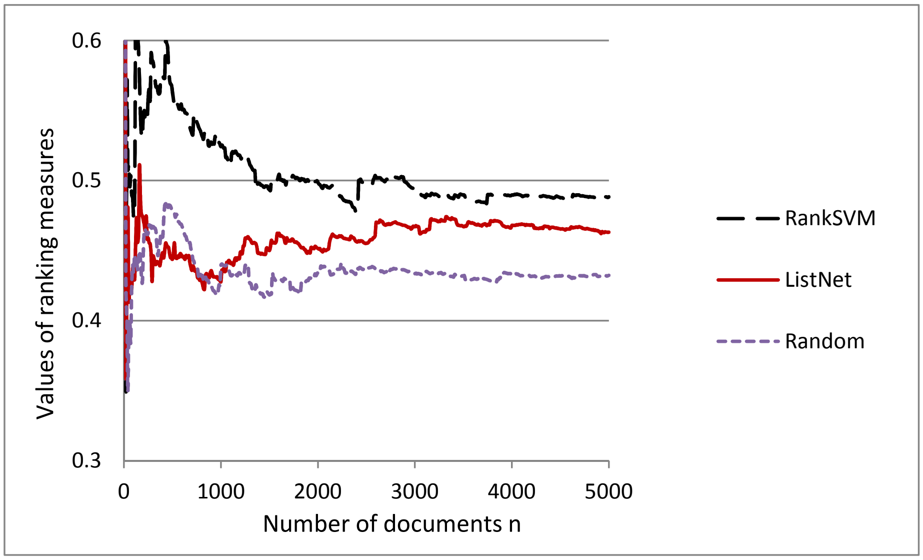

Figure 2: NDCG with feasible discount $D(r) = r^{-1/2}$: converges to different limits and distinguishes well the ranking functions.

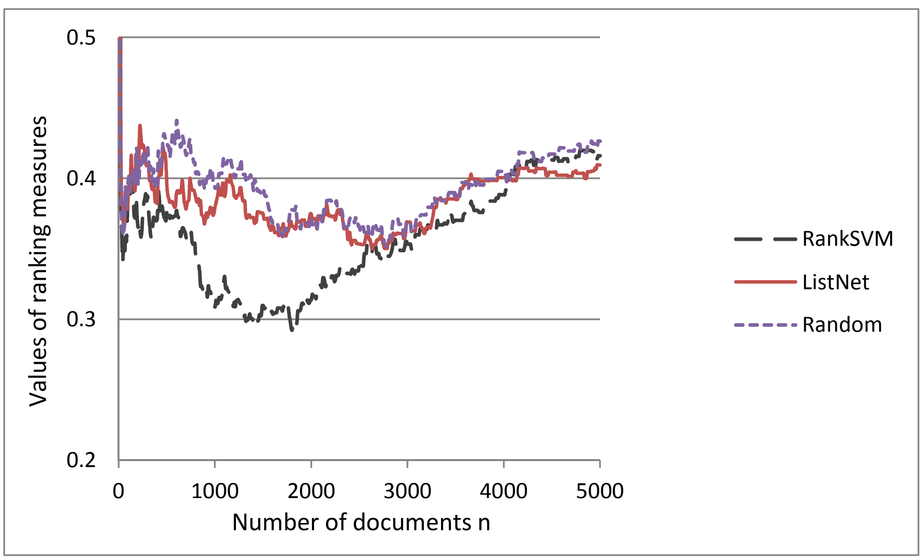

Figure 3: NDCG with too fast decay $D(r) = 2^{-r}$: does not converge; does not have good distinguishability power either.

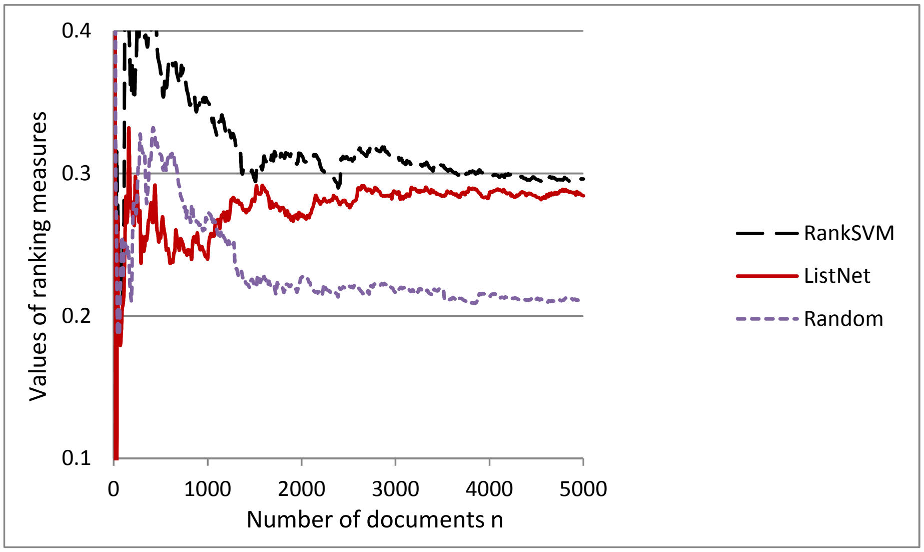

Figure 4: NDCG@k ($D(r) = \frac{1}{\log(1+r)}$, $k = n/5$): distinguishes well the ranking functions.

If one can find such a pair of ranking functions, it would be interesting to see how well the measure distinguish them.

Figure 3 shows the behavior of NDCG with a smooth discount which decays too fast. The measure cannot distinguish the three ranking functions very well. Even the randomly chosen function has an NDCG score similar to those of RankSVM and ListNet. From the figure, it is also likely that the measures do not converge.

Figure 4 depicts the result of NDCG@k, where k is a constant proportion of n. Before describing the result, let us first comparing Theorem 12 and Theorem 4. Note that although the discount are both the logarithmic one, NDCG@k for $k = cn$ can converge to different limits for different ranking functions, while standard NDCG always converges to 1. Figure 4 clearly demonstrate this result.

Appendix B. Proof of Theorem 6: the Key Lemmas

In this section we will prove Theorem 6. In fact we will prove a more complete result. The proof relies on a few key lemmas. In this section we only state these lemmas. Their proofs will be given in Appendix C. First we give a weaker definition of distinguishability, which guarantees that the ranking measure $\mathcal{M}$ gives consistent comparison results for two ranking functions only in expectation.

Definition 18 *Fix an underlying distribution P_{XY}. A pair of ranking functions f_0, f_1 is said to be* distinguishable in expectation *by a ranking measure $\mathcal{M}$, if there exist $b \in \{0,1\}$ and a positive integer N such that for all $n \geq N$,*

$$\mathbb{E}\big[\mathcal{M}(f_b, S_n)\big] > \mathbb{E}\big[\mathcal{M}(f_{1-b}, S_n)\big],$$

where the expectation is over the random draw of S_n.

Now we state a theorem which contains Theorem 6.

Theorem 19 *Assume that* $p = \Pr(Y = 1) > 0$. *For every pair of ranking functions* f_0, f_1, *Let* $\overline{y}^{f_i}(s) = \Pr[Y = 1|\tilde{f}_i(X) = s]$, $i = 0, 1$. *Unless* $\overline{y}^{f_0}(s) = \overline{y}^{f_1}(s)$ *almost surely on* $[0, 1]$, f_0, f_1 *are distinguishable in expectation by standard NDCG whose discount is* $D(r) = \frac{1}{\log(1+r)}$.

Moreover, if $\overline{y}^{f_0}(s)$ *and* $\overline{y}^{f_1}(s)$ *are Hölder continuous in* s, *then unless* $\overline{y}^{f_0}(s) = \overline{y}^{f_1}(s)$ *almost everywhere on* $[0, 1]$, f_0 *and* f_1 *are consistently distinguishable by standard NDCG.*

To prove Theorem 19, we need some notations.

Definition 20 *Suppose* $\mathcal{Y} = \{0, 1\}$. *Let* $\overline{y}^f(s) = \Pr[Y = 1|\tilde{f}(X) = s]$. *Also let* $F(t) = \int_1^t D(s)ds$. *We define the* unnormalized pseudo-expectation $\tilde{N}_D^f(n)$ *as*

$$\tilde{N}_D^f(n) = \int_1^n \overline{y}^f(1 - s/n)D(s)\mathrm{d}s = n\int_{\frac{1}{n}}^1 \overline{y}^f(1 - s)D(ns)\mathrm{d}s.$$

Assume that $p = \Pr(Y = 1) > 0$. *Define the* normalized pseudo-expectation $N_D^f(n)$ *as*

$$N_D^f(n) = \frac{\tilde{N}_D^f(n)}{F(np)}.$$

The proof of the first part of Theorem 19 (i.e., distinguishable in expectation) relies on the following two key lemmas, whose proofs will be given in Appendix C.

Lemma 21 *Let* $D(r) = \frac{1}{\log(1+r)}$. *Assume that* $p = \Pr(Y = 1) > 0$. *Then for every ranking function* f,

$$\left|\mathbb{E}[\mathrm{NDCG}_D(f, S_n)] - N_D^f(n)\right| \le \tilde{O}\left(n^{-1/3}\right).$$

Lemma 22 *Let* $D(r) = \frac{1}{\log(1+r)}$. *Assume that* $p = \Pr(Y = 1) > 0$. *let* $\overline{y}^{f_i}(s) = \Pr[Y = 1|\tilde{f}_i(X) = s]$, $i = 0, 1$. *Unless* $\overline{y}^{f_0}(\cdot) = \overline{y}^{f_1}(\cdot)$ *almost everywhere on* $[0, 1]$, *there must exist a nonnegative integer* K *and a constant* $a \neq 0$, *such that*

$$\left|N_D^{f_0}(n) - N_D^{f_1}(n) - \frac{a}{\log^K n}\right| \le O\left(\frac{1}{\log^{K+1} n}\right).$$

Lemma 21 says that the difference between the expectation of the NDCG measure of a ranking function and its pseudo-expectation is relatively small; while Lemma 22 says that the difference between the pseudo-expectations of two essentially different ranking functions are much larger.

To prove the "moreover" part of Theorem 19 (i.e., consistently distinguishable), we need the following key lemma, whose proof will be given in Section C. The lemma states that with high probability the NDCG measure of a ranking function is very close to its pseudo-expectation.

Lemma 23 *Let $D(r) = \frac{1}{\log(1+r)}$. Assume that $p = \Pr(Y = 1) > 0$. Suppose the ranking function f satisfies that $\overline{y}^f(s) = \Pr(Y = 1|\tilde{f}(X) = s)$ is Hölder continuous with constants $\alpha > 0$ and $C > 0$. That is, $|\overline{y}^f(s) - \overline{y}^f(s')| \le C|s - s'|^\alpha$ for all $s, s' \in [0, 1]$. Then*

$$\Pr\left[\left|\mathrm{NDCG}_D(f, S_n) - N_D^f(n)\right| \ge 5Cp^{-1}n^{-\min(\alpha/3,1)}\right] \le O\left(e^{-n^{1/4}}\right).$$

Proof of Theorem 19 That f_0 and f_1 are strictly distinguishable in expectation by standard NDCG is straightforward from Lemma 21 and Lemma 22. That f_0 and f_1 are strictly distinguishable with high probability follows immediately from Lemma 23, Lemma 22 and the observation that $\sum_{n \ge N} e^{-n^{1/4}} \le O\left(N^{3/4}e^{-N^{1/4}}\right) \le O\left(e^{-N^{1/5}}\right)$. ■

Appendix C. Proofs of the Key Lemmas in Appendix B

In this section, we give proofs of the three key lemmas in Appendix B (i.e., Lemma 21, Lemma 22 and Lemma 23) used to prove Theorem 6 and Theorem 19.

To prove the key lemmas, we need a few technical claims, whose proofs will be given in Appendix D. We first give four claims that will be used in the proof of Lemma 21.

Claim 24 *For any $s \in [0, 1]$,*

$$\sum_{r=1}^{n} \mathbb{P}\left[\tilde{f}(x_{(r)}^f) = s\right] = n. \tag{8}$$

Claim 25 *Recall that the DCG ranking measure with respect to discount $D(\cdot)$ was defined as*

$$\mathrm{DCG}_D(f, S_n) = \sum_{r=1}^{n} y_{(r)}^f D(r). \tag{9}$$

Let $D(r) = \frac{1}{\log(1+r)}$, and $\overline{y}^f(s) = \Pr[Y = 1|\tilde{f}(X) = s]$. Then

$$\mathbb{E}\left[\mathrm{DCG}_D(f, S_n)\right] = \sum_{r=1}^{n} \frac{1}{\log(1+r)} \int_0^1 \mathbb{P}\left[\tilde{f}(x_{(r)}^f) = 1 - s\right] \overline{y}^f(1 - s)\mathrm{d}s. \tag{10}$$

Claim 26 *For any positive integer n, define $E_{n,r} = [\frac{r}{n} - n^{-1/3}, \frac{r}{n} + n^{-1/3}]$ ($r \in [n]$). Then for any $r \in [n]$,*

$$\Pr\left[1 - \tilde{f}(x_{(r)}^f) \in E_{n,r}\right] \ge 1 - 2e^{-n^{1/3}}. \tag{11}$$

Claim 27 *Let $\mathcal{Y} = \{0, 1\}$. Assume $D(r) = \frac{1}{\log(1+r)}$. Let $F(t) = \int_1^t D(s)\mathrm{d}s$. Assume also $p = \Pr[Y = 1] > 0$. Then for every sufficiently large n, with probability $(1 - 2e^{-2n^{1/3}})$ the following inequality holds.*

$$\left|\mathrm{NDCG}_D(f, S_n) - \frac{\mathrm{DCG}_D(f, S_n)}{F(np)}\right| \le O\left(n^{-1/3}\right). \tag{12}$$

Now we are ready to prove Lemma 21.

Proof of Lemma 21. By the definition of $\tilde{N}_D^f(n)$ (see Definition 20) and eq.(8), we have

$$\tilde{N}_D^f(n) = n\int_{\frac{1}{n}}^{1} \frac{\overline{y}^f(1-s)\mathrm{d}s}{\log(1+ns)} = \sum_{r=1}^{n}\int_{\frac{1}{n}}^{1} \frac{\overline{y}^f(1-s)\mathbb{P}[\tilde{f}(x_{(r)}^f) = 1-s]}{\log(1+ns)}\mathrm{d}s. \tag{13}$$

By eq. (10) in Claim 25 and eq.(13), and note that $\overline{y}^f(s) \le 1$, we obtain

$$\begin{aligned}
&\left|\mathbb{E}[\mathrm{DCG}_D(f,S_n)] - \tilde{N}_D^f(n)\right| \\
\le\ & \sum_{r=1}^{n}\left|\int_{\frac{1}{n}}^{1} \overline{y}^f(1-s)\mathbb{P}\big[\tilde{f}(x_{(r)}^f) = 1-s\big]\left(\frac{1}{\log(1+r)} - \frac{1}{\log(1+ns)}\right)\mathrm{d}s\right| + \frac{1}{n}\sum_{r=1}^{n}\frac{1}{\log(1+r)} \\
\le\ & \sum_{r=1}^{n}\left|\int_{[\frac{1}{n},1]\setminus E_{n,r}} \mathbb{P}\big[\tilde{f}(x_{(r)}^f) = 1-s\big]\left|\frac{1}{\log(1+r)} - \frac{1}{\log(1+ns)}\right|\mathrm{d}s\right. \\
+\ & \sum_{r=1}^{n}\int_{E_{n,r}\cap[\frac{1}{n},1]}\left|\frac{1}{\log(1+r)} - \frac{1}{\log(1+ns)}\right|\mathrm{d}s + O\left(\frac{1}{\log n}\right).
\end{aligned} \tag{14}$$

We next bound the two terms in the RHS of the last inequality of (14) separately. By Claim 26, the first term can be upper bounded by

$$2e^{-n^{1/3}}\sum_{r=1}^{n}\sup_{s\in[\frac{1}{n},1]\setminus E_{n,r}}\left|\frac{1}{\log(1+r)} - \frac{1}{\log(1+ns)}\right| \le \frac{2}{\log 2}ne^{-2n^{1/3}}. \tag{15}$$

For the second term in the RHS of the last inequality of (14), it is easy to check that the following two inequalities hold:

$$\forall r > n^{2/3},\quad \sup_{s\in E_{n,r}\cap[\frac{1}{n},1]}\left|\frac{1}{\log(1+r)} - \frac{1}{\log(1+ns)}\right| \le \frac{n^{2/3}}{(1+r)\log^2(1+r)} + o\left(\frac{n^{2/3}}{(1+r)\log^2(1+r)}\right). \tag{16}$$

$$\forall r \le n^{2/3},\quad \sup_{s\in E_{n,r}\cap[\frac{1}{n},1]}\left|\frac{1}{\log(1+r)} - \frac{1}{\log(1+ns)}\right| \le \frac{1}{\log 2}. \tag{17}$$

Combining (14), (15), (16) and (17), we obtain

$$\left|\mathbb{E}[\mathrm{DCG}_D(f,S_n)] - \tilde{N}_D^f(n)\right| \le \frac{2ne^{-2ne^{1/3}}}{\log 2} + O\left(\frac{n^{2/3}}{\log 2} + \sum_{r=n^{2/3}}^{n}\frac{n^{2/3}}{(1+r)\log^2(1+r)}\right) \le \tilde{O}\left(n^{2/3}\right). \tag{18}$$

Finally, observe that $F(np) = \mathrm{Li}(1+np)$, where Li is the offset logarithmic integral function. By Claim 27 and the well-known fact $\mathrm{Li}(n) \sim \frac{n}{\log n}$, we have the following inequality and this completes the proof.

$$\left|\mathbb{E}[\mathrm{NDCG}_D(f,S_n)] - \frac{\mathbb{E}[\mathrm{DCG}_D(f,S_n)]}{\mathrm{Li}(1+np)}\right| \le \tilde{O}\left(n^{-1/3}\right) + O\left(e^{-2n^{1/3}}\right). \tag{19}$$

■

We next turn to prove Lemma 22. We need the following three claims.

Claim 28 *For sufficiently large n,*

$$\int_0^{\frac{2}{n}} \log^k x \mathrm{d}x = O\left(\frac{\log^k n}{n}\right). \tag{20}$$

Claim 29 *Fix an integer $k \in \mathbb{N}^* = \{0\} \cup \mathbb{N}$. For sufficiently large n,*

$$\int_{\frac{2}{n}}^1 \frac{\left|\log^k x\right| \mathrm{d}x}{(\log(nx))^{k+1}} \leq O\left(\frac{1}{\log^{k+1} n}\right). \tag{21}$$

Claim 30 $\mathrm{span}\left(\{\log^k x\}_{k\geq 0}\right)$*, is dense in $L^2[0,1]$.*

Now we are ready to prove Lemma 22.

Proof of Lemma 22. Let $\Delta y(s) = \overline{y}^{f_0}(s) - \overline{y}^{f_1}(s)$. By the definition of normalized pseudo expectation (see definition 20) and the fact that $|\Delta y(s)| \leq 1$, we have

$$\begin{aligned} N_D^{f_0}(n) - N_D^{f_1}(n) &= \frac{n}{\mathrm{Li}(1+np)} \int_{\frac{1}{n}}^1 \frac{\Delta y(1-s)\mathrm{d}s}{\log(1+ns)} \\ &= \frac{n}{\mathrm{Li}(1+np)} \int_{\frac{2}{n}}^1 \frac{\Delta y(1-s)\mathrm{d}s}{\log(1+ns)} + O\left(\frac{1}{\mathrm{Li}(n)}\right). \end{aligned} \tag{22}$$

Expanding $\frac{1}{\log(1+ns)}$ at the point ns, we obtain

$$\left| \int_{\frac{2}{n}}^1 \frac{\Delta y(1-s)\mathrm{d}s}{\log(1+ns)} - \int_{\frac{2}{n}}^1 \frac{\Delta y(1-s)\mathrm{d}s}{\log n + \log s} \right| \leq \int_{\frac{2}{n}}^1 \frac{\mathrm{d}s}{ns\log^2(ns)} \leq O\left(\frac{\log n}{n}\right). \tag{23}$$

Expanding $\frac{1}{\log n + \log s}$ at point $\log n$, we have that for all $m \in \mathbb{N}^*$, the following holds:

$$\begin{aligned} &\left| \int_{\frac{2}{n}}^1 \frac{\Delta y(1-s)\mathrm{d}s}{\log n + \log s} - \sum_{j=1}^m \frac{(-1)^{j-1}}{\log^j n} \int_{\frac{2}{n}}^1 \Delta y(1-s) \log^{j-1} s \ \mathrm{d}s \right| \\ = &\left| \int_{\frac{2}{n}}^1 \frac{\Delta y(1-s)\log^m s\mathrm{d}s}{(\log n + \xi_{n,s})^{m+1}} \right| \leq \int_{\frac{2}{n}}^1 \frac{\left|\Delta y(1-s)\log^m s\right|\mathrm{d}s}{(\log n + \log s)^{m+1}} \leq O\left(\frac{1}{\log^{m+1} n}\right). \end{aligned} \tag{24}$$

Note in above derivation that $\xi_{n,s} \in (\log s, 0))$, and the last inequality is due to Claim 29.

Furthermore, by Claim 30, unless $\Delta y(s) = 0$ a.e., there exist constants $k \in \mathbb{N}^*$ and $a \neq 0$ such that

$$(-1)^k \int_0^1 \Delta y(1-s) \log^k s \, \mathrm{d}s = a. \tag{25}$$

Let K be the smallest integer k that Eq. (25) holds. Combining (22), (23), (24), and (25) and noting Claim 28, we have the following and this completes the proof.

$$\left| N_D^{f_0}(n) - N_D^{f_1}(n) - \frac{a}{\log^K n} \right| \leq O\left(\frac{\log^K n}{n}\right) + O\left(\frac{1}{\log^{K+1} n}\right).$$

■

To prove the last key lemma, we need the following claim.

Claim 31 *Let $D(r) = \frac{1}{\log(1+r)}$. Let $F(t) = \int_1^t D(r)\mathrm{d}r$. Assume $\overline{y}^f(s)$ is Hölder continuous with constants α and C. Then*

$$\Big| \sum_{r=1}^n \overline{y}^f(1-r/n)D_r - \tilde{N}_D^f(n) \Big| \leq Cn^{-\alpha/3}F(n) + D(1) + |D'(1)|. \tag{26}$$

Now we prove the last key lemma.

Proof of Lemma 23. Let $x_1, \cdots, x_n$ be instances i.i.d. drawn according to P_X. Let $\tilde{x}_{(r)} = \tilde{f}(x^f_{(r)})$ and by definition $\tilde{x}_{(1)} \geq \tilde{x}_{(2)} \geq \cdots \geq \tilde{x}_{(n)}$. By Chernoff bound, for every r with probability $2e^{-2n^{1/3}}$ we have $|\tilde{x}_{(r)} - (1 - r/n)| > n^{-1/3}$. A union bound over r then yields

$$\Pr\left[\forall r \in [n], \left|\tilde{x}_{(r)} - \left(1 - \frac{r}{n}\right)\right| \leq n^{-1/3}\right] \geq 1 - 2ne^{-2n^{1/3}}. \tag{27}$$

Since y^f is Hölder continuous with constants α and C, eq. (27) implies

$$\Pr\left[\Big| \sum_{r=1}^n \overline{y}^f(\tilde{x}_{(r)})D(r) - \sum_{r=1}^n \overline{y}^f(1-r/n)D(r) \Big| \leq Cn^{-\alpha/3} \cdot \sum_{r=1}^n D(r)\right] \geq 1 - 2ne^{-2n^{1/3}}. \tag{28}$$

Combining Claim 31 and eq. (28), and note that $|D'(1)| + D(1) \leq 10$ we have

$$\Pr\left[\Big| \sum_{r=1}^n \overline{y}^f(\tilde{x}_{(r)})D(r) - \tilde{N}_D^f(n) \Big| \leq 2Cn^{-\alpha/3} \cdot F(n) + 10\right] \geq 1 - 2ne^{-2n^{1/3}}. \tag{29}$$

Fix $x_1, \ldots, x_n$. Let $x^f_{(1)}, \ldots, x^f_{(n)}$ be the induced ordered sequence. Also let $\tilde{x}_{(r)} = \tilde{f}(x^f_{(r)})$. Recall that $\overline{y}^f(s) = \mathbb{E}[Y|\tilde{f}(X) = s]$. Thus $\sum_{r=1}^n \overline{y}^f(\tilde{x}_{(r)})D(r)$ is the expectation of $\mathrm{DCG}_D(f, S_n) = \sum_{r=1}^n y^f_{(r)}D(r)$ conditioned on the fixed values $\tilde{x}_{(1)}, \ldots, \tilde{x}_{(n)}$. Also observe that conditioning on $\tilde{x}_{(1)}, \ldots, \tilde{x}_{(n)}$, $y^f_{(r)}$ $(r = 1, \ldots, n)$ are independent. By Hoeffding's inequality and taking into consideration that $x_1, \ldots, x_n$ are arbitrary and $(D(r))^2 \leq D(r)$ for all r, we have for every $\epsilon > 0$

$$\Pr\left[\left|\mathrm{DCG}_D(f, S_n) - \sum_{r=1}^{n} \overline{y}^f(\tilde{x}_{(r)})D(r)\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2}{F(n)}\right). \tag{30}$$

Set $\epsilon = F(n)^{2/3}$ in eq. (30) and combine eq. (29), we have

$$\Pr\left[\left|\mathrm{DCG}_D(f, S_n) - \tilde{N}_D^f(n)\right| > 2Cn^{-\alpha/3}F(n) + 2F(n)^{2/3}\right] \leq 2ne^{-2n^{1/3}} + 2e^{-2F(n)^{1/3}}. \tag{31}$$

Simple calculations yields

$$\Pr\left[\left|\frac{\mathrm{DCG}_D(f, S_n)}{F(np)} - N_D^f(n)\right| > 4Cp^{-1}n^{-\min(\alpha/3, 1)}\right] \leq 2ne^{-2n^{1/3}} + 2e^{-2F(n)^{1/3}}. \tag{32}$$

Combining eq. (12) and (32) The lemma follows. ■

Appendix D. Proof of the Technical Claims in Appendix C

Here we give proofs of the technical claims by which we prove the three key lemmas in Section C.

Proof of Claim 24.

Recall that for each $i \in [n]$, $\tilde{f}(x_i)$ is uniformly distributed on $[0, 1]$; and $x^f_{(1)}, \ldots, x^f_{(n)}$ are just reordering of $x_1, \ldots, x_n$. Thus

$$\sum_{r=1}^{n} \mathbb{P}\big[\tilde{f}(x^f_{(r)}) = s\big] = \sum_{i=1}^{n} \mathbb{P}\big[\tilde{f}(x_i) = s\big] = n.$$

■

Proof of Claim 25.

We have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{DCG}_D(f, S_n)\right] &= \sum_{r=1}^{n} D(r)\mathbb{E}\left[y^f_{(r)}\right] \\
&= \sum_{r=1}^{n} \frac{1}{\log(1+r)}\mathbb{E}\Big[\mathbb{E}\big[y^f_{(r)}|\tilde{f}(x^f_{(r)})\big]\Big] \\
&= \sum_{r=1}^{n} \frac{1}{\log(1+r)} \int_0^1 \mathbb{P}\left[\tilde{f}(x^f_{(r)}) = s\right] \overline{y}^f(s)\mathrm{d}s.
\end{aligned} \tag{33}$$

■

Proof of Claim 26. Just observe that $\tilde{f}(x^f_{(r)})$ is the r-th order statistic (r-th largest) of n uniformly distributed random variables on $[0, 1]$. Chernoff bound yields the result. ■

Proof of Claim 27.

Let $l = \sum_{(x,y)\in S_n} \mathbb{I}[y = 1]$ be the number of $y = 1$ in S_n. Since S_n is sampled i.i.d. and $\Pr[Y = 1] = p$, by Chernoff bound we have

$$\Pr\left[\left|l/n - p\right| > n^{-1/3}\right] \leq 2e^{-2n^{1/3}}. \tag{34}$$

Thus with probability at least $1 - 2e^{-2n^{1/3}}$

$$\begin{aligned}
&\left|\mathrm{NDCG}_D(f, S_n) - \frac{\mathrm{DCG}_D(f, S_n)}{F(np)}\right| \\
= \;& \left|\frac{\mathrm{DCG}_D(f, S_n)}{l} - \frac{\mathrm{DCG}_D(f, S_n)}{F(np)}\right| \\
\leq \;& \mathrm{DCG}_D(f, S_n) \cdot \max\left(\left|\frac{1}{F(n(p - n^{-1/3}))} - \frac{1}{F(np)}\right|, \left|\frac{1}{F(n(p + n^{-1/3}))} - \frac{1}{F(np)}\right|\right).
\end{aligned}$$

Recall that $F(t) = \int_1^t \frac{1}{\log(1+r)}\mathrm{d}r$, $p > 0$; and observe that $\mathrm{DCG}_D(f, S_n) \leq F(n)$. Taylor expansion of $\frac{1}{F((p\pm n^{-1/3})n)}$ at np and some simple calculations yields the result.

■

Proof of Claim 28.

Integration by part we have,

$$\int \log^k x \mathrm{d}x = k! \sum_{j=0}^{k} (-1)^{k-j} \frac{x \log^j x}{j!} + C. \tag{35}$$

The claim follows. ■

Proof of Claim 29.

Changing variable by letting $x = n^{-t}$ we have

$$\begin{aligned}
&\int_{\frac{2}{n}}^{1} \frac{\left|\log^k x\right| \mathrm{d}x}{(\log(nx))^{k+1}} \\
= \;& \int_0^{1-\frac{\log 2}{\log n}} \frac{t^k}{(1-t)^{k+1}} e^{-t \log n} \mathrm{d}t \\
= \;& \int_0^{1/2} \frac{t^k}{(1-t)^{k+1}} e^{-t \log n} \mathrm{d}t + \int_{1/2}^{1-\frac{\log 2}{\log n}} \frac{t^k}{(1-t)^{k+1}} e^{-t \log n} \mathrm{d}t.
\end{aligned} \tag{36}$$

Now we upper bound the two terms in the last line of eq. (36) separately. For the first term we have

$$\begin{aligned}
&\int_0^{1/2} \frac{t^k}{(1-t)^{k+1}} e^{-t\log n} \mathrm{d}t \le 2^{k+1} \int_0^{1/2} t^k e^{-t\log n} \mathrm{d}t \\
\le\ & \frac{2^{k+1}}{(\log n)^{k+1}} \int_0^{\infty} \tau^k e^{-\tau} \mathrm{d}\tau \le \frac{2^{k+1}\Gamma(k+1)}{(\log n)^{k+1}} \\
=\ & O\left(\frac{1}{(\log n)^{k+1}}\right),
\end{aligned} \tag{37}$$

where Γ is the gamma function, and the last inequality is due to that k is a fixed integer.

For the second term we have

$$\begin{aligned}
&\int_{1/2}^{1-\frac{\log 2}{\log n}} \frac{t^k}{(1-t)^{k+1}} e^{-t\log n} \mathrm{d}t \le \left(\frac{\log n}{\log 2}\right)^{k+1} \int_{1/2}^{1} e^{-t\log n} \mathrm{d}t \\
\le\ & \frac{1}{2} \cdot \frac{1}{\sqrt{n}} \cdot \left(\frac{\log n}{\log 2}\right)^{k+1} = \tilde{O}\left(\frac{1}{\sqrt{n}}\right),
\end{aligned} \tag{38}$$

where in $\tilde{O}$ we hide the polylog(n) terms.

Combining (37) and (38) we complete the proof.

■

Proof of Claim 30.

We only need to show that for any $f \in L^2[0,1]$, if

$$\int_0^1 f(x) \log^k x \mathrm{d}x = 0, \qquad k = 0, 1, \dots \tag{39}$$

then $f = 0$ a.e. on $[0,1]$.

Let $t = -\log x$, then eq.(39) becomes

$$\int_0^{\infty} f(e^{-t}) t^k e^{-t} \mathrm{d}t = 0, \qquad k = 0, 1, \dots$$

Note that Laguerre polynomials form a complete basis of $L^2[0,\infty)$ (cf. (Sansone, 1959), p.349) , thus $\{t^k\}_{k\ge 0}$ is complete in $L^2[0,\infty)$ with respect to measure e^{-t}. The claim follows. ■

Proof of Claim 31.

$$
\begin{aligned}
& \Big|\sum_{r=1}^{n}\overline{y}^f(1-r/n)D(r)-\tilde{N}_D^f(n)\Big| \\
= \; & \Big|\sum_{r=1}^{n}\overline{y}^f(1-r/n)D(r)-\int_1^n \overline{y}^f(1-s/n)D(s)\mathrm{d}s\Big| \\
= \; & \Big|\sum_{r=1}^{n-1}\int_r^{r+1}\Big(\overline{y}^f(1-r/n)D(r)-\overline{y}^f(1-s/n)D(s)\Big)\,\mathrm{d}s\Big|+\overline{y}^f(0)D(n) \\
\le \; & \Big|\sum_{r=1}^{n-1}\int_r^{r+1}\overline{y}^f(1-s/n)(D(r)-D(s))\mathrm{d}s\Big| \\
& +\sum_{r=1}^{n-1}\int_r^{r+1}\Big|\overline{y}^f(1-r/n)-\overline{y}^f(1-s/n)\Big|D(r)\mathrm{d}s+\overline{y}^f(0)D(n) \\
\le \; & \sum_{r=1}^{n-1}\int_r^{r+1}\Big|D(r)-D(s)\Big|\mathrm{d}s+Cn^{-\alpha/3}\sum_{r=1}^{n-1}D(r)+D(n) \\
\le \; & \sum_{r=1}^{n-1}|D'(r)|+Cn^{-\alpha/3}F(n)+D(n) \\
\le \; & Cn^{-\alpha/3}F(n)+|D'(1)|+\sum_{r=2}^{n}|D'(r)|+D(n) \\
\le \; & Cn^{-\alpha/3}F(n)+|D'(1)|+D(1)-D(n)+D(n) \\
= \; & Cn^{-\alpha/3}F(n)+|D'(1)|+D(1).
\end{aligned}
$$

Note that the sixth and the seventh line are both because $|D'(r)|$ is monotone decreasing; and second line from bottom is because $D(r)$ is monotone decreasing. ■

Appendix E. Proof of the Convergence Theorems

In this section we give the proof of the theorems considering convergence of NDCG with various discount and cut-off.

First we give the proof of Theorem 4, i.e., the standard NDCG converges to 1 almost surely for every ranking function.

Proof of Theorem 4. For notational simplicity we only prove for the case $\mathcal{Y}=\{0,1\}$. Generalization is straightforward. Recall that $S_n=\{(x_1,y_1),\cdots,(x_n,y_n)\}$ consists of n i.i.d. instance-label pairs drawn from an underlying distribution P_{XY}. Let $p=\Pr(Y=1)$. Also let $l=\sum_{i=1}^n y_i$. If $p=0$, the theorem trivially holds. Suppose $p>0$, by Chernoff bound we have

$$
\Pr\left(\left|\frac{l}{n}-p\right|>n^{-1/3}\right)\le 2e^{-2n^{1/3}}.
$$

For fixed n, conditioned on the event that $\left|\frac{l}{n} - p\right| \le n^{-1/3}$, by the definition of NDCG, it is easy to see that

$$\begin{aligned}
\mathrm{NDCG}_D(f, S_n) &= \frac{\sum_{r=1}^{n} y_{(r)}^f \frac{1}{\log(1+r)}}{\sum_{r=1}^{l} \frac{1}{\log(1+r)}} \\
&\ge \frac{\sum_{r=n-l+1}^{n} \frac{1}{\log(1+r)}}{\sum_{r=1}^{l} \frac{1}{\log(1+r)}} \\
&\ge \frac{\mathrm{Li}(n+1) - \mathrm{Li}(n(1-p+n^{-1/3})+1)}{\mathrm{Li}(n(p+n^{-1/3})+1)} - o(1) \\
&\ge 1 - o(1), \qquad (40)
\end{aligned}$$

where $\mathrm{Li}(t) = \int_2^t \frac{\mathrm{d}\tau}{\log \tau}$ is the offset logarithmic integral function; and the last step in eq.(40) is due to the well-known fact that $\mathrm{Li}(t) \sim \frac{t}{\log t}$. Thus for any $\epsilon > 0$, and for any sufficiently large n, conditioned on the event that $\left|\frac{1}{n}\sum_{i=1}^{n} y_i - p\right| \le n^{-1/3}$, we have

$$\mathrm{NDCG}_D(f, S_n) \ge 1 - \epsilon.$$

Also recall that $\mathrm{NDCG}_D(f, S_n) \le 1$. We have, for any $\epsilon > 0$ and every sufficiently large n

$$\Pr\left(|\mathrm{NDCG}_D(f, S_n) - 1| \ge \epsilon\right) \le 2e^{-2n^{1/3}}.$$

Since $\sum_{n\ge 1} 2e^{-2n^{1/3}} < \infty$, by Borel-Cantelli lemma $\mathrm{NDCG}_D(f, S_n)$ converges to 1 almost surely. ■

Next we give details of the other feasible discount functions as well as the cut-off versions. In particular, we provide proofs of Theorems 7, 9, 11, 12, 13. The proofs of these five theorems are quite similar. We only prove Theorem 7 to illustrate the ideas. The proof of the other four theorems require only minor modifications.

The proof of Theorem 7 relies on the following lemma, which is similar to Lemma 23.

Lemma 32 *Let $D(r) = r^{-\beta}$ for some $\beta \in (0, 1)$. Assume that $p = \Pr(Y = 1) > 0$. If the ranking function f satisfies that $\overline{y}^f(s) = \Pr(Y = 1|\tilde{f}(X) = s)$ is continuous, then for every $\epsilon > 0$ the following inequality holds for all sufficiently large n:*

$$\Pr\left[\left|\mathrm{NDCG}_D(f, S_n) - N_D^f(n)\right| \ge 5p^{-1}\epsilon\right] \le o(1).$$

Proof of Theorem 7. The theorem follows from Lemma 32 and simple calculations of $\lim_{n\to\infty} N_D^f(n)$. We omit the details. ■

Proof of Lemma 32. The proof is simple modification of the proof of Lemma 23. Note that the difference of Lemma 32 from Lemma 23 is that here we do not assume y^f is Hölder continuous. We only assume it is continuous.

Next observe that Claim 31 holds for $D(r) = r^{-\beta}$ $(0 < \beta < 1)$ as well. Because in the proof of Claim 31, we only use two properties of $D(r)$. That is, $D(r)$ is monotone decreasing and $|D'(r)|$ is monotone decreasing. Clearly $D(r) = r^{-\beta}$ satisfies these properties. But here $\overline{y}^f(s)$ is merely continuous rather than Hölder continuous. Thus we have a modified version of Claim 31. That is, for every $\epsilon > 0$, the following holds for all sufficiently large n:

$$\Big|\sum_{r=1}^{n} \overline{y}^f(1 - r/n)D(r) - \tilde{N}_D^f(n)\Big| \le \epsilon F(n) + D(1) + |D'(1)|.$$

The rest of the proof are almost identical to Lemma 23. We omit the details. ■

Finally, we give the proof of Theorem 10, i.e., if the discount decays substantially faster than r^{-1}, then the NDCG measure does not converge. Moreover, every pair of ranking functions are not strictly distinguishable with high probability by the measure.

Proof of Theorem 10. For notational simplicity we give a proof for $|\mathcal{Y}| = 2$ and $\mathcal{Y} = \{0, 1\}$. It is straightforward to generalize it to other cases.

In fact, we only need to show that for every ranking function f, there are constants $a, b, c > 0$ with $a > b$, such that for all sufficiently large n,

$$\Pr[\mathrm{NDCG}_D(f, S_n) \ge a] \ge c$$

and

$$\Pr[\mathrm{NDCG}_D(f, S_n) \le b] \ge c$$

both hold. Once we prove this, by definition the ranking measure does not converge (in probability). Also, it is clear that for every pair of ranking functions, there is at least a constant probability that the ranking measure of the two functions are "overlap". Therefore distinguishability is not possible.

For sufficiently large n, fix any $x_1, \ldots, x_n$. According to the assumption, the probability that the top-ranked m data all have label 1 is at least $(\delta/2)^m$, where m is the minimal integer such that

$$\sum_{r=1}^{m} D(r) \ge \frac{2}{3} \sum_{r=1}^{\infty} D(r).$$

Clearly we have

$$\Pr\left(\mathrm{NDCG}_D(f, S_n) \ge \frac{2}{3} \;\middle|\; x_1, \ldots, x_n\right) \ge (\delta/2)^m.$$

On the other hand, the probability that the top-ranked m elements all have label 0 and there are at least m elements in the list that have label 1 is at least $(\delta/2)^{2m}$. Note that

$$\frac{\sum_{r=m+1}^{n} D(r)}{\sum_{r=1}^{m} D(r)} \le \frac{1}{2}.$$

Thus we have

$$\Pr[\mathrm{NDCG}_D(f, S_n) \le \frac{1}{2} \mid x_1, \ldots, x_n] \ge (\delta/2)^{2m}.$$

Since $x_1, \ldots, x_n$ are arbitrary, the theorem follows. ■

Appendix F. Proof of Distinguishability for NDCG with $r^{-\beta}$ ($\beta \in (0,1)$) Discount

Here we give the proof of Theorem 8, i.e., NDCG with $r^{-\beta}$ $(0 < \beta < 1)$ discount has the power of distinguishability.

Proof of Theorem 8. The proof of distinguishability for polynomial discount is much easier than that of the logarithmic discount, because in the former case the pseudo-expectation has very simple form. If f_0 and f_1 satisfy the first condition $\int_0^1 \Delta y(s)(1-s)^{-\beta}\mathrm{d}s \neq 0$, then the theorem is trivially true since $\mathrm{NDCG}(f_0, S_n)$ and $\mathrm{NDCG}(f_1, S_n)$ converge to different limits. So we only need to prove the theorem assuming that $\int_0^1 \Delta y(s)(1-s)^{-\beta}\mathrm{d}s = 0$ and the second condition holds. The proof is similar to Theorem 6 by using the pseudo-expectation. We have the the next two lemmas for discount $D(r) = r^{-\beta}$, $\beta \in (0,1)$.

Lemma 33 *Let $D(r) = r^{-\beta}$, $\beta \in (0,1)$. Suppose that $\overline{y}^{f_0}(s)$ and $\overline{y}^{f_1}(s)$ are continuous. Also assume that $\int_0^1 \Delta y(s)(1-s)^{-\beta}\mathrm{d}s = 0$ and $\Delta y(1) \neq 0$. Then we have*

$$\left|N_D^{f_0}(n) - N_D^{f_1}(n)\right| \geq \left|\frac{\Delta y(1)}{2p^{1-\beta}}\right| \cdot n^{-(1-\beta)}. \tag{41}$$

Proof

$$\begin{aligned} N_D^{f_0}(n) - N_D^{f_1}(n) &= \frac{n}{F(np)} \int_{1/n}^{1} \Delta y(1-s) \cdot (ns)^{-\beta}\mathrm{d}s \\ &= \frac{1-\beta}{p^{1-\beta}} \int_{1/n}^{1} \Delta y(1-s) \cdot s^{-\beta}\mathrm{d}s \\ &= -\frac{1-\beta}{p^{1-\beta}} \int_{0}^{1/n} \Delta y(1-s) \cdot s^{-\beta}\mathrm{d}s. \end{aligned}$$

Since Δy is continuous, for any $\delta > 0$ there exists $\epsilon > 0$ such that for all $x \in [1-\epsilon, 1]$, $|\Delta y(x) - \Delta y(1)| \leq \delta$. Consequently, for sufficiently large n,

$$\Big| \int_0^{1/n} \Delta y(1-s) \cdot s^{-\beta}\mathrm{d}s - \Delta y(1) \cdot \int_0^{1/n} s^{-\beta}\mathrm{d}s \Big| \leq \delta \cdot \int_0^{1/n} s^{-\beta}\mathrm{d}s.$$

Let $\delta = \Delta y(1)/2$, we then have

$$\Big| \int_0^{1/n} \Delta y(1-s) \cdot s^{-\beta}\mathrm{d}s - \frac{\Delta y(1)}{1-\beta} \cdot n^{-(1-\beta)} \Big| \leq \frac{\Delta y(1)}{2(1-\beta)} \cdot n^{-(1-\beta)}.$$

The lemma follows. ∎

Lemma 34 *Let $D(r) = r^{-\beta}$, $\beta \in (0,1)$. Assume that $p = \Pr(Y=1) > 0$. If the ranking function f satisfies that $\overline{y}^f(s) = \Pr(Y = 1 | \tilde{f}(X) = s)$ is Hölder continuous with constants $\alpha > 0$ and $C > 0$ That is, $|\overline{y}^f(s) - \overline{y}^f(s')| \leq C|s - s'|^\alpha$ for all $s, s' \in [0,1]$. Then*

$$\Pr\left[\left|\mathrm{NDCG}_D(f, S_n) - N_D^f(n)\right| \geq 5Cp^{-1}n^{-\min(\alpha/3,1)}\right] \leq O\left(e^{-n^{(1-\beta)/3}}\right).$$

Proof The proof is almost the same as the proof of Lemma 23 ■

The theorem follows immediately from Lemma 33 and Lemma 34. ■

References

S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under an ROC curve. 2004.

A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *SIGIR*, pages 773–774, 2007.

J.A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34. ACM, 2005.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*, volume 82. Addison-Wesley New York, 1999.

M.F. Balcan, N. Bansal, A. Beygelzimer, D. Coppersmith, J. Langford, and G.B. Sorkin. Robust reductions from ranking to classification. *Machine learning*, 72(1):139–153, 2008.

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

D. Buffoni, C. Calauzenes, P. Gallinari, and N. Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

C.J.C. Burges, R. Ragno, and Q. V. Le. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, page 193. The MIT Press, 2007.

Clément Calauzènes, Nicolas Usunier, and Patrick Gallinari. On the (non-)existence of convex, calibrated surrogate losses for ranking. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 197–205. 2012.

Z. Cao, T. Qin, T.Y. Liu, M.F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136, 2007.

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*, pages 621–630. ACM, 2009.

Stéphan J.M. Clémençon and Nicolas Vayatis. Empirical performance maximization for linear rank statistics. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 305–312. 2009.

S. Clemençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

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

K. Crammer and Y. Singer. Pranking with ranking. In *Advances in Neural Information Processing Systems*, 2002.

W.B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley, 2010.

J. Duchi, L. Mackey, and M.I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, pages 327–334, 2010.

Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, 4:933–969, 2003. ISSN 1532-4435.

J. Hájek, Z. Šidák, and P.K. Sen. *Theory of rank tests*. Academic press New York, 1967.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Neural Information Processing Systems*, pages 115–132, 1999.

K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48. ACM, 2000.

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

T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

Evangelos Kanoulas and Javed A. Aslam. Empirical justification of the gain and discount function for NDCG. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 611–620. ACM, 2009.

M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

R. Nallapati. Discriminative models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71. ACM, 2004.

P. Ravikumar, A. Tewari, and E. Yang. On NDCG consistency of listwise ranking methods. In *Proceedings of 14th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2011.

C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *The Journal of Machine Learning Research*, 10:2233–2271, 2009.

T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 525–532. ACM, 2006.

G. Sansone. *Orthogonal Functions*. Interscience Publishers Inc., New York, 1959.

A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18. ACM, 2006.

H. Valizadegan, R. Jin, R. Zhang, and J. Mao. Learning to rank by optimizing NDCG measure. *Advances in Neural Information Processing Systems*, 22:1883–1891, 2009.

E.M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82. ACM, 2001.

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*, pages 1192–1199. ACM, 2008.

Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 271–278. ACM, 2007.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004a.

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

K. Zhou, H. Zha, Y. Chang, and G.R. Xue. Learning the gain values and discount factors of discounted cumulative gains. *IEEE Transactions on Knowledge and Data Engineering, to appear.*