On NDCG Consistency of Listwise Ranking Methods

Pradeep Ravikumar
University of Texas at Austin

Ambuj Tewari
University of Texas at Austin

Eunho Yang
University of Texas at Austin

Abstract

We study the consistency of listwise ranking methods with respect to the popular Normalized Discounted Cumulative Gain (NDCG) criterion. State of the art listwise approaches replace NDCG with a surrogate loss that is easier to optimize. We characterize NDCG consistency of surrogate losses to discover a surprising fact: several commonly used surrogates are NDCG inconsistent. We then show how to modify them so that they become NDCG consistent. We then state a stronger but more natural notion of strong NDCG consistency, and surprisingly are able to provide an explicit characterization of *all* strongly NDCG consistent surrogates. Going beyond qualitative consistency considerations, we also give quantitive statements that enable us to transform the excess error, as measured in the surrogate, to the excess error in comparison to the Bayes optimal ranking function for NDCG. Finally, we also derive improved results if a certain natural "low noise" or "large margin" condition holds.

Our experiments demonstrate that ensuring NDCG consistency does improve the performance of listwise ranking methods on real-world datasets. Moreover, a novel surrogate function suggested by our theoretical results leads to further improvements over even NDCG consistent versions of existing surrogates.

1 Introduction

Ranking a set of instances by their relative relevance arises in many contemporary problems, such as collaborative filtering, text mining and document retrieval. We[1] are interested in a particular formulation of this problem, natural in information retrieval (IR), where the ranking is at the resolution of a data item such as a query. Each query has a list of documents, and the task is to rank these documents in the order of relevance to the query. In the training set, the documents for each query are typically represented as feature vectors derived from the query-document pairs, and are annotated with relevance scores indicating the relative preference of the document in the list for that query. Given any new query, the goal is to rank its documents in an order that best respects their relevance scores according to some ranking evaluation measure. User studies have motivated specific ranking evaluation measures such as Mean Average Precision (MAP) [3], Expected Reciprocal Rank [8] and the popular Normalized Discounted Cumulative Gain (NDCG) [15].

In this paper, we study the NDCG evaluation measure, which evaluates the ranking of the entire list of documents by penalizing errors in higher ranked documents more strongly. While easy to *evaluate*, this is nonetheless a difficult measure to directly use for *training* a ranking model. A broad line of work has thus focused on breaking the ranking problem down into pointwise and pairwise problems [6]. In the *pointwise* approach, the ranking problem is viewed as a regression or classification problem of predicting the specific relevance score for any document [24]. The hope is that by minimizing some measure of the difference between each document's true relevance level and the model's estimate for it, the *listwise* NDCG measure of the ranking of the entire list of the documents would in turn be minimized. Examples include linear regression minimizing mean squared error, and ordinal regression. In the *pairwise* approach on the other hand, the ranking problem is reduced to the binary classification task of predicting the more relevant document amongst pairs of documents. Note that the training data for such an approach would only need pairwise relative preferences, which is easier to obtain than listwise relevance judgements, for instance using

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

[1] Author order is alphabetical reflecting equal contributions to this work.

query log click-through data [16]. But on the other hand, such training instances of document pairs and their pairwise relative relevances are typically not *iid* which impairs test performance.

The main caveat with such approaches is that they are ill-suited to the *listwise* NDCG evaluation measure that is a function of the entire list of ranked documents. Cao et al. [6], Xia et al. [23] in particular note that methods based on *listwise* loss functions outperform their pointwise and pairwise counterparts. Accordingly, one class of such *listwise* approaches attempt to optimize the NDCG (and such) evaluation measures directly using heuristics [5, 24, 20, 22, 21].

The state of the art set of *listwise* approaches however optimize surrogate listwise loss functions instead [19, 6, 23], motivated in part by successes of such an approach in classification. The use of such surrogate ranking loss functions gives rise to the main question of this paper: when are surrogate loss functions consistent with respect to the NDCG evaluation measure, and which classes of surrogate loss functions are better suited to the NDCG evaluation measure under finite samples? A line of recent results have studied consistency and Bayes optimality of estimates for the cases where the target evaluation measure is pointwise [10], and when it is pairwise [9, 12], and for the zero-one listwise loss [6]. In this paper, we study the consistency of any surrogate ranking loss function with respect to the listwise NDCG evaluation measure.

We first provide a characterization of **any** NDCG consistent ranking estimate: it has to match the sorted order of the expectation of the document relevance levels normalized by a particular *DCG* norm. As we show, this normalization provides a stabilizing effect on the ranking estimates, and suggests why user studies have validated the NDCG measure to some extent. However it turns out that many popular listwise surrogate loss functions such as Cosine [19] and ListNet [6] do **not** yield NDCG consistent ranking estimates, primarily because they employ a different normalization of the expected relevance scores. We then show how simple modifications of these methods then make them NDCG consistent. In our second set of results, we implicitly characterize the set of **all** NDCG consistent surrogate loss functions. We then define a slightly stronger, and as we point out more natural, notion of consistency called strong-NDCG consistency. We then provide an **explicit** characterization of the set of **all** such strong-NDCG consistent surrogate loss functions. In our final set of results, for the strong-NDCG family of surrogate loss functions, we provide explicit transform functions relating the excess surrogate error of their ranking estimates, to their deviation in NDCG error from the Bayes optimal ranking estimate. This not only proves consistency of these loss functions, but provides a means for quantitatively comparing different surrogate loss functions. Indeed, these transforms can be used to provide explicit convergence rate bounds though we defer this due to lack of space. Finally, we provide a notion of "low-noise" or "large-margin" distributions under which we are able to derive much tighter transforms.

In gist, we provide an extensive quantitative analysis of NDCG consistency of surrogate ranking loss functions. The resulting characterizations posit new methods as well. In Section 3.2, we show that the surrogate losses of linear regression (minimizing mean squared error), Cosine and Listnet are not NDCG consistent, and then provide simple modifications to make them NDCG consistent. In our first set of experiments, we compared the NDCG performance of these three loss functions with their counterparts with our modifications on multiple datasets, and largely show improvement across these datasets. In Section 4, we propose a family of NDCG consistent surrogates, and highlight a member of that family, which to the best of our knowledge has not been studied before. In our second set of experiments, we compare this novel surrogate loss to Cosine and squared-error loss functions, and again largely show improvement across datasets.

2 Preliminaries

Let m be the number of documents for each query. Let $\bar{\mathcal{X}}$ be the space of the feature vectors in which the documents are represented (typically derived from the query-document pairs). Let $\bar{\mathcal{R}} \subseteq \mathbb{R}$ be the space of the relevance scores each document receives. Thus for any query, we have a list $\mathbf{X} = (X_1, \ldots, X_m) \in \mathcal{X} := \bar{\mathcal{X}}^m$ of document feature vectors, and a corresponding list $\mathbf{R} = (R_1, \ldots, R_m) \in \mathcal{R} := \bar{\mathcal{R}}^m$ of document relevance scores. The dataset consists of n $(\mathbf{X}_i, \mathbf{R}_i)$ pairs which we assume to be drawn *iid* from some distribution over $\mathcal{X} \times \mathcal{R}$.

A permutation π is a bijection from $[m]$ to $[m]$. We interpret $\pi(i)$ as "the position of document i". Thus, according to π, the documents $\mathbf{x} = (x_1, \ldots, x_m)$ should be ordered as $(x_{\pi^{-1}(1)}, \ldots, x_{\pi^{-1}(i)}, \ldots, x_{\pi^{-1}(m)})$.

Let $\mathcal{P}_m$ be the set of all such degree m permutations. A listwise ranking evaluation metric measures the goodness of fit of any candidate ranking to the corresponding relevance scores, so that it is a map $\ell : \mathcal{P}_m \times \mathcal{R} \mapsto \mathbb{R}$. We are interested in the NDCG class of ranking loss functions:

Definition 1 (NDCG-like loss functions)**.**

$$\ell_{\mathrm{NDCG}}(\pi, \mathbf{r}) = -\frac{1}{Z(\mathbf{r})} \sum_{j=1}^{m} \frac{G(r_j)}{F(\pi(j))}, \tag{1}$$

where $G : \mathcal{R} \mapsto \mathbb{R}_+$ is a monotonically increasing function of the relevance judgments, and $F : \mathbb{R} \mapsto \mathbb{R}_+$ is also a monotonically increasing function. The normalization $Z(\mathbf{r})$ is the highest possible DCG value:

$$Z(\mathbf{r}) = \max_{\pi \in \mathcal{P}_m} \sum_{j=1}^{m} \frac{G(r_j)}{F(\pi(j))}$$

The NDCG criterion [15] uses $G(r) = 2^r - 1$, and $F(t) = \log(1+t)$, but in the sequel we allow $G(\cdot)$ and $F(\cdot)$ to be any general monotonic functions. We will use $G(\mathbf{r})$ to denote $(G(r_1), \ldots, G(r_m))^\top$.

We begin with a simple observation. All proofs omitted from the main paper can be found in the appendix.

Lemma 1. *The function $Z(\mathbf{r})$ can be written as $\|G(\mathbf{r})\|_D$ for a norm $\|\cdot\|_D$.*

We will need a notion of when the sorted order of one vector $\mathbf{s}$ is compatible with the sorted order of a given vector $\mathbf{r}$. This *assymetric* binary relation between $\mathbf{s}$ and $\mathbf{r}$ is denoted by $\mathbf{s} \rightsquigarrow \mathbf{r}$, and it holds precisely when, for all $i, j \in [m]$, $r_i > r_j$ implies $s_i > s_j$. We will call a map $g : \mathbb{R}^m \to \mathbb{R}^m$ *order preserving*, iff $g(\mathbf{r}) \rightsquigarrow \mathbf{r}$ for all $\mathbf{r} \in \mathbb{R}^m$. We will also need the following lemma whose proof is elementary.

Lemma 2. $\mathbf{s} \rightsquigarrow \mathbf{r}$ *iff there is an invertible order preserving map g such that $\mathbf{s} = g(\mathbf{r})$.*

3 NDCG Consistency

Note that the first argument of ℓ_{NDCG} as defined in (1) is a permutation. It is useful, both for learning and optimization, to define it as a function of a real-valued score vector instead. Indeed with some overloading of notation we can define $\ell_{\text{NDCG}}(\mathbf{s}, \mathbf{r}) = \ell_{\text{NDCG}}(\pi_{\mathbf{s}}, \mathbf{r})$ where $\pi_{\mathbf{s}}$ is a permutation such that $\pi_{\mathbf{s}}(j)$ is the position of s_j when elements of $\mathbf{s}$ are sorted in decreasing order of their values. Note that now the first argument is a real-valued score vector. Unfortunately, the above function is still difficult to optimize, since it depends in a complicated manner on $\mathbf{s}$, and is not convex in $\mathbf{s}$. This has thus motivated the search for convex surrogate ranking loss functions. A convex surrogate is simply a function $\phi : \mathbb{R}^m \times \mathbb{R}^m$ that is chosen as a proxy for the NDCG loss. To ascertain whether the surrogate is indeed a good proxy, we need the notion of NDCG-consistency.

Given any **ranking function** $\mathbf{f} : \mathcal{X} \to \mathcal{R}$, define the expected NDCG loss as,

$$L_{\text{NDCG}}(\mathbf{f}) = \mathbb{E}\left[\ell_{\text{NDCG}}(\mathbf{f}(\mathbf{X}), \mathbf{R})\right] \ .$$

Similarly, for a surrogate ϕ, define the expected surrogate loss as,

$$\Phi(\mathbf{f}) = \mathbb{E}\left[\phi(\mathbf{f}(\mathbf{X}), \mathbf{R})\right] \ .$$

Denote the minimum expected losses by

$$L^\star_{\text{NDCG}} = \min_{\mathbf{f}} L_{\text{NDCG}}(\mathbf{f}) \qquad \Phi^\star = \min_{\mathbf{f}} \Phi(\mathbf{f}) \ ,$$

where we are assuming, for simplicity, that the minimum over all measurable $\mathbf{f}$ is achieved.

Definition 2. *A surrogate ϕ is said to be NDCG consistent if for any distribution on $\mathcal{X} \times \mathcal{R}$, and for any sequence $\mathbf{f}_n$ such that*

$$\Phi(\mathbf{f}_n) \to \Phi^\star$$

we necessarily have

$$L_{\text{NDCG}}(\mathbf{f}_n) \to L^\star_{\text{NDCG}}$$

Thus a surrogate is NDCG-consistent if the ranking estimate minimizing the surrogate loss in turn is Bayes consistent with respect to the NDCG loss.

Some commonly used surrogates are:

$$\begin{aligned}
\phi_{\text{cos}}(\mathbf{s}, \mathbf{r}) &= 1 - \frac{\mathbf{s}}{\|\mathbf{s}\|_2} \cdot \frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_2} && \text{(Cosine)} \\
\phi_{\text{sq}}(\mathbf{s}, \mathbf{r}) &= \|\mathbf{s} - G(\mathbf{r})\|_2^2 && \text{(Least Squares)} \\
\phi_{\text{list}}(\mathbf{s}, \mathbf{r}) &= \text{KL}(\mathbf{r}' \| \mathbf{s}') && \text{(ListNet/Cross Entropy)}
\end{aligned}$$

In the last example, $\mathbf{r}', \mathbf{s}'$ are probability vectors derived from $\mathbf{r}, \mathbf{s}$ as follows: $r'_j = \exp(r_j) / \sum_k \exp(r_k)$, $s'_j = \exp(s_j) / \sum_k \exp(s_k)$.

3.1 Fisher Optimal Ranking Functions

Any probability distribution on $\mathcal{X} \times \mathcal{R}$ is fully specified by the marginal μ on $\mathcal{X}$ and the conditional $\eta_{\mathbf{x}}(\mathbf{r})$ on $\mathcal{R}$, i.e. $P((\mathbf{X}, \mathbf{R}) = (\mathbf{x}, \mathbf{r})) = \mu(\mathbf{x}) \cdot \eta_{\mathbf{x}}(\mathbf{r})$. For any score vector $\mathbf{s}$ and any ditribution η on $\mathcal{R}$, define

$$\bar{\ell}_{\text{NDCG}}(\mathbf{s}; \eta) = \mathbb{E}_{\mathbf{r} \sim \eta}\left[\ell_{\text{NDCG}}(\mathbf{s}, \mathbf{r})\right] \ ,$$

$$\bar{\phi}(\mathbf{s}; \eta) = \mathbb{E}_{\mathbf{r} \sim \eta}\left[\phi(\mathbf{s}, \mathbf{r})\right] \ .$$

Also define the minimum losses

$$\bar{\ell}^\star_{\text{NDCG}}(\eta) = \min_{\mathbf{s}} \bar{\ell}_{\text{NDCG}}(\mathbf{s}; \eta) \quad \bar{\phi}^\star(\eta) = \min_{\mathbf{s}} \bar{\phi}(\mathbf{s}; \eta) \ .$$

We have again assumed that these minima are achieved. To make the analysis less cumbersome, we will make a few more technical assumptions. First, we assume that the minimum of $\bar{\phi}(\mathbf{s}; \eta)$ is always achieved at a unique point $\mathbf{s}^\star_\phi(\eta)$. Moreover, for any sequence such that

$$\bar{\phi}(\mathbf{s}_n; \eta) \to \bar{\phi}^\star(\eta)$$

we assume that it must be the case that $\mathbf{s}_n \to \mathbf{s}^\star_\phi(\eta)$.

Note that, with these definitions, we have

$$L_{\text{NDCG}}(\mathbf{f}) = \mathbb{E}\left[\bar{\ell}_{\text{NDCG}}(\mathbf{f}(\mathbf{X}); \eta_{\mathbf{X}})\right] \quad \Phi(\mathbf{f}) = \mathbb{E}\left[\bar{\phi}(\mathbf{f}(\mathbf{X}); \eta_{\mathbf{X}})\right] \ .$$

and

$$L^\star_{\mathrm{NDCG}} = \mathbb{E}\left[\bar{\ell}^\star_{\mathrm{NDCG}}(\eta_{\mathbf{X}})\right] \qquad \Phi^\star = \mathbb{E}\left[\bar{\phi}^\star(\eta_{\mathbf{X}})\right] .$$

We now give an equivalent characterization of NDCG consistency of a surrogate.

Lemma 3. *A surrogate ϕ is NDCG consistent iff for any distribution η on $\mathcal{R}$ and any sequence $\mathbf{s}_n$ such that*

$$\bar{\phi}(\mathbf{s}_n; \eta) \to \bar{\phi}^\star(\eta)$$

we have

$$\bar{\ell}_{\mathrm{NDCG}}(\mathbf{s}_n; \eta) = \bar{\ell}^\star_{\mathrm{NDCG}}(\eta) ,$$

for n large enough.

The next lemma identifies the set of scores that maximize $\bar{\ell}_{\mathrm{NDCG}}(\cdot; \eta)$ for a given distribution η.

Lemma 4. *Fix a distribution η over $\mathcal{R}$. Then,*

$$\bar{\ell}_{\mathrm{NDCG}}(\mathbf{s}; \eta) = \bar{\ell}^\star_{\mathrm{NDCG}}(\eta)$$

iff

$$\mathbf{s} \rightsquigarrow \mathbb{E}_{\mathbf{r}\sim\eta}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right] .$$

An immediate corollary of the above lemma is the identification of the **Fisher optimal** ranking functions minimizing NDCG loss.

Corollary 5. *A function $\mathbf{f} : \mathcal{X} \to \mathcal{R}$ satisfies*

$$L_{\mathrm{NDCG}}(\mathbf{f}) = L^\star_{\mathrm{NDCG}}$$

iff

$$\mathbf{f}(\mathbf{X}) \rightsquigarrow \mathbb{E}_{\mathbf{r}\sim\eta_{\mathbf{X}}}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right] \quad \mu\text{-almost surely.}$$

This in turn gives a characterization of NDCG-consistent surrogates.

Theorem 6. *A surrogate ϕ is NDCG consistent iff for any distribution η on $\mathcal{R}$, there exists an invertible order preserving map $g : \mathbb{R}^m \to \mathbb{R}^m$ such that the unique minimizer $\mathbf{s}^\star_\phi(\eta)$ of $\bar{\phi}(\mathbf{s}; \eta)$ can be written as*

$$\mathbf{s}^\star_\phi(\eta) = g\left(\mathbb{E}_{\mathbf{r}\sim\eta}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right]\right) .$$

3.2 Inconsistency of common surrogates

We have seen above that the optimal score vector (for minimizing NDCG loss) is *not* obtained simply from the sorted order of $\mathbb{E}[G(\mathbf{r})]$, but rather from the sorted order of the expected "normalized" relevance score vector, i.e. $\mathbb{E}[G(\mathbf{r})/\|G(\mathbf{r})\|_D]$. Here, the normalization is achieved inversely scaling the raw relevance vector by its DCG norm. We argue below that, intuitively, some normalization of the raw relevance scores is needed to derive the optimal ordering in order to have a sort of "robustness" against high relevance values that show up due to noise, or only occasionally. Furthermore, we show that common surrogates are not NDCG consistent precisely because the normalizations they implicitly use are not the same as the one used by NDCG.

3.2.1 The Need for Normalization

To see how normalization helps, it is useful to consider the following simple example. Suppose that $m = 2$ (two documents to be ranked per query). Consider a conditional distribution η that supported on just two vectors: $\binom{5}{4}$ and $\binom{1}{3}$. The probability of the first vector is small, say 0.3 while that of the second is relatively larger, say 0.7. So, in this case, document 1 usually looks much less relevant (relevance level 1 versus level 3) than document 2 but 30% of the time, it is just slightly more relevant (relevance level 5 versus level 4). Should we prefer document 1 to 2? Intuitively, it seems clear that we should not.

But if we do not normalize and simply compute $\mathbb{E}[G(\mathbf{r})]$ for $G(r_i) = 2^{r_i} - 1$ as in NDCG, we get

$$\mathbb{E}[G(\mathbf{r})] = \begin{pmatrix} 0.3 \cdot 31 + 0.7 \cdot 1 \\ 0.3 \cdot 15 + 0.7 \cdot 7 \end{pmatrix} = \begin{pmatrix} 10 \\ 9.4 \end{pmatrix} .$$

According to this, document 1 will be ranked first. For the same example, the expected normalized relevance vector $\mathbb{E}[G(\mathbf{r})/\|G(\mathbf{r})\|_D]$ will be

$$\begin{pmatrix} 0.3 \cdot \frac{31}{31+\frac{15}{\log_2 3}} + 0.7 \cdot \frac{1}{7+\frac{1}{\log_2 3}} \\ 0.3 \cdot \frac{15}{31+\frac{15}{\log_2 3}} + 0.7 \cdot \frac{7}{7+\frac{1}{\log_2 3}} \end{pmatrix} = \begin{pmatrix} 0.3216 \\ 0.7533 \end{pmatrix} .$$

This matches out intuition that document 2 should be ranked first.

3.2.2 Are Common Surrogates Inconsistent?

If we compute the minimizers of $\bar{\phi}(\mathbf{s}; \eta)$ for $\phi = \phi_{\mathrm{cos}}, \phi_{\mathrm{sq}}$ and ϕ_{list}, we find that they rank the documents according to the sorted order of

$$\mathbb{E}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right], \mathbb{E}[G(\mathbf{r})], \text{ and } \mathbb{E}\left[\frac{\exp(\mathbf{r})}{\|\exp(\mathbf{r})\|_1}\right]$$

respectively. Thus, the least squares loss does not use any normalization, while the normalizations used by Cosine and Cross Entropy are different from that used by NDCG. We thus obtain the following surprising result.

Proposition 7. *The three surrogates $\phi_{\mathrm{cos}}, \phi_{\mathrm{sq}}$ and ϕ_{list} are not NDCG consistent.*

We provide explicitly worked out examples demonstrating inconsistency in the appendix.

3.3 Restoring consistency

The following surrogates, obtained by modifying $\phi_{\text{cos}}, \phi_{\text{sq}}$ and ϕ_{list} respectively, are NDCG consistent.

$$\tilde{\phi}_{\text{cos}}(\mathbf{s}, \mathbf{r}) = 1 - \frac{\mathbf{s}}{\|\mathbf{s}\|_2} \cdot \frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D} ,$$

$$\tilde{\phi}_{\text{sq}}(\mathbf{s}, \mathbf{r}) = \left\| \mathbf{s} - \frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D} \right\|_2^2 ,$$

$$\tilde{\phi}_{\text{list}}(\mathbf{s}, \mathbf{r}) = \text{KL}\left(\mathbf{r}' || \mathbf{s}'\right) ,$$

where, the last line, $\mathbf{r}', \mathbf{s}'$ are defined as $r'_j = G(r_j)/\|G(\mathbf{r})\|_D$, $s'_j = \exp(s_j)$ Moreover, we are using the extended definition of KL that defines it over all pairs of positive vectors (not necessarily probability vectors):

$$\text{KL}(\mathbf{p}||\mathbf{q}) = \sum_j p_j \log(p_j/q_j) - \sum_j p_j + \sum_j q_j .$$

Of these, the last two are convex. We will now see that these are just some examples from a large class of consistent surrogates.

4 A Family of NDCG Consistent Surrogates

The NDCG-consistent examples presented in the last section leads naturally to the question: are there other NDCG consistent surrogates? In order to answer this question, we first consider the following notion that we call *strong NDCG consistency.*

Definition 3. *A surrogate ϕ is said to be strongly NDCG consistent if there is an invertible order preserving map $g : \mathbb{R}^m \to \mathbb{R}^m$ such that for any distribution η on $\mathcal{R}$, the unique minimzer $\mathbf{s}^\star_\phi(\eta)$ of $\bar{\phi}(\mathbf{s}; \eta)$ can be written as*

$$\mathbf{s}^\star_\phi(\eta) = g\left(\mathbb{E}_{\mathbf{r}\sim\eta}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right]\right) .$$

Comparing to Theorem 6, the reader might wonder whether the above definition is any different from the usual NDCG consistency. There is, however, a subtle difference: in the above definition the same map g works for all distributions η. We expect any reasonable NDCG consistent surrogate to be actually strongly NDCG consistent: indeed g as a functional of the surrogate ϕ would typically not have knowledge of the distribution η. In fact, this has been true for all our positive examples (except for $\tilde{\phi}_{\text{cos}}$ which does not have a unique minimizers). We remark that any strongly NDCG consistent surrogate is also NDCG consistent.

The following results provides a *complete* characterization of strongly NDCG consistent surrogates. In other words, the family is exhaustive w.r.t. the property of strong NDCG consistency. We first setup some notation. Let $\psi : \mathbb{R}^m \mapsto \mathbb{R}$ be a strictly convex function. Any such function induces a *Bregman divergence* [7] $D_\psi : \mathbb{R}^m \times \mathbb{R}^m \mapsto \mathbb{R}$ as follows:

$$D_\psi(\mathbf{u}, \mathbf{v}) := \quad \psi(\mathbf{u}) - \psi(\mathbf{v}) - \langle (\nabla\psi)^{-1}(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle.$$

A Bregman divergence satisfies $D_\psi(\mathbf{u}, \mathbf{v}) \geq 0$ with equality if and only if $\mathbf{u} = \mathbf{v}$, but need not be symmetric or satisfy the triangle inequality, so it is only a generalized distance. With this notation, we can now state our surprising result: *any* strongly NDCG consistent surrogate has the form of a Bregman divergence:

Theorem 8. *Consider a surrogate of the form $\phi(\mathbf{s}, \mathbf{r}) = \Phi(\mathbf{s}, G(\mathbf{r})/\|G(\mathbf{r})\|_D)$. Then, ϕ is strongly NDCG consistent* **iff**

$$\Phi(\mathbf{s}, \mathbf{u}) = D_\psi(\mathbf{u}, g(\mathbf{s})), \tag{2}$$

for some Bregman divergence D_ψ for some strictly convex ψ and an invertible, order preserving g.

Proof. Since ϕ is strongly NDCG consistent, there is some invertible order preserving map h such that the unique minimizer of $\mathbb{E}\left[\Phi(\mathbf{s}, G(\mathbf{r})/\|G(\mathbf{r})\|_D\right]$ is $h(\mathbb{E}\left[G(\mathbf{r})/\|G(\mathbf{r})\|_D\right])$. Defining the random variable $\mathbf{u} = G(\mathbf{r})/\|G(\mathbf{r})\|_D$, we see that $\mathbb{E}\left[\Phi(\mathbf{s}, \mathbf{u})\right]$ being uniquely minimized at $h(\mathbb{E}\left[\mathbf{u}\right])$ for any η, is equivalent to: $\mathbb{E}\left[\Phi(h(\mathbf{s}), \mathbf{u})\right]$ being uniquely minimized at $\mathbb{E}\left[\mathbf{u}\right]$ for any η. Banerjee et al. [4] proved that this happens iff $\Phi(h(\mathbf{s}), \mathbf{u}) = D_\psi(\mathbf{u}, \mathbf{s})$ for some strictly convex ψ. ☐

Not every surrogate in the family identified above is convex. Below, we describe one large sub-family of *convex* NDCG consistent surrogates.

Theorem 9. *Let $\psi : \mathbb{R}^m \to \mathbb{R}$ be a strictly convex function whose gradient $\nabla\psi$ is order preserving. Then the surrogate defined as*

$$\phi(\mathbf{s}, \mathbf{r}) = D_\psi\left(\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}, (\nabla\psi)^{-1}(\mathbf{s})\right) \tag{3}$$

is convex (in $\mathbf{s}$) and NDCG consistent.

Proof. Since ψ is strictly convex, its gradient $\nabla\psi$ is invertible. By assumption, it is order preserving. Hence, by Theorem 8, ϕ is strongly NDCG consistent. To see that this surrogate is convex, simply rewrite it as

$$\phi(\mathbf{s}, \mathbf{r}) = D_{\psi^\star}\left(\mathbf{s}, \nabla\psi\left(\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right)\right) ,$$

where ψ^* is the Fenchel conjugate [14] of ψ. This is easily seen to be convex in $\mathbf{s}$ because any Bregman divergence is convex in its first argument. ☐

Note that it is easy to find ψ's whose gradients are order preserving maps. Any ψ of the form

$$\psi(\mathbf{r}) = \psi_{\text{out}}\left(\sum_{j=1}^{m} h(r_j)\right)$$

for a strictly convex function $h : \mathbb{R} \to \mathbb{R}$, and a strictly increasing function $\psi_{\text{out}} : \mathbb{R} \to \mathbb{R}$ has the property. Note that $\tilde{\phi}_{\text{sq}}$ and $\tilde{\phi}_{\text{list}}$ are in this family. They arise by choosing $\psi_{\text{out}}(x) = x, h(x) = x^2$ and $\psi_{\text{out}}(x) = x, h(x) = \exp(x)$ respectively.

Note that this family contains only convex surrogates with unique minimizers (of expected loss). As such, it does not include $\tilde{\phi}_{\cos}$ which is neither convex nor has unique minimizers. But $\tilde{\phi}_{\cos}$ is actually closely related to $\tilde{\phi}_{\text{sq}}$. Ignoring terms independent of $\mathbf{s}$, $\tilde{\phi}_{\text{sq}}$ can be written as:

$$\|\mathbf{s}\|_2^2 - 2\left\langle \mathbf{s}, \frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right\rangle$$

while $\tilde{\phi}_{\cos}$ can be written as

$$-\left\langle \frac{\mathbf{s}}{\|\mathbf{s}\|_2}, \frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right\rangle .$$

Thus, we see that *penalization* by $\|\mathbf{s}\|_2^2$ in $\tilde{\phi}_{\text{sq}}$ is replaced with *normalization* by $\|\mathbf{s}\|_2$ in $\tilde{\phi}_{\cos}$.

An interesting sub-family (that includes $\tilde{\phi}_{\text{sq}}$ but not $\tilde{\phi}_{\text{list}}$) is obtained by choosing $\psi(\mathbf{r}) = \|\mathbf{r}\|_p^2$ for $p > 1$. This corresponds to $\psi_{\text{out}}(x) = x^{2/p}, h(x) = |x|^p$. It is most interesting to focus on the range $p \in (1, 2]$, where ψ is strongly convex w.r.t. $\|\cdot\|_p$ and the excess risk transform of the next section applies. To the best of our knowledge, this subfamily has so far not been used as surrogates for NDCG. Thus, using this $\psi(\mathbf{r}) = \|\mathbf{r}\|_p^2$ for $p \in (1, 2]$ we obtain the family of loss functions:

$$\|\mathbf{s}\|_q^2 - 2\left\langle \mathbf{s}, \frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right\rangle,$$

where $q = p/(p-1)$ is the dual exponent of p and lies in the range $[2, \infty)$. Correspondingly, given this penalized form, we again can define the normalized version:

$$-\left\langle \frac{\mathbf{s}}{\|\mathbf{s}\|_q}, \frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right\rangle . \tag{4}$$

In the experiments, we compare this novel surrogate loss to Cosine and Cross Entropy loss functions, and show that it largely leads to improvement in NDCG performance across datasets.

5 Excess Risk Transforms

Recall that a function ψ is strongly convex w.r.t. a norm $\|\cdot\|$ if $D_\psi(\mathbf{s}, \mathbf{r}) \geq C_\phi \|\mathbf{s} - \mathbf{r}\|^2$ for some $C_\phi > 0$. The next result shows that under a strong convexity assumption, we can relate the excess error as measured in the surrogate to the excess NDCG error over the Bayes optimal error. Thus, it provides a quantified form of NDCG consistency.

Theorem 10. *Suppose we're using the surrogate ϕ as defined in* (2). *Further, assume that the function ψ is C_ϕ-strongly convex w.r.t a norm $\|\cdot\|$. Then, for any $\mathbf{f}$, for a constant C_F defined as $C_F = 2\left\|\left(\frac{1}{F(1)}, \ldots, \frac{1}{F(j)}, \ldots, \frac{1}{F(m)}\right)^\top\right\|_\star$, it holds that*

$$L_{\text{NDCG}}(\mathbf{f}) - L^\star_{\text{NDCG}} \leq \frac{C_F}{\sqrt{C_\phi}} \cdot \sqrt{\Phi(\mathbf{f}) - \Phi^\star} .$$

5.1 Better Transforms under "Low Noise" Conditions

We can improve the bound in Theorem 10 under a "low noise" condition that is reminiscent of similar assumptions for classification [18, 25]. We first define a notion of "margin" for the conditional distribution $\eta_{\mathbf{x}}(\mathbf{r})$ on $\mathcal{R}$ as follows. Let $\bar{\mathbf{r}} = \text{sort}(\mathbb{E}_{\mathbf{r}\sim\eta_{\mathbf{x}}}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right])$. We then define

$$\gamma_{\mathbf{x}} = \min_{l=1}^{m-1} (\bar{\mathbf{r}}_l - \bar{\mathbf{r}}_{l+1})\left(\frac{1}{F(l)} - \frac{1}{F(l+1)}\right). \tag{5}$$

It can be verified that for any $\mathbf{s} \in \mathcal{R}$,

$$\bar{\ell}_{\text{NDCG}}(\mathbf{s}; \eta_{\mathbf{x}}) - \bar{\ell}^\star_{\text{NDCG}}(\eta_{\mathbf{x}}) \geq \gamma_{\mathbf{x}},$$

so that $\gamma_{\mathbf{x}}$ is the minimum margin by which the conditional NDCG loss of any score vector would differ from that of the Fisher optimal score vector. Note that $\gamma_{\mathbf{x}} \leq 1$ by the definition of the DCG norm. But the closer it is to one, the larger the margin between the Fisher optimal score vector and any other score vector — which we would hope would entail that the ranking problem be easier. Let $\alpha \geq 0$ be such that

$$C_\gamma = \mathbb{E}\left[\left(\frac{1}{\gamma_{\mathbf{x}}}\right)^\alpha\right] < \infty. \tag{6}$$

Note that larger the margin $\gamma_{\mathbf{x}}$ (i.e. closer to one), the larger the value of α, so that the latter provides an alternative quantification of the size of the margin. We then define $\mathbf{v}_{\mathbf{x}} = \gamma_{\mathbf{x}}^{-\alpha}/\mathbf{C}_\gamma$. Note that by construction $\mathbb{E}[\mathbf{v}_{\mathbf{x}}] = 1$. The next theorem quantifies the advantage of a distribution with large margin.

Theorem 11. *Suppose we're using the surrogate ϕ as defined in* (2). *Assume that the function ψ is α strongly convex w.r.t a norm $\|\cdot\|$. Further, let C_γ and α be defined as in (5-6). Then, for any $\mathbf{f}$, and for a constant C_F defined as $C_F = 2\left\|\left(\frac{1}{F(1)}, \ldots, \frac{1}{F(j)}, \ldots, \frac{1}{F(m)}\right)^\top\right\|_\star$, it holds that*

$$L_{\text{NDCG}}(\mathbf{f}) - L^\star_{\text{NDCG}} \leq (\Phi(\mathbf{f}) - \Phi^\star)^{\frac{\alpha+1}{\alpha+2}} \cdot [C_F^2 C_\phi^{-1}]^{\frac{\alpha+1}{\alpha+2}} C_\gamma^{\frac{1}{\alpha+2}} .$$

Proof. Proceeding as in the proof of Theorem 10, we arrive at

$$\left(\bar{\ell}_{\text{NDCG}}(\mathbf{s};\eta) - \bar{\ell}^{\star}_{\text{NDCG}}(\eta)\right)^2 \leq C_F^2\, C_\phi^{-1} \cdot \bar{\phi}(\mathbf{s},\eta) - \bar{\phi}^{\star}(\eta)\ . \tag{7}$$

We then proceed along the lines of the proof of Theorem 13 in [25]. Since

$$\left(\bar{\ell}_{\text{NDCG}}(\mathbf{s};\eta) - \bar{\ell}^{\star}_{\text{NDCG}}(\eta)\right) \geq \gamma_{\mathbf{x}},$$

it can then be shown that

$$\begin{aligned}\left[\left(\bar{\ell}_{\text{NDCG}}(\mathbf{s};\eta) - \bar{\ell}^{\star}_{\text{NDCG}}(\eta)\right)^2 / \gamma_{\mathbf{x}}^{-\alpha}\right]^{\frac{\alpha+1}{\alpha+2}} \\ \geq \left(\bar{\ell}_{\text{NDCG}}(\mathbf{s};\eta) - \bar{\ell}^{\star}_{\text{NDCG}}(\eta)\right) / \gamma_{\mathbf{x}}^{-\alpha}.\end{aligned} \tag{8}$$

Recalling that $\mathbf{v}_{\mathbf{x}} = \gamma_{\mathbf{x}}^{-\alpha}/\mathbf{C}_\gamma$, and using the inequalities (7),(8) we get

$$\begin{aligned}C_F^2\, C_\phi^{-1} \cdot \frac{\bar{\phi}(\mathbf{s},\eta) - \bar{\phi}^{\star}(\eta)}{\mathbf{v}_{\mathbf{x}}} &\geq \frac{\left(\bar{\ell}_{\text{NDCG}}(\mathbf{s};\eta) - \bar{\ell}^{\star}_{\text{NDCG}}(\eta)\right)^2}{\mathbf{v}_{\mathbf{x}}} \\ &\geq \left[\frac{\bar{\ell}_{\text{NDCG}}(\mathbf{s};\eta) - \bar{\ell}^{\star}_{\text{NDCG}}(\eta)}{\mathbf{v}_{\mathbf{x}}}\right]^{\frac{\alpha+2}{\alpha+1}} C_\gamma^{-\frac{1}{\alpha+2}}.\end{aligned}$$

Taking expectations of both sides with respect to $\mu\, \mathbf{v}_{\mathbf{x}}$ and using Jensen's inequality completes the proof. □

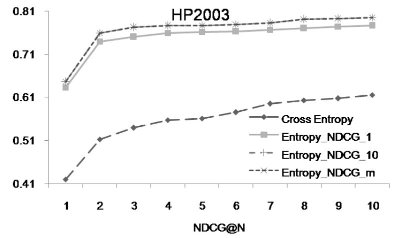
6 Experiments

This section reports two types of experiments. Firstly, we demonstrate the effectiveness of NDCG consistent variants of existing surrogates on the various datasets. Secondly, we give the performance of one novel loss function, as an example, from our proposed normalized family of loss functions (4).

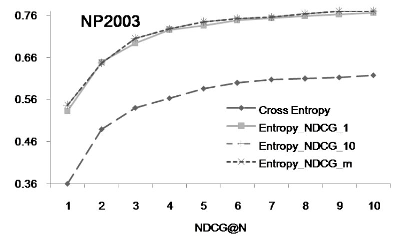
As our data, we used ten typical LETOR [17] datasets. These included LETOR 3.0, with three datasets from the 2003-2004 TREC Web track [11], as well as the older OHSUMED collection [13]. We also used LETOR 4.0 which is based on the 2007-2008 TREC Million Query tracks [2]. Finally, we also used Microsoft Learning to Rank Dataset [1] from the commercial web search engine with 10,000 queries (MS10K). We computed the NDCG metric with various truncated positions from 1 to 10, since these are the most popular metrics in Information Retrieval. Note that all the analysis above was based on the non-truncated version of NDCG. Therefore, we used $Z(\mathbf{r})_1$,$Z(\mathbf{r})_{10}$ as the approximations of DCG norm where $Z(\mathbf{r})_k = \max_\pi \sum_{j=1}^{k} \frac{G(r_{\pi^{-1}(j)})}{F(j)}$.

Figure 1 shows the improvements from the NDCG-consistency modifications in Section 3 of each surrogate over various datasets. For each surrogate as a baseline, three 'NDCG' labels indicate the modified versions with $Z(\mathbf{r})_1$,$Z(\mathbf{r})_{10}$ and $Z(\mathbf{r})$. As the best case, restoring NDCG consistency led to almost 30% improvement for cross-entropy surrogate on the HP2003 1(a). However, there were (small) performance degradations on some datasets; especially on the NP2004 for cross-entropy surrogate 1(i). We believe that this is because of the lack of the training and test queries: there are only 45 and 15 queries in the NP2004 dataset. Due to space limitations, we only present the cases where restoring each surrogate has pronounced effect. Plots for the rest of the cases are included in the appendix. Note that the original papers proposing listwise surrogates employed different loss functions and techniques for optimizing those loss functions. For example, ListNet [6] used gradient descent to minimize the cross-entropy loss, with the number of iterations and learning rate as parameters tuned on the validation set. On the other hand, RankCosine [19] minimized the cosine loss with the additive model. To evaluate across methods in a fair manner, we adopt the same optimization technique for all loss functions: gradient descent. In particular, we used the MATLAB implementation of gradient descent without any parameter tuning. And we did not include any other parameter for each surrogate and cross-validation for it. Finally, to avoid confusion, please note that in developing the theory we followed the convention used in Statistics of working with *losses* instead of *gains*. However, for reporting our results we adhere to reporting NDCG as a gain. Thus, *higher* NDCG values are *better*. We also ran random permutation significance tests for these comparisons, which we present in detail in Table 1 in the appendix. We note that the "large" changes are all one-sided: the only changes larger than 3% are all improvements; some of them as large as 30% as noted above.

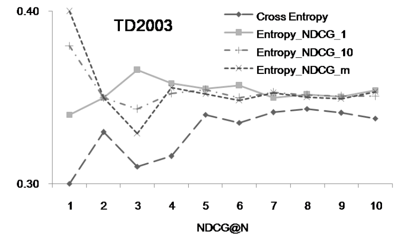
Secondly, we give the performance of one loss function, as an example, from the normalized family of loss functions (4). We chose $q = \log(m_i) + 2$ to make p close to 1. Figure 2 describes the NDCG@10 metrics of three surrogates: original Cosine, our new proposal with this choice of q, and original Cross Entropy, on the LETOR v3.0 and v4.0 datasets. Our surrogate loss function was much better than the cosine loss or even better than the state ofthe art cross-entropy loss on the 3 datasets, while comparable performance is seen on the other 6 datasets. Since Cross Entropy is strongly convex w.r.t. the ℓ_1-norm (for probability vectors), there is some hope that choosing $p \approx 1$ will make our proposed surrogate competitive with NDCG consistent version of Cross Entropy. Indeed the performance is comparable and sometime even better as Figure 6 in the Appendix demonstrates.



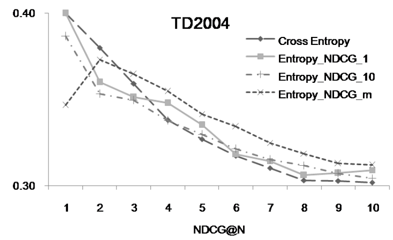
(a) Cross Entropy on the HP2003



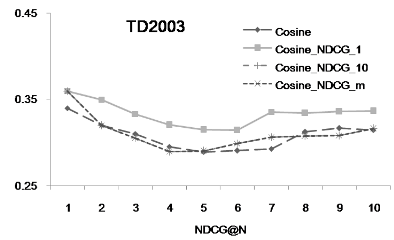
(b) Cross Entropy on the NP2003



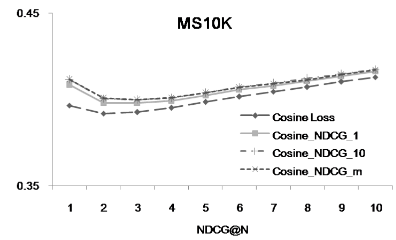
(c) Cross Entropy on the TD2003



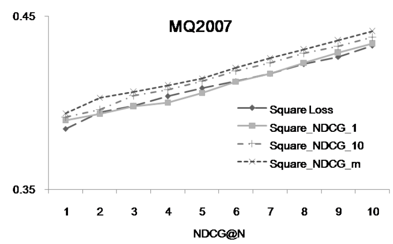
(d) Cross Entropy on the TD2004



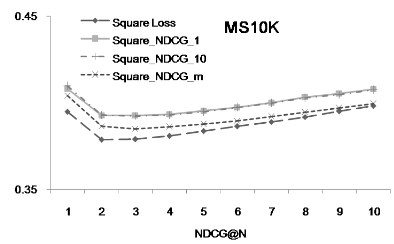
(e) Cosine on the TD2003



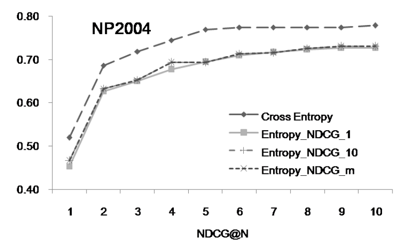
(f) Cosine on the MS10K



(g) Square on the MQ2007



(h) Square on the MS10K



(i) Cross Entropy on the NP2004

Figure 1: Selected results for NDCG@1-10: original surrogate vs. modifications to be NDCG consistent surrogate with different DCG norm approximations.

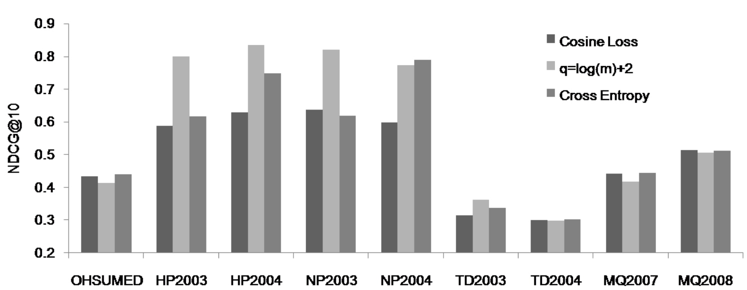


Figure 2: One example of normalized loss functions, $q = \log(m_i) + 2$ vs. existing listwise loss functions

References

[1] Microsoft learning to rank datasets. `http://research.microsoft.com/mslr`.

[2] J. Allan, J. Aslam, B. Carterette, V. Pavlu, and E. Kanoulas. Million query track 2008 overview. In *Proceedings of the 16th Text REtrieval Conference (TREC)*, 2009.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. Addison Wesley, 1999.

[4] A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Info. Theory*, 51(7):2664–2669, 2005.

[5] CJ Burges, QV Le, and R Ragno. Learning to Rank with Nonsmooth Cost Functions. In *Neural Information Processing Systems*, 2007.

[6] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *International Conference on Machine learning 24*, pages 129–136. ACM, 2007.

[7] Y. Censor and S. A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Numerical Mathematics and Scientific Computation. Oxford University Press, 1988.

[8] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Conference on Information and Knowledge Management (CIKM)*, 2009.

[9] S. Clemencon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In *Conference on Learning Theory (COLT)*, 2005.

[10] D. Cossock and T. Zhang. Statistical analysis of bayes optimal subset ranking. *IEEE Trans. Info. Theory*, 54:4140–5154, 2008.

[11] N. Craswell and D. Hawking. Overview of the TREC-2004 Web track. In *Proceedings of the 2004 Text REtreival Conference (TREC)*, 2005.

[12] J. Duchi, L. Mackey, and M. Jordan. On the consistency of ranking algorithms. In *International Conference on Machine Learning (ICML)*, 2010.

[13] W. Hersh, C. Buckley, TJ Leone, and D. Hickam. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 1994.

[14] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*, volume 1. Springer-Verlag, New York, 1993.

[15] Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, pages 41–48, New York, NY, USA, 2000. ACM.

[16] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 142, 2002.

[17] T.Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pages 3–10, 2007.

[18] Enno Mammen and Alexandre B. Tsybakov. Smoooth discrimination analysis. *Annals of Statistics*, 27(6):1808–1829, 1999.

[19] 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*, 2007.

[20] M Taylor, J Guiver, S Robertson, and T Minka. Softrank: Optimising Non-smooth Rank Metrics. In *International Conference on Web Search and Web Data Mining*, pages 77–86. ACM, 2008.

[21] H Valizadegan, R Jin, R Zhang, and J Mao. Learning to Rank by Optimizing NDCG Measure. In *Neural Information Processing Systems*, 2010.

[22] M.N. Volkovs and R.S. Zemel. Boltzrank: Learning to maximize expected ranking gain. In *International Conference on Machine learning 26*, pages 1089–1096, 2009.

[23] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: theory and algorithm. In *International Conference on Machine learning 25*, pages 1192–1199, 2008.

[24] 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*, page 398. ACM, 2007.

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

Supplementary Material

Proofs

Proof of Lemma 1

The norm in question is just

$$\mathbf{g} \mapsto \max_{\pi \in \mathcal{P}_m} \sum_{j=1}^{m} \frac{|g_i|}{F(\pi(j))} .$$

Each term inside the max is a weighted ℓ_1-norm. Pointwise maximum of any number of norms is also a norm.

Proof of Lemma 2

We note that the lemma would be easy to show if all entries of $\mathbf{r}$ were distinct. It was more delicate however to handle the general case where this need not hold.

The reverse implication is straightforward so we only prove the forward direction. Assume $\mathbf{s} \rightsquigarrow \mathbf{r}$. This means there is a permutation σ such that

$$s_{\sigma(1)} \geq s_{\sigma(2)} \geq \ldots \geq s_{\sigma(m)} ,$$
$$r_{\sigma(1)} \geq r_{\sigma(2)} \geq \ldots \geq r_{\sigma(m)} .$$

Now, define the map g as follows. Given $\mathbf{x}$, let τ be any permutation that sorts x in decreasing order, i.e.

$$x_{\tau(1)} \geq x_{\tau(2)} \geq \ldots \geq x_{\tau(m)} .$$

Define $g(x)$ to be the vector y defined as:

$$\begin{aligned} y_{\tau(1)} &= s_{\sigma(1)} + \tan^{-1}\left(x_{\tau(1)} - r_{\sigma(1)}\right) , \text{ and} \\ y_{\tau(k+1)} &= y_{\tau(k)} - \left[s_{\sigma(k)} - s_{\sigma(k+1)} \right. \\ &\quad \left. + \tan^{-1}\left(x_{\tau(k+1)} - r_{\sigma(k+1)}\right)\right] , \end{aligned}$$

for $k \geq 1$. Here, $\tan^{-1}(z)$, for $z \in \mathbb{R}$ is a non-negative function defined as the unique $\theta \in [0, \pi)$ such that $\tan(\theta) = z$. It is easy to check that g is order preserving and invertible, and that $g(\mathbf{r}) = \mathbf{s}$.

Proof of Lemma 3

There are two directions to prove: NDCG consistency $\leftrightarrow$ condition in Lemma 3. The forward direction is trivial: just take a marginal distribution μ that puts all mass on a single $\mathbf{x}$. For the other direction, assume condition in Lemma 3. Suppose we have a sequence $\mathbf{f}_n$ such that

$$\Phi(\mathbf{f}_n) \to \Phi^\star .$$

Then, it must be the case that, for μ-almost all $\mathbf{x}$,

$$\bar{\phi}(\mathbf{f}_n(\mathbf{x}), \eta_{\mathbf{x}}) \to \bar{\phi}^\star(\eta_{\mathbf{x}}) .$$

This implies that

$$\bar{\ell}_{\text{NDCG}}(\mathbf{f}_n(\mathbf{x}), \eta_{\mathbf{x}}) \to \bar{\ell}^\star_{\text{NDCG}}(\eta_{\mathbf{x}})$$

for μ-almost all $\mathbf{x}$. Thus we have

$$L_{\text{NDCG}}(\mathbf{f}_n) \to L^\star_{\text{NDCG}} .$$

which shows that ϕ is NDCG consistent.

Proof of Lemma 4

Assume that is not the case that

$$\mathbf{s} \rightsquigarrow \mathbb{E}\left[\mathbf{u}\right]$$

where

$$\mathbf{u} = \mathbb{E}_{\mathbf{r} \sim \eta}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right] .$$

Then, there exist i, j such that $u_i > u_j$ but $s_i \leq s_j$. Thus, there exists a permutation $\pi_{\mathbf{s}}$ that respects the sorted order of $\mathbf{s}$ and which ranks j higher than i, i.e. $\pi_{\mathbf{s}}(j) < \pi_{\mathbf{s}}(i)$. This means that

$$\frac{1}{F(\pi_{\mathbf{s}}(j))} - \frac{1}{F(\pi_{\mathbf{s}}(i))} > 0 .$$

Multiplying by $u_i - u_j > 0$ gives

$$-\left(\frac{u_i}{F(\pi_{\mathbf{s}}(i))} + \frac{u_j}{F(\pi_{\mathbf{s}}(j))}\right) > -\left(\frac{u_j}{F(\pi_{\mathbf{s}}(i))} + \frac{u_i}{F(\pi_{\mathbf{s}}(j))}\right) .$$

That is, we can decrease the NDCG loss by swapping i with j. Thus $\bar{\ell}_{\text{NDCG}}(\mathbf{s}; \eta) < \bar{\ell}^\star_{\text{NDCG}}(\eta)$.

Now assume that $\mathbf{s} \rightsquigarrow \mathbb{E}\left[\mathbf{u}\right]$ with $\mathbf{u}$ as defined above. Using the same argument we can show that the NDCG loss does not decrease no matter which i, j we swap. Hence $\bar{\ell}_{\text{NDCG}}(\mathbf{s}; \eta) = \bar{\ell}^\star_{\text{NDCG}}(\eta)$.

Proof of Theorem 6

Again there are two directions to prove: condition of Lemma 3 $\leftrightarrow$ condition of Theorem 6. Let us prove the forward direction first. By definition of $\mathbf{s}^\star_\phi(\eta)$, we have

$$\bar{\phi}(\mathbf{s}^\star_\phi(\eta); \eta) = \bar{\phi}^\star(\eta)$$

and hence, under the condition of Lemma 3,

$$\bar{\ell}_{\text{NDCG}}(\mathbf{s}^\star_\phi(\eta); \eta) = \bar{\ell}^\star_{\text{NDCG}}(\eta) ,$$

which implies

$$\mathbf{s}^\star_\phi(\eta) \rightsquigarrow \mathbb{E}_{\mathbf{r} \sim \eta}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right] .$$

Now, by Lemma 2, there is an invertible order preserving g such that

$$\mathbf{s}^\star_\phi(\eta) = g\left(\mathbb{E}_{\mathbf{r} \sim \eta}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right]\right) .$$

For the reverse direction, assume condition of Theorem 6 and that

$$\bar{\phi}(\mathbf{s}_n;\eta) \to \bar{\phi}^\star(\eta) \tag{9}$$

for some sequence $\mathbf{s}_n$. We want to show that

$$\bar{\ell}_{\mathrm{NDCG}}(\mathbf{s}_n,\eta) = \bar{\ell}^\star_{\mathrm{NDCG}}(\eta) \tag{10}$$

for n large enough. By our regularity assumption on ϕ, (9) implies that $\mathbf{s}_n \to \mathbf{s}^\star_\phi(\eta)$. By the condition of Theorem 6, we have

$$\mathbf{s}^\star_\phi(\eta) \rightsquigarrow \mathbb{E}_{\mathbf{r}\sim\eta}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right] .$$

Again abbreviate the vector on the right to $\mathbf{u}$. We want to claim that for each fixed pair i,j such that $u_i > u_j$, we have $s_{n,i} > s_{n,j}$ for n large enough. But this follows from

$$[\mathbf{s}^\star_\phi(\eta)]_i > [\mathbf{s}^\star_\phi(\eta)]_j$$

and the fact that $\mathbf{s}_n \to \mathbf{s}^\star_\phi(\eta)$. Since there are only finitely many pairs i,j, we can now claim that

$$\mathbf{s}_n \rightsquigarrow \mathbf{u} = \mathbb{E}_{\mathbf{r}\sim\eta}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right]$$

for n large enough. Thus, by Lemma 4, we know that (10) is true for n large enough. This proves the reverse direction and finishes the proof.

Proof of Proposition 7

To show NDCG inconsistency of a surrogate ϕ, it is enough to exhibit one distribution η, where the sorted order of the minimizer of $\bar{\phi}(\mathbf{s};\eta)$ is different from the sorted order of $\mathbb{E}\left[G(\mathbf{r})/\|G(\mathbf{r})\|_D\right]$.

We have already done that for $\phi = \phi_{\mathrm{sq}}$ in Section 3.2.1. For both $\phi_{\cos}$ and ϕ_{list}, the distribution exhibiting inconsistency will be supported on two vectors

$$\begin{pmatrix}1\\x\end{pmatrix} \qquad \begin{pmatrix}y\\1\end{pmatrix}$$

with probabilities p and $1-p$ respectively. One can simply verify that we get NDCG inconsistency if we choose $p = 0.38, x = 5, y = 2$ (for Cosine) or $p = 0.35, x = 5, y = 2$ (for Cross Entropy). The geometric picture behind what is causing inconsistency for these distributions is given in Figures 3 and 4.

Proof of Theorem 10

We will show that for any $\mathbf{s}$ and any distribution η over $\mathcal{R}$, we have

$$\bar{\ell}_{\mathrm{NDCG}}(\mathbf{s};\eta) - \bar{\ell}^\star_{\mathrm{NDCG}}(\eta) \le \frac{C_F}{\sqrt{C_\phi}} \cdot \sqrt{\bar{\phi}(\mathbf{s};\eta) - \bar{\phi}^\star(\eta)}$$

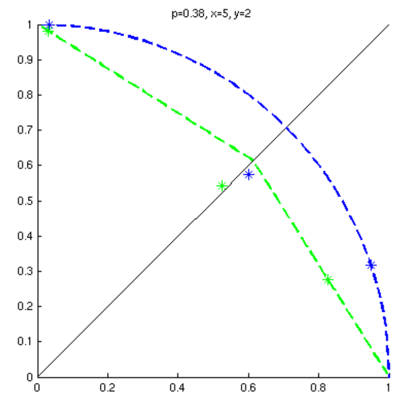


Figure 3: **Inconsistency of Cosine** The distribution is supported on $u = (1,x)$ and $v = (y,1)$ with probability p and $1-p$ respectively. The 3 green points are $G(u)/\|G(u)\|_D$, $G(v)/\|G(v)\|_D$ and their weighted mean. The 3 blue points are $G(u)/\|G(u)\|_2$, $G(v)/\|G(v)\|_2$ and their weighted mean. Note that the weighted means lie on different sides of the black diagonal line.

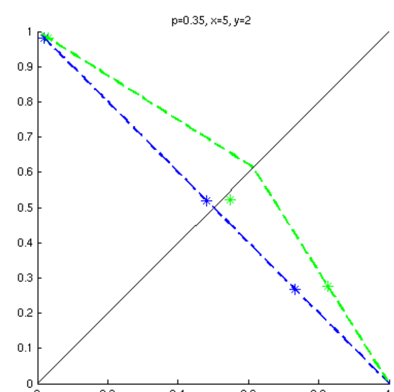


Figure 4: **Inconsistency of Cross Entropy** The distribution is supported on $u = (1,x)$ and $v = (y,1)$ with probability p and $1-p$ respectively. The 3 green points are $G(u)/\|G(u)\|_D$, $G(v)/\|G(v)\|_D$ and their weighted mean. The 3 blue points are $\exp(u)/\mathbf{1}^\top\exp(u)$, $\exp(v)/\mathbf{1}^\top\exp(v)$ and their weighted mean. Note that the weighted means lie on different sides of the black diagonal line.

from which the result follows after taking expectations and using Jensen's inequality.

To keep notation simple, we will omit subscripts under expectations. All expectations are w.r.t. $\mathbf{r}$ drawn from η. Let π be an arbitrary permutation. We have,

$$\bar{\ell}_{\mathrm{NDCG}}(\mathbf{s};\eta) = \mathbb{E}\left[-\frac{1}{\|G(\mathbf{r})\|_D}\sum_{j=1}^{m}\frac{G(r_j)}{F(\pi_{\mathbf{s}}(j))}\right]$$

$$= \mathbb{E}\left[-\frac{1}{\|G(\mathbf{r})\|_D}\sum_{j=1}^{m}\frac{G(r_{\pi_{\mathbf{s}}^{-1}(j)})}{F(j)}\right] \tag{11}$$

$$= \mathbb{E}\left[-\sum_{j=1}^{m}\frac{(g(\mathbf{s}))_{\pi_{\mathbf{s}}^{-1}(j)}}{F(j)}\right] + T_1$$

$$\leq \mathbb{E}\left[-\sum_{j=1}^{m}\frac{(g(\mathbf{s}))_{\pi^{-1}(j)}}{F(j)}\right] + T_1$$

$$= \mathbb{E}\left[-\frac{1}{\|G(\mathbf{r})\|_D}\sum_{j=1}^{m}\frac{G(r_{\pi^{-1}(j)})}{F(j)}\right] + T_2 + T_1$$

$$= \mathbb{E}\left[-\frac{1}{\|G(\mathbf{r})\|_D}\sum_{j=1}^{m}\frac{G(r_j)}{F(\pi(j))}\right] + T_2 + T_1$$

$$= \bar{\ell}_{\mathrm{NDCG}}(\pi;\eta) + T_2 + T_1\ . \tag{12}$$

where

$$T_1 := \mathbb{E}\left[\sum_{j=1}^{m}\frac{1}{F(j)}\cdot\left((g(\mathbf{s}))_{\pi_{\mathbf{s}}^{-1}(j)} - \frac{G(r_{\pi_{\mathbf{s}}^{-1}(j)})}{\|G(\mathbf{r})\|_D}\right)\right]\ ,$$

$$T_2 := \mathbb{E}\left[\sum_{j=1}^{m}\frac{1}{F(j)}\cdot\left(\frac{G(r_{\pi^{-1}(j)})}{\|G(\mathbf{r})\|_D} - (g(\mathbf{s}))_{\pi^{-1}(j)}\right)\right]\ .$$

The inequality above holds because the sorted order of $\mathbf{s}$ and $g(\mathbf{s})$ match (i.e. $\mathbf{s} \rightsquigarrow g(\mathbf{s})$) since g is an order-preserving map. Note that both T_1 and T_2 are bounded by

$$\frac{C_F}{2}\cdot\left\|g(\mathbf{s}) - \mathbb{E}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right]\right\|$$

using the inequality $\langle \mathbf{u},\mathbf{v}\rangle \leq \|\mathbf{u}\|\cdot\|\mathbf{v}\|_\star$ and definition of C_F. Plugging this into (12), we get

$$\bar{\ell}_{\mathrm{NDCG}}(\mathbf{s};\eta) - \bar{\ell}_{\mathrm{NDCG}}(\pi;\eta) \leq C_F\left\|g(\mathbf{s}) - \mathbb{E}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right]\right\|$$
$$\leq \frac{C_F}{\sqrt{C_\phi}}\cdot\sqrt{\bar{\phi}(\mathbf{s},\eta) - \bar{\phi}^\star(\eta)}\ .$$

The last inequality above follows because by C_ϕ-strong convexity of ψ w.r.t. $\|\cdot\|$, we have

$$\bar{\phi}(\mathbf{s},\eta) = \mathbb{E}\left[D_\psi\left(\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}, g(\mathbf{s})\right)\right]$$
$$= \min_{\mathbf{s}'}\mathbb{E}\left[D_\psi\left(\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}, g(\mathbf{s}')\right)\right]$$
$$+ D_\psi\left(\mathbb{E}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right], g(\mathbf{s})\right)$$
$$\geq \bar{\phi}^\star(\eta) + C_\phi\left\|g(\mathbf{s}) - \mathbb{E}\left[\frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}\right]\right\|^2\ .$$

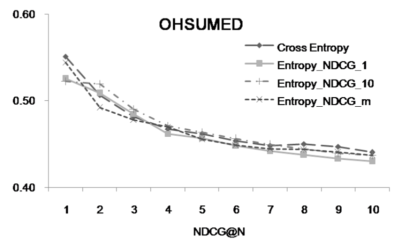
Taking maximum over π yields,

$$\bar{\ell}_{\mathrm{NDCG}}(\mathbf{s};\eta) - \bar{\ell}^\star_{\mathrm{NDCG}}(\eta) \leq \frac{C_F}{\sqrt{C_\phi}}\cdot\sqrt{\bar{\phi}(\mathbf{s},\eta) - \bar{\phi}^\star(\eta)}\ .$$

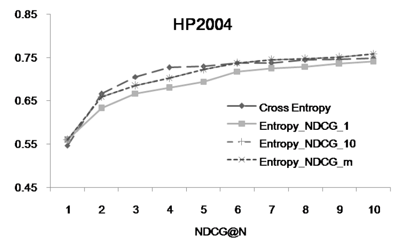
and this completes the proof.

Plots

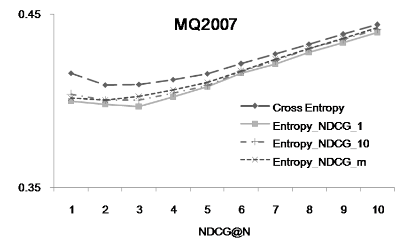
In Figure 6, we present the rest of the plots comparing the NDCG consistent and unmodified versions of existing surrogates, and where the differences were not that pronounced. We also ran significance tests for these comparisons; presented in Figure 1. We modified the Lemur toolkit to compute NDCG@10 and used the random permutation test with 5% significance level for each test. We were able to test 9 out of 10 datasets in the paper; we were not able to run the Lemur toolkit for the MS10K dataset due to memory limits. Out of 81 evaluation points (9 datasets x 3 loss functions x 3 metrics (NDCG@1,5,10)), NDCG recovery performed significantly better in 11 and worse in 9 cases. One interesting thing here was that 5 cases out of 9 "worse" cases came from only one dataset (MQ2008). Further, the "large" changes were all one-sided: the only changes larger than 3% were all improvements; some of them as large as 30%.



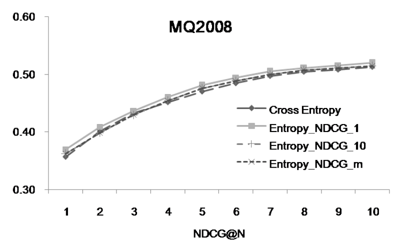
(a) Cross Entropy on the OHSUMED



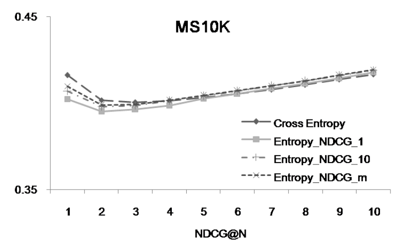
(b) Cross Entropy on the HP2004



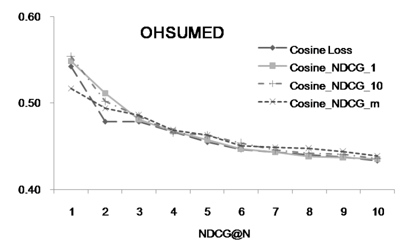
(c) Cross Entropy on the MQ2007



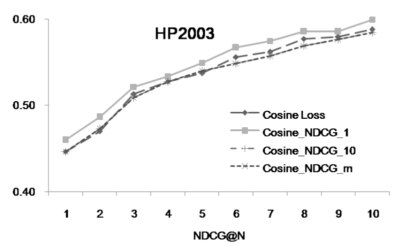
(d) Cross Entropy on the MQ2008



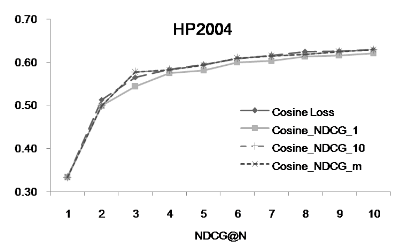
(e) Cross Entropy on the MS10K



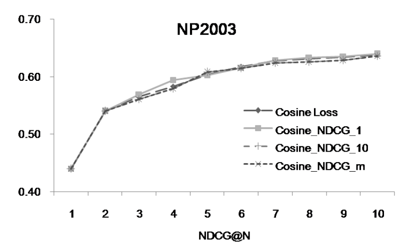
(f) Cosine on the OHSUMED



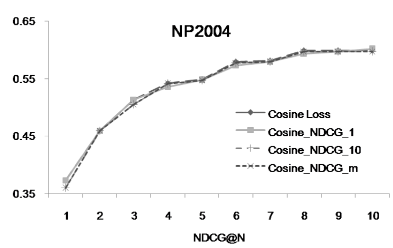
(g) Cosine on the HP2003



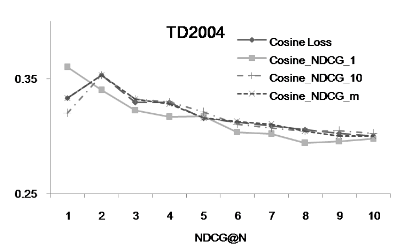
(h) Cosine on the HP2004



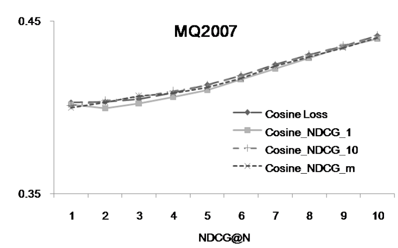
(i) Cosine on the NP2003



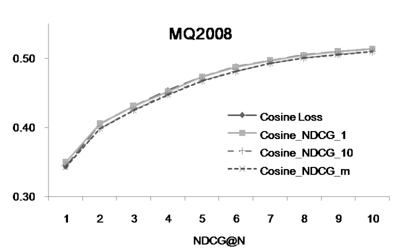
(j) Cosine on the NP2004



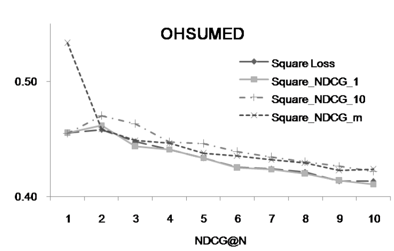
(k) Cosine on the TD2004



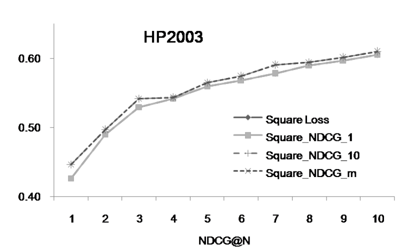
(l) Cosine on the MQ2007



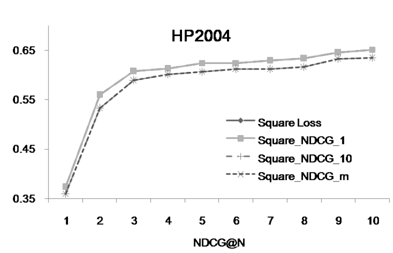
(m) Cosine on the MQ2008



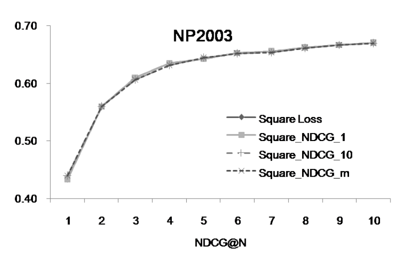
(n) Square on the OHSUMED



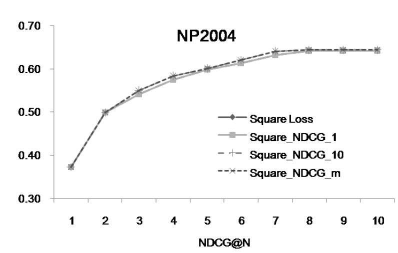
(o) Square on the HP2003



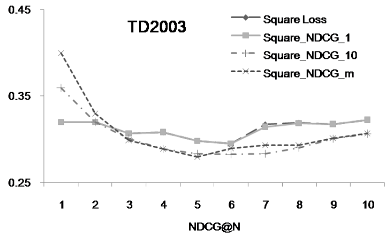
(p) Square on the HP2004



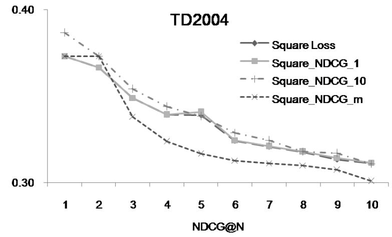
(q) Square on the NP2003



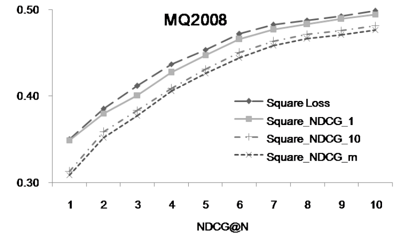
(r) Square on the NP2004



(s) Square on the TD2003



(t) Square on the TD2004



(u) Square on the MQ2008

Figure 5: NDCG@1-10 for original surrogate vs. NDCG consistent surrogate

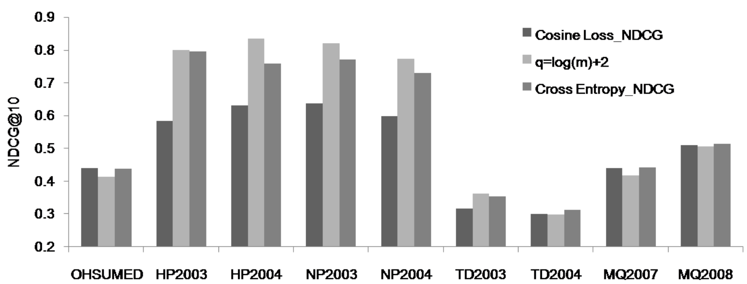


Figure 6: One example of normalized loss functions, $q = \log(m_i) + 2$ vs. existing listwise loss functions w/ recovering NDCG consistency

Table 1: Comparison of the 'NDCG-consistent' version (with $Z(r)_{10}$) to the baseline across 81 evaluation points: 9 datasets, 3 loss functions (cross-entropy, cosine and squared), and 3 metrics (NDCG@{1,5,10}). For each case, we report whether our method performed better, same, or worse than the baseline (with statistical significance). We also report average change in relative accuracy across the 9 evaluation points for each dataset.

Dataset	OHSUMED	HP2003	HP2004	NP2003	NP2004	TD2003	TD2004	MQ2007	MQ2008	Total	%
Better	2	5	0	3	0	0	0	1	0	11	13.6%
Same	7	4	9	6	8	8	9	6	4	61	75.3%
Worse	0	0	0	0	1	1	0	2	5	9	11.1%
Total	9	9	9	9	9	9	9	9	9	63	100%
Avg. change	0.81%	14.9%	-0.78%	11.68%	-2.77%	4.73%	1.82%	-0.09%	-2.02%	4.34%	