Active evaluation of ranking functions based on graded relevance

Christoph Sawade · Steffen Bickel · Timo von Oertzen ·
Tobias Scheffer · Niels Landwehr

Received: 17 November 2012 / Accepted: 26 April 2013 / Published online: 22 May 2013
© The Author(s) 2013

Abstract Evaluating the quality of ranking functions is a core task in web search and other information retrieval domains. Because query distributions and item relevance change over time, ranking models often cannot be evaluated accurately on held-out training data. Instead, considerable effort is spent on manually labeling the relevance of query results for test queries in order to track ranking performance. We address the problem of estimating ranking performance as accurately as possible on a fixed labeling budget. Estimates are based on a set of most informative test queries selected by an active sampling distribution. Query labeling costs depend on the number of result items as well as item-specific attributes such as document length. We derive cost-optimal sampling distributions for the commonly used performance measures Discounted Cumulative Gain and Expected Reciprocal Rank. Experiments on web search engine data illustrate significant reductions in labeling costs.

Keywords Information retrieval · Ranking · Active evaluation

Editors: Tijl De Bie and Peter Flach.

C. Sawade (✉) · T. Scheffer · N. Landwehr
Department of Computer Science, University of Potsdam, August-Bebel-Strasse 89, 14482 Potsdam,
Germany
e-mail: sawade@cs.uni-potsdam.de

T. Scheffer
e-mail: scheffer@cs.uni-potsdam.de

N. Landwehr
e-mail: landwehr@cs.uni-potsdam.de

S. Bickel
Nokia gate5 GmbH, Invalidenstrasse 117, 10115 Berlin, Germany
e-mail: steffen.bickel@nokia.com

T. von Oertzen
Department of Psychology, University of Virginia, Charlottesville, VA 22903, USA
e-mail: timo@virginia.edu

1 Introduction

This paper addresses the problem of estimating the performance of a given ranking function in terms of graded relevance measures such as Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen 2002) and Expected Reciprocal Rank (ERR) (Chapelle et al. 2009). In informational retrieval domains, ranking models often cannot be evaluated on held-out training data. For example, older training data might not represent the distribution of queries the model is currently exposed to, or the ranking model might be procured from a third party that does not provide any training data.

In practice, ranking performance is estimated by applying a given ranking model to a representative set of test queries and manually assessing the relevance of all retrieved items for each query. We study the problem of estimating ranking performance as accurately as possible on a fixed budget for labeling item relevance, or, equivalently, minimizing labeling costs for a given level of estimation accuracy. We also study the related problem of cost-efficiently comparing the ranking performance of two models; this is required, for instance, to evaluate the result of an index update.

We assume that drawing unlabeled data $x \sim p(x)$ from the distribution of queries that the model is exposed to is inexpensive, whereas obtaining relevance labels is costly. The standard approach to estimating ranking performance is to draw a sample of test queries from $p(x)$, obtain relevance labels, and compute the empirical performance. However, recent results on *active risk estimation* (Sawade et al. 2010) and *active comparison* (Sawade et al. 2012) indicate that estimation accuracy can be improved by drawing test examples from an appropriately engineered instrumental distribution $q(x)$ rather than $p(x)$, and correcting for the discrepancy between p and q by importance weighting. Motivated by these results, we study active estimation processes for ranking performance. In analogy to active learning, they select specific queries from a pool of unlabeled test queries, obtain relevance labels for all items returned for these queries by the ranking function to be evaluated, and then compute appropriate performance estimates based on the observed item relevance. Test queries are selected according to an instrumental distribution $q(x)$. The actively selected sample is weighted appropriately to compensate for the discrepancy between instrumental and test distributions which leads to a consistent—that is, asymptotically unbiased—performance estimate.

We derive instrumental distributions that asymptotically minimize the estimation error for either evaluating a single ranking function or comparing two given ranking functions. In ranking problems, the labeling costs for a query depend on the number of items returned by the ranking function as well as item-specific attributes such as document length; this is in contrast to the active estimation settings discussed previously (Sawade et al. 2010, 2012). We show how these costs can be taken into account when deriving optimal instrumental distributions. Moreover, in the ranking setting a naïve computation of optimal instrumental distributions is exponential; a central contribution of this paper is to show how these distributions can be computed in polynomial time using dynamic programming.

An initial version of this paper appeared at the 2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (Sawade et al. 2012). The current paper extends the earlier conference version by more detailed proofs for central theorems, additional empirical results, and a more self-contained presentation. We now briefly summarize the extensions made with respects to the earlier conference version.

A central theoretical contribution of the paper is the method for efficient computation of optimal sampling distributions (Theorem 3). There are four different instantiations of this result: estimation of the absolute performance of a single ranking function, and comparison of the relative performance of ranking functions, in combination with the performance

measures discounted cumulative gain (DCG) and expected reciprocal rank (ERR). The conference version of the paper only contained the proof for one of the four cases (absolute estimation, ERR). The current paper contains detailed proofs for all four cases.

While the conference paper derived active performance estimation techniques for the ranking performance measures ERR and DCG, only results for the ERR measure were included in the empirical study. In the current paper, we also include comprehensive empirical results for the DCG measure. In the empirical study, we have also added a visualization of the sampling distribution employed by our method as a function of sample cost and intrinsic expectation of ranking performance.

The rest of the paper is organized as follows. Section 2 details the problem setting. Section 3 derives cost-optimal sampling distributions for the estimation of DCG and ERR. Section 4 discusses empirical sampling distributions in a pool-based setting. Section 5 presents empirical results. Section 6 discusses related work, Sect. 7 concludes.

2 Problem setting

Let $\mathcal{X}$ denote a space of queries, and $\mathcal{Z}$ denote a finite space of items. We study ranking functions

$$\begin{aligned} \mathbf{r} : \mathcal{X} &\rightarrow \bigcup_{k=1}^{\infty} \mathcal{Z}^k, \\ \mathbf{r} : x &\mapsto (r_1(x), \dots, r_{|\mathbf{r}(x)|}(x))^T \end{aligned}$$

that, given a query $x \in \mathcal{X}$, return a list of $|\mathbf{r}(x)|$ items $r_i(x) \in \mathcal{Z}$ ordered by relevance. The number of items in a ranking $\mathbf{r}(x)$ can vary depending on the query and application domain from thousands (web search) to ten or fewer (mobile applications that have to present results on a small screen). Ranking performance of $\mathbf{r}$ is defined in terms of graded relevance labels $y_z \in \mathcal{Y}$ that represent the relevance of an item $z \in \mathcal{Z}$ for the query x , where $\mathcal{Y} \subset \mathbb{R}$ is a finite space of relevance labels with minimum zero (irrelevant) and maximum $y_{\max}$ (perfectly relevant). We summarize the graded relevance of all $z \in \mathcal{Z}$ in a label vector $\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}$ with components y_z for $z \in \mathcal{Z}$.

In order to evaluate the quality of a ranking $\mathbf{r}(x)$ for a single query x , we employ two commonly used ranking performance measures: *Discounted Cumulative Gain* (DCG), given by

$$\begin{aligned} L_{dcg}(\mathbf{r}(x), \mathbf{y}) &= \sum_{i=1}^{|\mathbf{r}(x)|} \ell_{dcg}(y_{r_i(x)}, i) \\ \ell_{dcg}(y, i) &= \frac{2^y - 1}{\log_2(i + 1)}, \end{aligned} \quad (1)$$

and *Expected Reciprocal Rank* (ERR), given by

$$\begin{aligned} L_{err}(\mathbf{r}(x), \mathbf{y}) &= \sum_{i=1}^{|\mathbf{r}(x)|} \frac{1}{i} \ell_{err}(y_{r_i(x)}) \prod_{l=1}^{i-1} (1 - \ell_{err}(y_{r_l(x)})) \\ \ell_{err}(y) &= \frac{2^y - 1}{2^{y_{\max}}} \end{aligned} \quad (2)$$

as introduced by Järvelin and Kekäläinen (2002) and Chapelle et al. (2009), respectively.

DCG scores a ranking by summing over the relevance of all documents discounted by their position in the ranking. ERR is based on a probabilistic user model: the user scans a list of documents in the order defined by $\mathbf{r}(x)$ and chooses the first document that appears sufficiently relevant; the likelihood of choosing a document z is a function of its graded relevance score y_z . If s denotes the position of the chosen document in $\mathbf{r}(x)$, then $L_{err}(\mathbf{r}(x), \mathbf{y})$ is the expectation of the reciprocal rank $1/s$ under the probabilistic user model. Both DCG and ERR discount relevance with ranking position, ranking quality is thus most strongly influenced by documents that are ranked highly. If $\mathbf{r}(x)$ includes many items, L_{dcg} and L_{err} are in practice often approximated by only labeling items up to a certain position in the ranking or a certain relevance threshold and ignoring the contribution of lower-ranked items.

Let $p(x, \mathbf{y}) = p(x)p(\mathbf{y}|x)$ denote the joint distribution over queries $x \in \mathcal{X}$ and label vectors $\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}$ the model is exposed to. We assume that the individual relevance labels y_z for items z are drawn independently given a query x :

$$p(\mathbf{y}|x) = \prod_{z \in \mathcal{Z}} p(y_z|x, z). \quad (3)$$

This assumption is common in pointwise ranking approaches, e.g., regression based ranking models (Cossock and Zhang 2008; Mohan et al. 2011). The ranking performance of $\mathbf{r}$ with respect to $p(x, \mathbf{y})$ is given by

$$R[\mathbf{r}] = \int \int L(\mathbf{r}(x), \mathbf{y}) p(x, \mathbf{y}) dx d\mathbf{y}, \quad (4)$$

where $L \in \{L_{dcg}, L_{err}\}$ denotes the performance measure under study. We use integrals for notational convenience, for discrete spaces the corresponding integral is replaced by a sum. If the context is clear, we refer to $R[\mathbf{r}]$ simply by R .

Since $p(x, \mathbf{y})$ is unknown, ranking performance is typically approximated by an empirical average

$$\hat{R}_n[\mathbf{r}] = \frac{1}{n} \sum_{j=1}^n L(\mathbf{r}(x_j), \mathbf{y}_j), \quad (5)$$

where a set of test queries $x_1, \dots, x_n$ and graded relevance labels $\mathbf{y}_1, \dots, \mathbf{y}_n$ are drawn *iid* from $p(x, \mathbf{y})$. The empirical performance $\hat{R}_n$ consistently estimates the true ranking performance; that is, $\hat{R}_n$ converges to R with $n \rightarrow \infty$.

Test queries x_i need not necessarily be drawn according to the input distribution p . When instances are drawn according to an instrumental distribution q , an estimator can be defined as

$$\hat{R}_{n,q}[\mathbf{r}] = \left(\sum_{j=1}^n \frac{p(x_j)}{q(x_j)} \right)^{-1} \sum_{j=1}^n \frac{p(x_j)}{q(x_j)} L(\mathbf{r}(x_j), \mathbf{y}_j), \quad (6)$$

where $(x_j, \mathbf{y}_j)$ are drawn from $q(x)p(\mathbf{y}|x)$ and again $L \in \{L_{dcg}, L_{err}\}$. Weighting factors $\frac{p(x_j)}{q(x_j)}$ compensate for the discrepancy between test and instrumental distributions, and the normalizer is the sum of weights. Because of the weighting factors, Eq. (6) defines a consistent estimator (see, e.g. Liu 2001, p. 35). Note that Eq. (5) is a special case of Eq. (6), using the instrumental distribution $q = p$.

For certain choices of the sampling distribution q , $\hat{R}_{n,q}$ may be a more label-efficient estimator of the true performance R than $\hat{R}_n$ (Sawade et al. 2010).

A crucial feature of ranking domains is that labeling costs for queries $x \in \mathcal{X}$ vary with the number of items $|\mathbf{r}(x)|$ returned and item-specific features such as the length of a document

whose relevance has to be determined. We denote labeling costs for a query x by $\lambda(x)$, and assume that $\lambda(x)$ is bounded away from zero by $\lambda(x) \geq \epsilon > 0$. Our goal is to minimize the deviation of $\hat{R}_{n,q}$ from R under the constraint that expected overall labeling costs stay below a budget $\Lambda \in \mathbb{R}$:

$$(q^*, n^*) = \arg \min_{q,n} \mathbb{E}[(\hat{R}_{n,q} - R)^2], \quad \text{s.t. } \mathbb{E}\left[\sum_{j=1}^n \lambda(x_j)\right] \leq \Lambda. \quad (7)$$

Note that Eq. (7) represents a trade-off between labeling costs and informativeness of a test query: optimization over n implies that many inexpensive or few expensive queries could be chosen.

To estimate *relative* performance of two ranking functions $\mathbf{r}_1$ and $\mathbf{r}_2$, Eq. (7) can be replaced by

$$(q^*, n^*) = \arg \min_{q,n} \mathbb{E}[(\hat{\Delta}_{n,q} - \Delta)^2], \quad \text{s.t. } \mathbb{E}\left[\sum_{j=1}^n \lambda(x_j)\right] \leq \Lambda, \quad (8)$$

where $\hat{\Delta}_{n,q} = \hat{R}_{n,q}[\mathbf{r}_1] - \hat{R}_{n,q}[\mathbf{r}_2]$ and $\Delta = R[\mathbf{r}_1] - R[\mathbf{r}_2]$. In the next section, we derive sampling distributions q^* asymptotically solving Eqs. (7) and (8).

3 Asymptotically optimal sampling

A bias-variance decomposition (Geman et al. 1992) applied to Eq. (7) results in

$$\mathbb{E}[(\hat{R}_{n,q} - R)^2] = \mathbb{E}[\hat{R}_{n,q}^2 - 2R\hat{R}_{n,q} + R^2] + \mathbb{E}[\hat{R}_{n,q}]^2 - \mathbb{E}[\hat{R}_{n,q}]^2 \quad (9)$$

$$= (\mathbb{E}[\hat{R}_{n,q}] - R)^2 + (\mathbb{E}[\hat{R}_{n,q}^2] - \mathbb{E}[\hat{R}_{n,q}]^2) \quad (10)$$

$$= \text{Bias}[\hat{R}_{n,q}]^2 + \text{Var}[\hat{R}_{n,q}]. \quad (11)$$

In Eq. (9), we expand the square, and add and subtract the expected value of $\hat{R}_{n,q}$. Reordering terms and factorizing yield Eq. (10). Equation (11) expresses the expected squared deviation of the estimation as a sum of a squared bias and a variance term.

Because $\hat{R}_{n,q}$ is consistent, $\mathbb{E}[(\hat{R}_{n,q} - R)^2]$ vanishes for $n \rightarrow \infty$. According to Liu (2001), Chap. 2.5.3, the squared bias term is of order $\frac{1}{n^2}$, while the variance is of order $\frac{1}{n}$. For large n , the expected deviation is thus dominated by the variance, and $\sigma_q^2 = \lim_{n \rightarrow \infty} n \text{Var}[\hat{R}_{n,q}]$ exists. For large n , we can thus approximate

$$\mathbb{E}[(\hat{R}_{n,q} - R)^2] \approx \frac{1}{n} \sigma_q^2, \quad \mathbb{E}[(\hat{\Delta}_{n,q} - \Delta)^2] \approx \frac{1}{n} \tau_q^2,$$

where $\tau_q^2 = \lim_{n \rightarrow \infty} n \text{Var}[\hat{\Delta}_{n,q}]$.

Let $\delta(x, \mathbf{y}) = L(\mathbf{r}_1(x), \mathbf{y}) - L(\mathbf{r}_2(x), \mathbf{y})$ denote the performance difference of the two ranking models for a test query $(x, \mathbf{y})$ and $L \in \{L_{dcg}, L_{err}\}$. The following theorems derive sampling distributions minimizing the quantities $\frac{1}{n} \sigma_q^2$ and $\frac{1}{n} \tau_q^2$, thereby approximately solving Problems (7) and (8).

Theorem 1 (Optimal sampling for evaluation of a ranking function) *Let $L \in \{L_{dcg}, L_{err}\}$ and $\sigma_q^2 = \lim_{n \rightarrow \infty} n \text{Var}[\hat{R}_{n,q}]$. The optimization problem*

$$(q^*, n^*) = \arg \min_{q,n} \frac{1}{n} \sigma_q \quad \text{s.t. } \mathbb{E}\left[\sum_{j=1}^n \lambda(x_j)\right] \leq \Lambda$$

is solved by

$$q^*(x) \propto \frac{p(x)}{\sqrt{\lambda(x)}} \sqrt{\int (L(\mathbf{r}(x), \mathbf{y}) - R)^2 p(\mathbf{y}|x) d\mathbf{y}}, \quad n^* = \frac{\Lambda}{\int \lambda(x) q(x) dx}. \quad (12)$$

Theorem 2 (Optimal sampling for comparison of ranking functions) *Let $L \in \{L_{dcg}, L_{err}\}$ and $\tau_q^2 = \lim_{n \rightarrow \infty} n \text{Var}[\hat{\Delta}_{n,q}]$. The optimization problem*

$$(q^*, n^*) = \arg \min_{q, n} \frac{1}{n} \tau_q \quad \text{s.t.} \quad \mathbb{E} \left[\sum_{j=1}^n \lambda(x_j) \right] \leq \Lambda$$

is solved by

$$q^*(x) \propto \frac{p(x)}{\sqrt{\lambda(x)}} \sqrt{\int (\delta(x, \mathbf{y}) - \Delta)^2 p(\mathbf{y}|x) d\mathbf{y}}, \quad n^* = \frac{\Lambda}{\int \lambda(x) q(x) dx}. \quad (13)$$

Before we prove Theorems 1 and 2, we state the following two lemmata:

Lemma 1 (Asymptotic variance) *Let $\hat{R}_{n,q}$, $\hat{\Delta}_{n,q}$, R , and Δ be defined as above. Then*

$$\lim_{n \rightarrow \infty} n \text{Var}[\hat{R}_{n,q}] = \int \int \frac{p^2(x)}{q^2(x)} (L(\mathbf{r}(x), \mathbf{y}) - R)^2 p(\mathbf{y}|x) q(x) dx d\mathbf{y} \quad (14)$$

$$\lim_{n \rightarrow \infty} n \text{Var}[\hat{\Delta}_{n,q}] = \int \int \frac{p(x)^2}{q(x)^2} (\delta(x, \mathbf{y}) - \Delta)^2 p(\mathbf{y}|x) q(x) dx d\mathbf{y} \quad (15)$$

for $L \in \{L_{dcg}, L_{err}\}$.

The proof of Lemma 1 closely follows results from Sawade et al. (2010, 2012) and is included in the [Appendix](#).

Lemma 2 (Maximizing distribution) *Let $a : \mathcal{X} \rightarrow \mathbb{R}$ and $\lambda : \mathcal{X} \rightarrow \mathbb{R}$ denote functions on the query space such that $\int \sqrt{a(x)} dx$ exists and $\lambda(x) \geq \epsilon > 0$. The functional*

$$G[q] = \left(\int \frac{a(x)}{q(x)} dx \right) \left(\int \lambda(x) q(x) dx \right),$$

where $q(x)$ is a distribution over the query space $\mathcal{X}$, is minimized over q by setting

$$q(x) \propto \sqrt{\frac{a(x)}{\lambda(x)}}.$$

A proof of Lemma 2 is included in the [Appendix](#). We now prove Theorems 1 and 2.

Proof of Theorems 1 and 2 We first study the minimization of $\frac{1}{n} \sigma_q^2$ in Theorem 1. Since

$$\mathbb{E} \left[\sum_{j=1}^n \lambda(x_j) \right] = n \int \lambda(x) q(x) dx,$$

the minimization problem can be reformulated as

$$\min_q \min_n \frac{1}{n} \sigma_q^2 \quad \text{s.t.} \quad n \leq \frac{\Lambda}{\int \lambda(x) q(x) dx}.$$

Clearly $n^* = \Lambda / \int \lambda(x)q(x)dx$ solves the inner optimization. Since Λ is a constant, the remaining minimization over q is

$$\begin{aligned} q^* &= \arg \min_q \frac{\int \lambda(x)q(x)dx}{\Lambda} \sigma_q^2 \\ &= \arg \min_q \sigma_q^2 \int \lambda(x)q(x)dx. \end{aligned}$$

Equation (14) in Lemma 1 implies

$$\sigma_q^2 = \iint \frac{p^2(x)}{q^2(x)} (L(\mathbf{r}(x), \mathbf{y}) - R)^2 p(\mathbf{y}|x)q(x)dx d\mathbf{y}.$$

Setting $a(x) = p^2(x) \int (L(\mathbf{r}(x), \mathbf{y}) - R)^2 p(\mathbf{y}|x) d\mathbf{y}$ and applying Lemma 2 implies Eq. (12).

For the minimization of $\frac{1}{n} \tau_q^2$ in Theorem 2 we analogously derive

$$q^* = \arg \min_q \tau_q^2 \int \lambda(x)q(x)dx.$$

Equation (15) in Lemma 1 implies

$$\tau_q^2 = \iint \frac{p(x)^2}{q(x)^2} (\delta(x, \mathbf{y}) - \Delta)^2 p(\mathbf{y}|x)q(x)dx d\mathbf{y}.$$

Setting $a(x) = p^2(x) \int (\delta(x, \mathbf{y}) - \Delta)^2 p(\mathbf{y}|x) d\mathbf{y}$ and applying Lemma 2 implies Eq. (13). $\square$

4 Empirical sampling distribution

The sampling distributions prescribed by Theorems 1 and 2 depend on the unknown test distribution $p(x)$. We now turn towards a setting in which a pool D of m unlabeled queries is available. Queries from this pool can be sampled and then labeled at a cost. Drawing queries from the pool replaces generating them under the test distribution; that is, $p(x) = \frac{1}{m}$ for all $x \in D$.

The optimal sampling distribution also depends on the true conditional distribution $p(\mathbf{y}|x) = \prod_{z \in \mathcal{Z}} p(y_z|x, z)$ (Eq. (3)). To implement the method, we approximate $p(y_z|x, z)$ by a model $p(y_z|x, z; \theta)$ of graded relevance. For the large class of pointwise ranking methods—that is, methods that produce a ranking by predicting graded relevance scores for query-document pairs and then sorting documents according to their score—such a model can typically be derived from the graded relevance predictor. Finally, the sampling distributions depend on the true performance $R[\mathbf{r}]$ as given by Eq. (4), or $\Delta = R[\mathbf{r}_1] - R[\mathbf{r}_2]$. $R[\mathbf{r}]$ is replaced by an introspective performance $R_\theta[\mathbf{r}]$ calculated from Eq. (4), where the integral over $\mathcal{X}$ is replaced by a sum over the pool, $p(x) = \frac{1}{m}$, and $p(\mathbf{y}|x) = \prod_{z \in \mathcal{Z}} p(y_z|x, z; \theta)$. Note that because the integral over $\mathcal{X}$ in Eq. (4) is replaced by a sum over a pool of unlabeled data D , the empirical sampling distribution is chosen such that we accurately estimate ranking performance over this pool rather than over the entire instance space $\mathcal{X}$. However, assuming that a large enough pool of unlabeled data is available, ranking performance over the entire pool D will closely approximate ranking performance over $\mathcal{X}$.

The performance difference Δ is approximated by $\Delta_\theta = R_\theta[\mathbf{r}_1] - R_\theta[\mathbf{r}_2]$. Note that as long as $p(x) > 0$ implies $q(x) > 0$, the weighting factors ensure that such approximations do not introduce an asymptotic bias in our estimator (Eq. (6)).

With these approximations, we arrive at the following empirical sampling distributions.

Corollary 1 When relevance labels for individual items are independent given the query (Eq. (3)), and $p(y_z|x, z)$ is approximated by a model $p(y|x, z; \theta)$ of graded relevance, the sampling distributions minimizing $\frac{1}{n}\sigma_q^2$ and $\frac{1}{n}\tau_q^2$ in a pool-based setting resolve to

$$q^*(x) \propto \frac{1}{\sqrt{\lambda(x)}} \sqrt{\mathbb{E}[(L(\mathbf{r}(x), \mathbf{y}) - R_\theta)^2|x; \theta]} \quad (16)$$

and

$$q^*(x) \propto \frac{1}{\sqrt{\lambda(x)}} \sqrt{\mathbb{E}[(\delta(x, \mathbf{y}) - \Delta_\theta)^2|x; \theta]}, \quad (17)$$

respectively. Here, for any function $g(x, \mathbf{y})$ of a query x and label vector $\mathbf{y}$,

$$\mathbb{E}[g(x, \mathbf{y})|x; \theta] = \sum_{\mathbf{y} \in \mathcal{Y}^Z} g(x, \mathbf{y}) \prod_{z \in Z} p(y_z|x, z; \theta) \quad (18)$$

denotes expectation of $g(x, \mathbf{y})$ with respect to label vectors $\mathbf{y}$ generated according to $p(y_z|x, z, \theta)$.

The corollary directly follows from inserting the approximations $p(x) \approx \frac{1}{n}$ and $p(\mathbf{y}|x) \approx \prod_{z \in Z} p(y_z|x, z; \theta)$ into Eqs. (12) and (13).

We observe that for the evaluation of a single given ranking function $\mathbf{r}$ (Eq. (16)), the empirical sampling distribution gives preference to queries x with low costs $\lambda(x)$ and for which the expected ranking performance deviates strongly from the average expected ranking performance R_θ ; the expectation is taken with respect to the available graded relevance model θ . For the comparison of two given ranking functions $\mathbf{r}_1$ and $\mathbf{r}_2$ (Eq. (17)), preference is given to queries x with low costs and for which the difference in performance $L(\mathbf{r}_1(x), \mathbf{y}) - L(\mathbf{r}_2(x), \mathbf{y})$ is expected to be high (note that Δ_θ will typically be small).

Computation of the empirical sampling distributions given by Eqs. (16) and (17) requires the computation of $\mathbb{E}[g(x, \mathbf{y})|x; \theta]$, which is defined in terms of a sum over exponentially many relevance label vectors $\mathbf{y} \in \mathcal{Y}^Z$. The following theorem states that the empirical sampling distributions can nevertheless be computed in polynomial time:

Theorem 3 (Polynomial-time computation of sampling distributions) *The sampling distribution given by Eq. (16) can be computed in time*

$$\mathcal{O}\left(|\mathcal{Y}||D| \max_x |\mathbf{r}(x)|\right) \quad \text{for } L \in \{L_{deg}, L_{err}\}.$$

The sampling distribution given by Eq. (17) can be computed in time

$$\begin{aligned} &\mathcal{O}\left(|\mathcal{Y}||D| \max_x (|\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|)\right) \quad \text{for } L = L_{deg}, \\ &\mathcal{O}\left(|\mathcal{Y}||D| \max_x (|\mathbf{r}_1(x)| \cdot |\mathbf{r}_2(x)|)\right) \quad \text{for } L = L_{err}. \end{aligned}$$

A detailed proof of Theorem 3 is given in the [Appendix](#). The general idea behind the proof is to compute the required quantities using dynamic programming. Specifically, after substituting Eqs. (1) and (2) into Eqs. (16) and (17) and exploiting the independence assumption given by Eq. (3), Eqs. (16) and (17) can be decomposed into cumulative sums and products of expectations over individual item labels $y \in \mathcal{Y}$. These sums and products can be computed in polynomial time.

Algorithm 1 summarizes the active estimation algorithm. It samples queries $x_1, \dots, x_n$ with replacement from the pool according to the distribution prescribed by Corollary 1 and

Algorithm 1 Active estimation of ranking performance

input Ranking function $\mathbf{r}$ or pair of ranking functions $\mathbf{r}_1, \mathbf{r}_2$; graded relevance model $p(y_z|x, z; \theta)$; pool D , labeling budget Λ .

- 1: Compute sampling distribution q^* (Corollary 1, Eq. (16) or (17)).
- 2: Initialize $n \leftarrow 0$.
- 3: Draw $x_1 \sim q^*(x)$ from D with replacement.
- 4: **while** $\sum_{j=1}^{n+1} \lambda(x_j) \leq \Lambda$ **do**
- 5: Obtain $\mathbf{y}_{n+1} \sim p(\mathbf{y}|x_{n+1})$ from human labeler (restrict to items in rankings).
- 6: Update number of labeled instances $n \leftarrow n + 1$.
- 7: Draw $x_{n+1} \sim q^*(x)$ from D with replacement.
- 8: **end while**
- 9: Compute $\hat{R}_{n,q}[\mathbf{r}]$ or $\hat{\Delta}_{n,q} = \hat{R}_{n,q}[\mathbf{r}_1] - \hat{R}_{n,q}[\mathbf{r}_2]$ (Eq. (6)).

output $\hat{R}_{n,q}[\mathbf{r}]$ or $\hat{\Delta}_{n,q}$.

obtains relevance labels from a human labeler for all items included in $\mathbf{r}(x_i)$ or $\mathbf{r}_1(x_i) \cup \mathbf{r}_2(x_i)$ until the labeling budget Λ is exhausted. Note that queries can be drawn more than once; in the special case that the labeling process is deterministic, recurring labels can be looked up rather than be queried from the deterministic labeling oracle repeatedly. Hence, the actual labeling costs may stay below $\sum_{j=1}^n \lambda(x_j)$. In this case, the loop is continued until the labeling budget Λ is exhausted.

5 Empirical studies

We compare active estimation of ranking performance (Algorithm 1, labeled *active*) to estimation based on a test sample drawn uniformly from the pool (Eq. (5), labeled *passive*). Algorithm 1 requires a model $p(y_z|x, z; \theta)$ of graded relevance in order to compute the sampling distribution q^* from Corollary 1. If no such model is available, a uniform distribution $p(y_z|x, z; \theta) = \frac{1}{|Y|}$ can be used instead (labeled *active_{uniD}*). To quantify the effect of modeling costs, we also study a variant of Algorithm 1 that assumes uniform costs $\lambda(x) = 1$ in Eqs. (16) and (17) (labeled *active_{uniC}*). This variant implements active risk estimation (Sawade et al. 2010) and active comparison (Sawade et al. 2012) for ranking; we have shown how the resulting sampling distributions can be computed in polynomial time (Corollary 1 and Theorem 3).

Experiments are performed on the Microsoft Learning to Rank data set MSLR-WEB30k (Microsoft Research 2010). It contains 31,531 queries, and a set of documents for each query whose relevance for the query has been determined by human labelers in the process of developing the Bing search engine. The resulting 3,771,125 query-document pairs are represented by 136 features widely used in the information retrieval community (such as query term statistics, page rank, and click counts). Relevance labels take values from 0 (irrelevant) to 4 (perfectly relevant).

The data are split into five folds. On one fold, we train ranking functions using different graded relevance models (details below). The remaining four folds serve as a pool of unlabeled test queries; we estimate (Sect. 5.1) or compare (Sect. 5.2) the performance of the ranking functions by drawing and labeling queries from this pool according to Algorithm 1 and the baselines discussed above. Test queries are drawn until a labeling budget Λ is exhausted. To quantify the human effort realistically, we model the labeling costs $\lambda(x)$ for a query x as proportional to a sum of costs incurred for labeling individual documents

Fig. 1 Distribution of the number of documents per query (*left*) and the query labeling costs $\lambda(x)$ (*right*) in the MSLR-WEB30k data set

$z \in \mathbf{r}(x)$; labeling costs for a single document z are assumed to be logarithmic in the document length.

All evaluation techniques, both active and passive, can approximate L_{dcg} and L_{err} for a query x by requesting labels only for the first k documents in the ranking. The number of documents for which the MSLR-WEB30k data set provides labels varies over the queries at an average of 119 documents per query. In our experiments, we use all documents for which labels are provided for each query and for all evaluation methods under investigation. Figure 1 (left) shows the distribution of the number of documents per query over the entire data set. The cost unit is chosen such that average labeling costs for a query are one. Figure 1 (right) shows the distribution of labeling costs $\lambda(x)$ of queries over the entire dataset. All results are averaged over the five folds and 5,000 repetitions of the evaluation process. Error bars indicate the standard error.

5.1 Estimating ranking performance

Based on the outcome of the 2010 Yahoo ranking challenge (Chapelle and Chang 2011; Mohan et al. 2011), we choose a pointwise ranking approach and employ Random Forest regression (Breiman 2001) to train graded relevance models on query-document pairs. Specifically, we learn Random Forest regression models for the binary problems $\{y_z \leq k\}$ versus $\{y_z > k\}$ for relevance labels $k \in \{1, \dots, 4\}$. The regression values can be interpreted as probabilities $p(y_z \leq k|x, z; \theta)$ (Li et al. 2007, Sect. 4). This allows us to derive probability estimates $p(y_z = k|x, z; \theta) = p(y_z \leq k|x, z; \theta) - p(y_z \leq k-1|x, z; \theta)$ required by Algorithm 1. The ranking function is obtained by returning all documents associated with a query sorted according to their predicted graded relevance (Li et al. 2007, Sect. 3.3). We denote this model by RF_{class} . As an alternative graded relevance model, we also study a MAP version of ordered logistic regression (McCullagh 1980). The Ordered Logit model (denoted by *Logit*) is a modification of the logistic regression model which accounts for the ordering of the relevance labels; the target variable is $y_z \leq k$ rather than a particular relevance level $y_z = k$. This model directly provides probability estimates $p(y_z|x, z; \theta)$. Half of the available training fold is used for model training, the other half is used as a validation set to tune hyperparameters of the respective ranking model.

Figure 2 shows absolute deviation between true ranking performance and estimated ranking performance as a function of the labeling budget Λ for the performance measures ERR

Fig. 2 Estimation error over Λ when evaluating Random Forest (top row) and Ordered Logit (bottom row) with performance measure ERR (left) and DCG (right). Error bars indicate the standard error

and DCG. True performance is taken to be the performance over all queries in the pool D . We observe that active estimation is more accurate than passive estimation; the labeling budget can be reduced from $\Lambda = 300$ by between 10 % and 20 % depending on the ranking method and performance measure under study. Note that DCG and ERR have different ranges; L_{err} takes values between zero and one, whereas L_{dgc} is potentially unbounded and increases with the number of ranked items $|\mathbf{r}(x)|$. For Ordered Logit models the true performance is between 26.5 and 33.0 for DCG and between 0.31 to 0.34 for ERR. For RF_{class} models true performance is between 27.2 and 33.4 for DCG and between 0.35 and 0.38 for ERR.

5.2 Comparing ranking performance

We additionally train linear Ranking SVM (Herbrich et al. 2000) (denoted by SVM) and a Random Forests regression model (Li et al. 2007; Mohan et al. 2011) directly estimating relevance labels (denoted by RF_{reg}), and compare the resulting ranking functions to those of the RF_{class} and $Logit$ models. For the comparison of RF_{class} vs. $Logit$ both models provide us with estimates $p(y_z|x, z; \theta)$; in this case a mixture model is employed as proposed in Sawade et al. (2012). We measure *model selection error*, defined as the fraction of experiments in which an evaluation method does not correctly identify the model with higher true performance. Figure 3 shows model selection error as a function of the available labeling budget for different pairwise comparisons and the performance measures ERR and DCG. Active estimation more reliably identifies the model with higher ranking performance, saving between 30 % and 55 % of labeling effort compared to passive estimation at $\Lambda = 300$. We observe that the gains of active versus passive estimation are not only due to differences

Fig. 3 Model selection error over Λ when comparing Random Forest regression vs. classification (*top row*), and Ordered Logit vs. Ranking SVM (*middle row*) or Random Forest classification (*bottom row*). The performance measure is ERR (*left*) or DCG (*right*). Error bars indicate the standard error

in query costs: the baseline *active_{uniC}*, which does not take into account query costs for computing the sampling distribution, performs almost as well as *active*.

As a further comparative evaluation we simulate an index update. An outdated index with lower coverage is simulated by randomly removing 10 % of all query-document pairs from each result list $\mathbf{r}(x)$ for all queries. We use *RF_{class}* as the ranking model. Active and passive estimation methods are applied to estimate the difference in performance between models based on the outdated and current index. Figure 4 shows absolute deviation of estimated from true performance difference over labeling budget Λ for the performance measures ERR and DCG. We observe that active estimation quantifies the impact of the index update more accurately than passive estimation, saving approximately 75 % of labeling effort for the performance measure ERR and about 30 % labeling effort for DCG at $\Lambda = 300$.

We finally simulate the incorporation of novel sources of training data by comparing a Random Forest model trained on 100,000 query-document pairs ($\mathbf{r}_1$) to a Random Forest

Fig. 4 Absolute estimation error over labeling costs Λ for a simulated index update affecting 10 % of items for each query. The performance measure is ERR (left) or DCG (right). Error bars indicate the standard error

Fig. 5 Absolute estimation error comparing ranking functions trained on 100,000 vs. 200,000 query-document pairs over Λ (top row), and over training set size of second model at $\Lambda = 100$ (bottom row). The performance measure is ERR (left) or DCG (right). Error bars indicate the standard error

model trained on between 120,000 and 200,000 query-document pairs ($\mathbf{r}_2$). The difference in performance between $\mathbf{r}_1$ and $\mathbf{r}_2$ is estimated using active and passive methods. Figure 5 (top row) shows absolute deviation of estimated from true performance difference for models trained on 100,000 and 200,000 instances as a function of Λ . Active estimation quantifies the performance gain from additional training data more accurately, reducing labeling costs by approximately 45 % (ERR) and 30 % (DCG). Figure 5 (bottom row) shows estimation error as a function of the number of query-document pairs the model $\mathbf{r}_2$ is trained on for

Fig. 6 Heatmap of the sampling distribution $q^*(x)$ over the pool D when evaluating a Random Forest model in terms of DCG, plotted into a two-dimensional space with axes $\lambda(x)$ and $\mathbb{E}[L(x, \mathbf{y})|x; \theta]$ (left). Number of queries drawn by passive and active estimation methods for a labeling budget of $\Lambda = 100$ when evaluating a Random Forest model in terms of DCG (right)

$\Lambda = 100$. Active estimation significantly reduces estimation error compared to passive estimation for all training set sizes.

5.3 Influence of query costs on the sampling distribution

The empirical sampling distributions prescribed by Corollary 1 select queries $x \in D$ based on their cost $\lambda(x)$ and intrinsic expectations of ranking performance given by R_θ (Eq. (16)) and Δ_θ (Eq. (17)), respectively. Figure 6 (left) shows a representative example of the sampling distribution $q^*(x)$ given by Eq. (16) plotted into a two-dimensional space with axes $\lambda(x)$ and intrinsic expected ranking performance $\mathbb{E}[L(x, \mathbf{y})|x; \theta]$ for the query x . We observe that low-cost queries are preferred over high-cost queries, and queries for which the intrinsic ranking performance $\mathbb{E}[L(x, \mathbf{y})|x; \theta]$ for query x is far away from the intrinsic overall performance R_θ are more likely to be chosen (in the example, approximately $R_\theta = 26$).

Optimization problems (7) and (8) constitute a trade-off between labeling costs and informativeness of a test query: optimization over n implies that many inexpensive or few expensive queries could be chosen. On average, active estimation prefers to draw more (but cheaper) queries than passive estimation (Eqs. (16) and (17)). Figure 6 (right) shows the number of queries actually drawn by the different evaluation methods for a labeling budget of $\Lambda = 100$ when estimating absolute DCG for RF_{class} . We observe that the active estimation methods that take into account costs in the computation of the optimal sampling distribution (*active* and *active_{uniD}*) draw more instances than *passive* and *active_{uniC}*.

6 Related work

There has been significant interest in learning ranking functions from data in order to improve the relevance of search results (Borges 2010; Li et al. 2007; Mohan et al. 2011; Zheng et al. 2007). This has partly been driven by the recent release of large-scale datasets derived from commercial search engines, such as the Microsoft Learning to Rank datasets (Microsoft Research 2010) and the Yahoo Learning to Rank Challenge datasets (Chapelle and Chang 2011). These datasets serve as realistic benchmarks for evaluating and comparing the performance of different ranking algorithms.

Most approaches for learning ranking functions employ the pointwise ranking model (e.g. Li et al. 2007; Mohan et al. 2011). In pointwise ranking, the problem of learning rankings is reduced to the problem of predicting graded relevance scores for individual query-document pairs; a ranking is then obtained by sorting documents according to their relevance score for a given query. The performance measures DCG (see Li et al. 2007, Lemma 1) and ERR (see Mohan et al. 2011, Theorem 1) can be upper-bounded by the classification error; these upper bounds can be seen as one justification for pointwise ranking approaches. However, there has also been work on directly maximizing performance measures defined over entire ranking such as normalized DCG (Valizadegan et al. 2010; Burges 2010).

To reduce the amount of training data that needs to be relevance-labeled, several approaches for active learning of ranking functions have been proposed (Long et al. 2010; Radlinski and Joachims 2007). The active performance estimation problem discussed in this paper can be considered a dual problem of active learning: in active learning, the goal of the selection process is to obtain maximally accurate predictions or model parameters; our goal is to obtain maximally accurate performance estimates for existing ranking models.

There are two possible directions for actively selecting training or evaluation data. One direction is to actively select most informative queries for which documents should be retrieved and labeled (the approach followed in this paper). Another possible direction to reduce labeling effort is to actively select a subset of documents that should be labeled for each query. Long et al. (2010) employ both document sampling and query sampling to minimize prediction variance of ranking models in active learning. Document sampling is also used by Carterette et al. (2006) in order to decide which of two ranking functions achieves higher *precision at k*, by Aslam et al. (2006) to obtain unbiased estimates of mean average precision and mean R-precision for ranking models, and by Carterette and Smucker (2007) to perform statistical significance testing for pairs of ranking models.

For the estimation of DCG and ERR studied in this paper, document sampling is not directly applicable, since the considered performance measures depend on the relevance labels of all individual documents for a given query (see Eqs. (1) and (2)). For ERR even the discounting factor for a specific ranking position can only be determined if the relevance of all higher-ranked documents is known. Consequently, if arbitrary subsets of documents are sampled, the sum in Eq. (2) could only be computed up to the first factor associated with a document not contained in the subset. Substituting missing relevance labels with some estimate or default value would yield a final performance estimate that can be strongly biased, as a single incorrect relevance label affects the discounting factors for all subsequent positions. This problem could be avoided if we label a prefix of the list of all documents associated with a query, thereby truncating the sums in Eqs. (1) and (2). While truncation would still introduce an estimation bias, the size of this bias could be bounded for a given ranking and traded off against the reduction in variance that can be achieved by labeling additional queries. Studying the bias-variance trade-off associated with such truncated label vectors is an interesting direction for future work.

The active performance estimation approach presented in this paper employs importance weighting to compensate for the bias introduced by the instrumental distribution. Importance weighting is a widely studied technique that has been used in several active learning approaches, for example with exponential family models (Bach 2007) or SVMs (Beygelzimer et al. 2009).

7 Conclusions

We have studied the problem of estimating or comparing the performance of given ranking functions as accurately as possible on a fixed budget for labeling item relevance. Starting with an analysis of the sources of estimation error, we have derived distributions for sampling test queries that asymptotically minimize the error of importance-weighted performance estimators for the ranking measures Discounted Cumulative Gain and Expected Reciprocal Rank (Theorems 1 and 2). To apply the method in practice, we assume that a large pool of unlabeled test queries drawn from the true query distribution is available; Corollary 1 prescribes an empirical distribution for sampling queries from this pool. A naïve computation of this empirical sampling distribution is exponential; however, we have shown that the distribution can be computed in polynomial time using dynamic programming (Theorem 3).

A crucial feature of ranking domains is that labeling costs vary for different queries as a function of the number of documents returned for a query and document-specific features. This is in contrast to settings studied in existing approaches for active evaluation (Sawade et al. 2010, 2012). Our cost-sensitive analysis has shown that to minimize the estimation error, sampling weights should be inversely proportional to the square root of the labeling costs of a document (Theorems 1 and 2).

Empirically, we observed that active estimates of ranking performance are more accurate than passive estimates. In different experimental settings—estimation of the performance of a single ranking model, comparison of different types of ranking models, simulated index updates—performing active estimation resulted in saved labeling efforts of between 10 % and 75 %. The advantage of using an active estimation strategy is particularly strong when evaluating the relative performance of different ranking models; here, the sampling strategy is based on choosing queries that highlight the differences between the models (Eq. (17)). In contrast, the gains for absolute performance estimation for a single ranking model were relatively small in our experiments. Furthermore, we observed that the gains of active estimation are slightly more pronounced when estimating ERR as compared to estimating DCG.

Acknowledgements We would like to thank the anonymous reviewers for their helpful comments and suggestions for improving the paper. We gratefully acknowledge that this work was supported by a Google Research Award.

Appendix

Proof of Lemma 1 This proof closely follows Sawade et al. (2010, 2012), but is included here to make the presentation more self-contained. Let

$$G = \iint g(x, \mathbf{y}) p(x, \mathbf{y}) dx d\mathbf{y},$$

$$\hat{G}_{n,q} = \left(\sum_{j=1}^n \frac{p(x_j)}{q(x_j)} \right)^{-1} \sum_{j=1}^n \frac{p(x_j)}{q(x_j)} g(x_j, \mathbf{y}_j)$$

for a function $g : \mathcal{X} \times \mathcal{Y}^Z \rightarrow \mathbb{R}$ and $(x_j, \mathbf{y}_j) \sim q(x)p(\mathbf{y}|x)$. Let $\hat{G}_{n,q}^0 = \sum_{j=1}^n w_j g_j$ and $W_n = \sum_{j=1}^n w_j$ with $w_j = \frac{p(x_j)}{q(x_j)}$ and $g_j = g(x_j, \mathbf{y}_j)$. We note that for examples drawn according to $q(x)$, $\mathbb{E}[\hat{G}_{n,q}^0] = nG$ and $\mathbb{E}[W_n] = n$. The random variables $w_1, \dots, w_n$ and

$w_1 g_1, \dots, w_n g_n$ are *iid*, therefore the central limit theorem implies that $\frac{1}{n} \hat{G}_{n,q}^0$ and $\frac{1}{n} W_n$ are asymptotically normally distributed with

$$\begin{aligned}\sqrt{n} \left(\frac{1}{n} \hat{G}_{n,q}^0 - G \right) &\xrightarrow{n \rightarrow \infty} \mathcal{N}(0, \text{Var}[w_j g_j]) \\ \sqrt{n} \left(\frac{1}{n} W_n - 1 \right) &\xrightarrow{n \rightarrow \infty} \mathcal{N}(0, \text{Var}[w_j]),\end{aligned}$$

where $\xrightarrow{n \rightarrow \infty}$ denotes convergence in distribution. We employ the multivariate *delta method* (see, e.g. Wasserman 2004, Chap. 5) to extend the convergence results for $\hat{G}_{n,q}^0$ and W_n to a convergence result for the normalized estimator $\hat{G}_{n,q}$. The delta method allows to derive the asymptotic distribution of a differentiable function f whose input variables are asymptotically normally distributed. Applying it to the function $f(a, b) = \frac{a}{b}$ with $a = \frac{1}{n} \hat{G}_{n,q}^0$ and $b = \frac{1}{n} W_n$ yields

$$\sqrt{n} \left(\frac{\frac{1}{n} \hat{G}_{n,q}^0}{\frac{1}{n} W_n} - G \right) \xrightarrow{n \rightarrow \infty} \mathcal{N}(0, \nabla f(G, 1)^\top \Sigma \nabla f(G, 1)), \quad (19)$$

where ∇f denotes the gradient of f and Σ is the (asymptotic) covariance matrix of the input arguments

$$\Sigma = \begin{pmatrix} \text{Var}[w_j g_j] & \text{Cov}[w_j g_j, w_j] \\ \text{Cov}[w_j g_j, w_j] & \text{Var}[w_j] \end{pmatrix}.$$

Taking the variance of both sides of Eq. (19), we observe that

$$\lim_{n \rightarrow \infty} n \text{Var}[\hat{G}_{n,q}] = \nabla f(G, 1)^\top \Sigma \nabla f(G, 1). \quad (20)$$

Finally,

$$\begin{aligned}\nabla f(G, 1)^\top \Sigma \nabla f(G, 1) &= \text{Var}[w_j g_j] - 2G \text{Cov}[w_j, w_j g_j] + G^2 \text{Var}[w_j] \\ &= \mathbb{E}[w_j^2 g_j^2] - 2G \mathbb{E}[w_j^2 g_j] + G^2 \mathbb{E}[w_j^2] \\ &= \int \int \frac{p(x)^2}{q(x)^2} (g(x, \mathbf{y}) - G)^2 p(\mathbf{y}|x) q(x) dx d\mathbf{y}.\end{aligned} \quad (21)$$

For $g(x, \mathbf{y}) = L(\mathbf{r}(x), \mathbf{y})$, $L \in \{L_{deg}, L_{err}\}$, Eqs. (20) and (21) imply Eq. (14). For $g(x, \mathbf{y}) = \delta(x, \mathbf{y})$, Eqs. (20) and (21) imply Eq. (15). $\square$

Proof of Lemma 2 We have to minimize the functional

$$\left(\int \frac{a(x)}{q(x)} dx \right) \left(\int \lambda(x) q(x) dx \right) \quad (22)$$

in terms of q under the constraints $\int q(x) dx = 1$ and $q(x) > 0$. We first note that Objective (22) is invariant under multiplicative rescaling of $q(x)$, thus the constraint $\int q(x) dx = 1$ can be dropped during optimization and enforced in the end by normalizing the unconstrained solution. We reformulate the problem as

$$\min_q C \int \frac{a(x)}{q(x)} dx \quad \text{s.t. } C = \int \lambda(x) q(x) dx \quad (23)$$

which we solve using a Lagrange multiplier α by

$$\min_q C \int \frac{a(x)}{q(x)} dx + \alpha \left(\int \lambda(x) q(x) dx - C \right).$$

The optimal point for the constrained problem satisfies the Euler-Lagrange equation

$$\alpha \lambda(x) = C \frac{a(x)}{q(x)^2},$$

and therefore

$$q(x) = \sqrt{C \frac{a(x)}{\alpha \lambda(x)}}. \quad (24)$$

Resubstitution of Eq. (24) into the constraint (Eq. (23)) leads to

$$C = \int \sqrt{C \frac{a(x)}{\alpha \lambda(x)}} \lambda(x) dx,$$

solving for α we obtain

$$\alpha = \frac{(\int \sqrt{C a(x) \lambda(x)} dx)^2}{C^2}. \quad (25)$$

Finally, resubstitution of Eq. (25) into Eq. (24) proves the claim. $\square$

Proof of Theorem 3 In order to show that the empirical sampling distributions given by Eqs. (16) and (17) can be computed efficiently, we have to show that Eq. (18) can be computed efficiently. This can be done by suitable algebraic manipulation, exploiting the independence assumption given by Eq. (3).

Empirical sampling distribution for absolute estimation (Eq. (16)) with $L = L_{dcg}$ It suffices to show that the intrinsic risk R_θ can be computed in time

$$\mathcal{O}(|\mathcal{Y}| |D| \max_x |\mathbf{r}(x)|),$$

and that for any $x \in D$ the quantity $\mathbb{E}[(L(\mathbf{r}(x), \mathbf{y}) - R_\theta)^2 | x; \theta]$ can be computed in time $\mathcal{O}(|\mathcal{Y}| |\mathbf{r}(x)|)$ given R_θ . We first note that for any $z \in \mathcal{Z}$, it holds that

$$\begin{aligned} & \mathbb{E}[\ell_{dcg}(y_z, i) | x; \theta] \\ &= \sum_{\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}} \ell_{dcg}(y_z, i) \prod_{z' \in \mathcal{Z}} p(y_{z'} | x, z'; \theta) \\ &= \sum_{y_z} \sum_{\mathbf{y} \in \mathcal{Y}^{\mathcal{Z} \setminus \{z\}}} \ell_{dcg}(y_z, i) \prod_{z' \in \mathcal{Z}} p(y_{z'} | x, z'; \theta) \\ &= \sum_{y_z} \ell_{dcg}(y_z, i) p(y_z | x, z; \theta) \sum_{\mathbf{y} \in \mathcal{Y}^{\mathcal{Z} \setminus \{z\}}} \prod_{z' \in \mathcal{Z} \setminus \{z\}} p(y_{z'} | x, z'; \theta) \\ &= \sum_{y_z} \ell_{dcg}(y_z, i) p(y_z | x, z; \theta), \end{aligned} \quad (26)$$

where $\mathbf{y} \in \mathcal{Y}^{\mathcal{Z} \setminus \{z\}}$ is a vector of relevance labels $y_{z'}$ for all $z' \in \mathcal{Z} \setminus \{z\}$. The quantity $\mathbb{E}[\ell_{dcg}(y_z, i) | x; \theta]$ can thus be computed in time $\mathcal{O}(|\mathcal{Y}|)$.

Let $\ell_i = \ell_{dcg}(y_{r_i(x)}, i)$, such that the DCG can be expressed as $L_{dcg}(\mathbf{r}(x), \mathbf{y}) = \sum_{i=1}^{|\mathbf{r}(x)|} \ell_i$. Then, for $L = L_{dcg}$ it holds that

$$\begin{aligned}
R_\theta &= \frac{1}{|D|} \sum_{x \in D} \mathbb{E}[L(\mathbf{r}(x), \mathbf{y})|x; \theta] \\
&= \frac{1}{|D|} \sum_{x \in D} \sum_{\mathbf{y} \in \mathcal{Y}^Z} \sum_{i=1}^{|\mathbf{r}(x)|} \ell_i \prod_{z \in Z} p(y_z|x, z; \theta) \\
&= \frac{1}{|D|} \sum_{x \in D} \sum_{i=1}^{|\mathbf{r}(x)|} \mathbb{E}[\ell_i|x; \theta],
\end{aligned}$$

therefore R_θ can be computed in time $\mathcal{O}(|\mathcal{Y}||D|\max_x |\mathbf{r}(x)|)$. We furthermore derive

$$\begin{aligned}
&\mathbb{E}[(L(\mathbf{r}(x), \mathbf{y}) - R_\theta)^2|x; \theta] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{|\mathbf{r}(x)|} \ell_i - R_\theta\right)^2|x; \theta\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{|\mathbf{r}(x)|} \ell_i^2 + 2 \sum_{i=1}^{|\mathbf{r}(x)|} \sum_{l=i+1}^{|\mathbf{r}(x)|} \ell_i \ell_l - 2R_\theta \sum_{i=1}^{|\mathbf{r}(x)|} \ell_i + R_\theta^2|x; \theta\right] \quad (27)
\end{aligned}$$

$$= \sum_{i=1}^{|\mathbf{r}(x)|} \left(\mathbb{E}[\ell_i^2|x; \theta] + 2\mathbb{E}[\ell_i|x; \theta] \sum_{l=i+1}^{|\mathbf{r}(x)|} \mathbb{E}[\ell_l|x; \theta] - 2R_\theta \mathbb{E}[\ell_i|x; \theta] \right) + R_\theta^2. \quad (28)$$

Equation (27) expands the square of sums twice and in Eq. (28) we make use of the independence assumption for item relevance (Eq. (3)) from which follows that $\mathbb{E}[\ell_i \ell_l|x; \theta] = \mathbb{E}[\ell_i|x; \theta]\mathbb{E}[\ell_l|x; \theta]$. Equation (28) can now be evaluated in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$: cumulative sums over l and all terms of the form $\mathbb{E}[\ell_i|x; \theta]$ as well as $\mathbb{E}[\ell_i^2|x; \theta]$ can be precomputed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$ (see Eq. (26)), and the second summand of Eq. (28) can then be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$. Thus, the empirical sampling distribution can overall be computed in $\mathcal{O}(|\mathcal{Y}||D|\max_x |\mathbf{r}(x)|)$. $\square$

Empirical sampling distribution for absolute estimation (Eq. (16)) with $L = L_{err}$ It suffices to show that the intrinsic risk R_θ can be computed in time

$$\mathcal{O}\left(|\mathcal{Y}||D|\max_x |\mathbf{r}(x)|\right),$$

and that for any $x \in \mathcal{X}$ the quantity $\mathbb{E}[(L(\mathbf{r}(x), \mathbf{y}) - R_\theta)^2|x; \theta]$ can be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$ given R_θ . Let $\ell_i = \ell_{err}(y_{r_i(x)})$ and $\ell_{1:i} = \frac{1}{i} \ell_i \prod_{l=1}^{i-1} (1 - \ell_l)$, such that the ERR can be expressed as $L_{err}(\mathbf{r}(x), \mathbf{y}) = \sum_{i=1}^{|\mathbf{r}(x)|} \ell_{1:i}$. From the independence assumption for item relevance (Eq. (3)), it follows that

$$\mathbb{E}[\ell_{1:i}|x; \theta] = \frac{1}{i} \mathbb{E}[\ell_{err}(y_{r_i(x)})|x; \theta] \prod_{l=1}^{i-1} (1 - \mathbb{E}[\ell_{err}(y_{r_l(x)})|x; \theta]).$$

In analogy to $L = L_{deg}$ (see Eq. (26)) it follows that

$$\mathbb{E}[\ell_{err}(y_z)|x; \theta] = \sum_{y_z} \ell_{err}(y_z) p(y_z|x, z; \theta) \quad (29)$$

for any $z \in \mathcal{Z}$ and can thus be computed in time $\mathcal{O}(|\mathcal{Y}|)$. Thus, all terms $\mathbb{E}[\ell_{1:i}|x; \theta]$ for $i = 1, \dots, |\mathbf{r}(x)|$ can together be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$, since all terms of the form $\mathbb{E}[\ell_{err}(y_z)|x; \theta]$ can be precomputed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$ and subsequently the cumulative

products over l can be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$. Therefore, R_θ for $L = L_{err}$ can be computed in time $\mathcal{O}(|\mathcal{Y}||D|\max_x |\mathbf{r}(x)|)$, since it holds that

$$\begin{aligned} R_\theta &= \frac{1}{|D|} \sum_{x \in D} \mathbb{E}[L(\mathbf{r}(x), \mathbf{y})|x; \theta] \\ &= \frac{1}{|D|} \sum_{x \in D} \sum_{\mathbf{y} \in \mathcal{Y}^Z} \sum_{i=1}^{|\mathbf{r}(x)|} \ell_{1:i} \prod_{z \in \mathcal{Z}} p(y_z|x, z; \theta) \\ &= \frac{1}{|D|} \sum_{x \in D} \sum_{i=1}^{|\mathbf{r}(x)|} \mathbb{E}[\ell_{1:i}|x; \theta]. \end{aligned}$$

We now turn towards the quantity

$$\mathbb{E}[(L(\mathbf{r}(x), \mathbf{y}) - R_\theta)^2|x; \theta].$$

We expand the square of sums twice in Eq. (30). Equation (31) follows from the independence assumption (Eq. (3)):

$$\begin{aligned} &\mathbb{E}[(L(\mathbf{r}(x), \mathbf{y}) - R_\theta)^2|x; \theta] \\ &= \mathbb{E}\left[\sum_{i=1}^{|\mathbf{r}(x)|} \ell_{1:i}^2 + 2 \sum_{i=1}^{|\mathbf{r}(x)|} \sum_{l=i+1}^{|\mathbf{r}(x)|} \ell_{1:i} \ell_{1:l} - 2R_\theta \sum_{i=1}^{|\mathbf{r}(x)|} \ell_{1:i} + R_\theta^2|x; \theta\right] \end{aligned} \quad (30)$$

$$= \sum_{i=1}^{|\mathbf{r}(x)|} \left(\mathbb{E}[\ell_{1:i}^2|x; \theta] + 2 \sum_{l=i+1}^{|\mathbf{r}(x)|} \mathbb{E}[\ell_{1:i} \ell_{1:l}|x; \theta] - 2R_\theta \mathbb{E}[\ell_{1:i}|x; \theta] \right) + R_\theta^2. \quad (31)$$

In contrast to Eq. (28) in the proof of $L = L_{dcg}$, items used to calculate $\ell_{1:i}$ and $\ell_{1:l}$ in the second summand overlap. However, we note that for $l > i$ the following decomposition holds:

$$\ell_{1:l} = \frac{1}{l} \ell_l \left(\prod_{k=1}^{i-1} (1 - \ell_k) \right) (1 - \ell_i) \left(\prod_{k=i+1}^{l-1} (1 - \ell_k) \right). \quad (32)$$

Thus $\sum_{l=i+1}^{|\mathbf{r}(x)|} \mathbb{E}[\ell_{1:i} \ell_{1:l}|x; \theta]$ can be expressed as follows:

$$\begin{aligned} &\sum_{l=i+1}^{|\mathbf{r}(x)|} \mathbb{E}\left[\frac{1}{i} \ell_i \left(\prod_{k=1}^{i-1} (1 - \ell_k) \right) \right. \\ &\quad \times \left. \frac{1}{l} \ell_l \left(\prod_{k=1}^{i-1} (1 - \ell_k) \right) (1 - \ell_i) \left(\prod_{k=i+1}^{l-1} (1 - \ell_k) \right) |x; \theta \right] \\ &= \frac{1}{i} \sum_{l=i+1}^{|\mathbf{r}(x)|} \mathbb{E}[\ell_i (1 - \ell_i)|x; \theta] \mathbb{E}\left[\prod_{k=1}^{i-1} (1 - \ell_k)^2 |x; \theta\right] \mathbb{E}\left[\frac{1}{l} \ell_l \prod_{k=i+1}^{l-1} (1 - \ell_k) |x; \theta\right] \\ &= \frac{1}{i} \mathbb{E}[\ell_i (1 - \ell_i)|x; \theta] \left(\prod_{k=1}^{i-1} \mathbb{E}[(1 - \ell_k)^2|x; \theta] \right) \\ &\quad \cdot \sum_{l=i+1}^{|\mathbf{r}(x)|} \frac{1}{l} \mathbb{E}[\ell_l|x; \theta] \frac{\prod_{k=i+1}^{|\mathbf{r}(x)|} \mathbb{E}[(1 - \ell_k)|x; \theta]}{\prod_{k=l}^{|\mathbf{r}(x)|} \mathbb{E}[(1 - \ell_k)|x; \theta]} \end{aligned} \quad (33)$$

$$\begin{aligned}
&= \frac{1}{i} \mathbb{E}[\ell_i(1 - \ell_i)|x; \theta] \left(\prod_{k=1}^{i-1} \mathbb{E}[(1 - \ell_k)^2|x; \theta] \right) \left(\prod_{k=i+1}^{|\mathbf{r}(x)|} \mathbb{E}[(1 - \ell_k)|x; \theta] \right) \\
&\quad \cdot \sum_{l=i+1}^{|\mathbf{r}(x)|} \frac{\mathbb{E}[\ell_l|x; \theta]}{l \prod_{k=l}^{|\mathbf{r}(x)|} \mathbb{E}[(1 - \ell_k)|x; \theta]}. \quad (34)
\end{aligned}$$

In Eq. (33), we insert the definition of $\ell_{1:i}$ and the decomposition given by Eq. (32) for $\ell_{1:l}$. The rest follows from the independence assumption (see Eq. (3)) and reordering terms.

Equation (34) and thus Eq. (31) can be evaluated in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$, since all cumulative products over k , cumulative sums over l , and expected values can together be pre-computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$. Thus, the empirical sampling distribution can overall be computed in time $\mathcal{O}(|\mathcal{Y}||D|\max_x |\mathbf{r}(x)|)$. $\square$

Empirical sampling distribution for comparative estimation (Eq. (17)) with $L = L_{d_{cg}}$ Since the intrinsic risks $R_\theta[\mathbf{r}_i]$ can be computed in time $\mathcal{O}(|\mathcal{Y}||D|\max_x |\mathbf{r}_i(x)|)$ and hence the difference $\Delta_\theta = R_\theta[\mathbf{r}_1] - R_\theta[\mathbf{r}_2]$ can be computed in time

$$\mathcal{O}\left(|\mathcal{Y}||D|\max_x (|\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|)\right)$$

it is sufficient to show that for any $x \in \mathcal{X}$ the quantity

$$\mathbb{E}\left[\left(L(\mathbf{r}_1(x), \mathbf{y}) - L(\mathbf{r}_2(x), \mathbf{y}) - \Delta_\theta\right)^2|x; \theta\right]$$

can be computed in time $\mathcal{O}(|\mathcal{Y}||D||\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|)$. In order to simplify notation we define an inverse ranking function

$$\mathbf{r}^{-1} : \mathcal{X} \times \mathcal{Z} \rightarrow \{1, \dots, |\mathbf{r}(x)|\} \cup \{\infty\}. \quad (35)$$

Given a query $x \in \mathcal{X}$ and an item $z \in \mathcal{Z}$ it returns the position in the list of results $\mathbf{r}^{-1} : (x, r_i(x)) \mapsto i$ if item z is contained in the results of ranking function $\mathbf{r}$ and infinity otherwise. Then, $L_{d_{cg}}$ (see Eq. (1)) can be equivalently written as

$$L_{d_{cg}}(\mathbf{r}(x), \mathbf{y}) = \sum_{z \in \mathcal{Z}} \ell_{d_{cg}}(y_z, \mathbf{r}^{-1}(x, z)),$$

since $\ell_{d_{cg}}(y, \infty) = 0$.

Let $\ell_z = \ell_{d_{cg}}(y_z, \mathbf{r}_1^{-1}(x, z)) - \ell_{d_{cg}}(y_z, \mathbf{r}_2^{-1}(x, z))$ be the difference of the loss functions for two ranking functions. Then, we derive

$$\begin{aligned}
&\mathbb{E}\left[\left(L(\mathbf{r}_1(x), \mathbf{y}) - L(\mathbf{r}_2(x), \mathbf{y}) - \Delta_\theta\right)^2|x; \theta\right] \\
&= \mathbb{E}\left[\left(\sum_{z \in \mathcal{Z}} \ell_z - \Delta_\theta\right)^2|x; \theta\right] \\
&= \mathbb{E}\left[\sum_{z \in \mathcal{Z}} \ell_z^2 + \sum_{z \in \mathcal{Z}} \sum_{\substack{\bar{z} \in \mathcal{Z} \\ \bar{z} \neq z}} \ell_z \ell_{\bar{z}} - 2 \sum_{z \in \mathcal{Z}} \ell_z \Delta_\theta + \Delta_\theta^2|x; \theta\right] \quad (36)
\end{aligned}$$

$$= \sum_{z \in \mathcal{Z}} \left(\mathbb{E}[\ell_z^2|x; \theta] + \mathbb{E}[\ell_z|x; \theta] \left(\sum_{\substack{\bar{z} \in \mathcal{Z} \\ \bar{z} \neq z}} \mathbb{E}[\ell_{\bar{z}}|x; \theta] - 2\Delta_\theta \right) \right) + \Delta_\theta^2. \quad (37)$$

Equation (36) expands the square of sums twice and in Eq. (37) we make use of the independence between different items (see Eq. (3)).

A single quantity of the form $[\ell_{\bar{z}}|x; \theta]$ can be computed in time $\mathcal{O}(|\mathcal{Y}|)$ (see Eq. (26)). Since the expectation $\mathbb{E}[\ell_{deg}(y_z, \mathbf{r}_i^{-1}(x, z))|x; \theta]$ equals zero for all $z \in \mathcal{Z}$ not contained in the results of $\mathbf{r}_i(x)$, the sums over $z \in \mathcal{Z}$ can be computed in time $\mathcal{O}(|\mathcal{Y}| \max_x (|\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|))$. Thus, the empirical sampling distribution can overall be computed in time $\mathcal{O}(|\mathcal{Y}| |D| \max_x (|\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|))$. $\square$

Empirical sampling distribution for comparative estimation (Eq. (17)) with $L = L_{err}$
 Since the intrinsic risks $R_\theta[\mathbf{r}_i]$ can be computed in time $\mathcal{O}(|\mathcal{Y}| |D| \max_x |\mathbf{r}_i(x)|)$ and hence the difference $\Delta_\theta = R_\theta[\mathbf{r}_1] - R_\theta[\mathbf{r}_2]$ can be computed in

$$\mathcal{O}\left(|\mathcal{Y}| |D| \max_x (|\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|)\right)$$

it is sufficient to show that for any $x \in \mathcal{X}$ the quantity

$$\mathbb{E}\left[\left(L(\mathbf{r}_1(x), \mathbf{y}) - L(\mathbf{r}_2(x), \mathbf{y}) - \Delta_\theta\right)^2 |x; \theta\right]$$

can be computed in time $\mathcal{O}(|\mathcal{Y}| |\mathbf{r}_1(x)| |\mathbf{r}_2(x)|)$. From the independence assumption (Eq. (3)) it follows that

$$\begin{aligned} & \mathbb{E}\left[\left(L(\mathbf{r}_1(x), \mathbf{y}) - L(\mathbf{r}_2(x), \mathbf{y}) - \Delta_\theta\right)^2 |x; \theta\right] \\ &= \sum_{k=1}^2 \left([L_{err}(\mathbf{r}_k(x), \mathbf{y})^2 |x; \theta] + 2\Delta_\theta (-1)^k \mathbb{E}[L_{err}(\mathbf{r}_k(x), \mathbf{y}) |x; \theta] \right) + \Delta_\theta^2 \\ & \quad - 2\mathbb{E}[L_{err}(\mathbf{r}_1(x), \mathbf{y}) L_{err}(\mathbf{r}_2(x), \mathbf{y}) |x; \theta]. \end{aligned} \quad (38)$$

In analogy to Eq. (31) in the proof of Theorem 3, it can be shown that all terms except $\mathbb{E}[L_{err}(\mathbf{r}_1(x), \mathbf{y}) L_{err}(\mathbf{r}_2(x), \mathbf{y}) |x; \theta]$ in Eq. (38) can be computed in linear time. In the following we show, that

$$\mathbb{E}[L_{err}(\mathbf{r}_1(x), \mathbf{y}) L_{err}(\mathbf{r}_2(x), \mathbf{y}) |x; \theta]$$

can be computed in $\mathcal{O}(|\mathcal{Y}| |\mathbf{r}_1(x)| |\mathbf{r}_2(x)|)$. Let $\ell_z = \ell_{err}(y_z)$ and the inverse ranking function $\mathbf{r}^{-1}$ be defined as in proof for $L = L_{deg}$ (see Eq. (35)). Furthermore, we define

$$\bar{\ell}_{z' < z, t} = (1 - \ell_{z'}) \mathbb{I}[\mathbf{r}_t^{-1}(x, z') < \mathbf{r}_t^{-1}(x, z)],$$

where $\mathbb{I}[\cdot] \rightarrow \{0, 1\}$ denotes the indicator function. Then, $L_{err}(\mathbf{r}(x), \mathbf{y})$ can be expressed as sum over the pool of items $\mathcal{Z}$:

$$L_{err}(\mathbf{r}(x), \mathbf{y}) = \sum_{z \in \mathcal{Z}} \frac{1}{\mathbf{r}^{-1}(x, z)} \ell_z \prod_{z' \in \mathcal{Z}} \bar{\ell}_{z' < z, t}.$$

Then we derive

$$\begin{aligned} & \mathbb{E}[L_{err}(\mathbf{r}_1(x), \mathbf{y}) L_{err}(\mathbf{r}_2(x), \mathbf{y}) |x; \theta] \\ &= \sum_{z \in \mathcal{Z}} \sum_{\bar{z} \in \mathcal{Z}} \frac{1}{\mathbf{r}_1^{-1}(x, z)} \frac{1}{\mathbf{r}_2^{-1}(x, \bar{z})} \mathbb{E}\left[\left(\ell_z \prod_{z' \in \mathcal{Z}} \bar{\ell}_{z' < z, 1}\right) \left(\ell_{\bar{z}} \prod_{z' \in \mathcal{Z}} \bar{\ell}_{z' < \bar{z}, 2}\right) |x; \theta\right]. \end{aligned} \quad (39)$$

The expectation in Eq. (39) can be decomposed as follows:

$$\begin{aligned} & \mathbb{E}\left[\left(\ell_z \prod_{z' \in \mathcal{Z}} \bar{\ell}_{z' < z, 1}\right) \left(\ell_{\bar{z}} \prod_{z' \in \mathcal{Z}} \bar{\ell}_{z' < \bar{z}, 2}\right) |x; \theta\right] \\ &= \mathbb{E}\left[\ell_z \bar{\ell}_{\bar{z} < z, 1} \left(\prod_{z' \neq \bar{z}} \bar{\ell}_{z' < z, 1}\right)\right] \end{aligned}$$

$$\times \ell_{\bar{z}} \bar{\ell}_{z' < \bar{z}, 2} \left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x, z') < \mathbf{r}_1^{-1}(x, z)}} \bar{\ell}_{z' < \bar{z}, 2} \right) \left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x, z') > \mathbf{r}_1^{-1}(x, z)}} \bar{\ell}_{z' < \bar{z}, 2} \right) | x; \theta \Big] \quad (40)$$

$$\begin{aligned} &= \mathbb{E}[\ell_z \bar{\ell}_{z' < \bar{z}, 1} \ell_{\bar{z}} \bar{\ell}_{z' < \bar{z}, 2} | x; \theta] \\ &\times \mathbb{E} \left[\left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x, z') < \mathbf{r}_1^{-1}(x, z)}} \bar{\ell}_{z' < \bar{z}, 1} \right) \left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x, z') > \mathbf{r}_1^{-1}(x, z)}} \bar{\ell}_{z' < \bar{z}, 2} \right) | x; \theta \right] \\ &\times \mathbb{E} \left[\left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x, z') > \mathbf{r}_1^{-1}(x, z)}} \bar{\ell}_{z' < \bar{z}, 2} \right) | x; \theta \right] \quad (41) \end{aligned}$$

$$= \ell_{z, \bar{z}}^- \left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x, z') < \mathbf{r}_1^{-1}(x, z)}} \mathbb{E}[\bar{\ell}_{z' < \bar{z}, 1} \bar{\ell}_{z' < \bar{z}, 2} | x; \theta] \right) \left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x, z') > \mathbf{r}_1^{-1}(x, z)}} \mathbb{E}[\bar{\ell}_{z' < \bar{z}, 2} | x; \theta] \right) \quad (42)$$

$$= \ell_{z, \bar{z}}^- \ell_{z, \bar{z}}^< \ell_{z, \bar{z}}^>, \quad (43)$$

where

$$\ell_{z, \bar{z}}^- = \begin{cases} \mathbb{E}[\ell_z^2 | x; \theta], & \text{if } z = \bar{z} \\ \mathbb{E}[\ell_z (1 - \ell_z) \mathbb{I}[\mathbf{r}_2^{-1}(x, z) < \mathbf{r}_2^{-1}(x, \bar{z})] | x; \theta] \\ \quad \cdot \mathbb{E}[\ell_{\bar{z}} (1 - \ell_{\bar{z}}) \mathbb{I}[\mathbf{r}_1^{-1}(x, \bar{z}) < \mathbf{r}_1^{-1}(x, z)] | x; \theta], & \text{if } z \neq \bar{z}, \end{cases} \quad (44)$$

$$\ell_{z, \bar{z}}^< = \prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x, z') < \mathbf{r}_1^{-1}(x, z)}} \mathbb{E}[(1 - \ell_{z'}) \mathbb{I}[\mathbf{r}_2^{-1}(x, z') < \mathbf{r}_2^{-1}(x, \bar{z}) + 1] | x; \theta], \quad (45)$$

$$\ell_{z, \bar{z}}^> = \prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x, z') > \mathbf{r}_1^{-1}(x, z)}} \mathbb{E}[(1 - \ell_{z'}) \mathbb{I}[\mathbf{r}_2^{-1}(x, z') < \mathbf{r}_2^{-1}(x, \bar{z})] | x; \theta]. \quad (46)$$

In analogy to the proof of Theorem 3 for absolute estimation with $L = L_{err}$ (see Eq. (32)), we decompose the cumulative products into disjoint item sets: In Eq. (40), we exclude factors depending on $\bar{z}$ and z , respectively. Equations (41) and (43) follow from the independence assumption (Eq. (3)). Furthermore, we summarize factors over the same items in Eq. (42). Finally, we again make use of the independence assumption for item relevance and abbreviate the products over the disjoint factors (Eq. (43)).

Individual expected values in Eqs. (44)–(46) can be computed in time $\mathcal{O}(|\mathcal{Y}|)$ (see Eq. (29)). Cumulative products in Eqs. (45) and (46) can be precomputed incrementally for all $z \in \mathcal{Z}$ using the order given by $\mathbf{r}_1^{-1}$ in time $\mathcal{O}(|\mathcal{Y}||\mathcal{Z}|)$. Since $\frac{1}{\mathbf{r}_i^{-1}(x, z)}$ equals zero for all $z \in \mathcal{Z}$ not contained in the results of $\mathbf{r}_i(x)$, Eq. (39) can be computed in $\mathcal{O}(|\mathcal{Y}||\mathbf{r}_1(x)||\mathbf{r}_2(x)|)$. Thus, the empirical sampling distribution can overall be computed in time $\mathcal{O}(|\mathcal{Y}||D| \max_x (|\mathbf{r}_1(x)||\mathbf{r}_2(x)|))$. $\square$

References

- Aslam, J., Pavlu, V., & Yilmaz, E. (2006). A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development on information retrieval*.
- Bach, F. (2007). Active learning for misspecified generalized linear models. In *Advances in neural information processing systems*.

- Beygelzimer, A., Dasgupta, S., & Langford, J. (2009). Importance weighted active learning. In *Proceedings of the international conference on machine learning*.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Burges, C. (2010). *RankNet to LambdaRank to LambdaMART: an overview* (Tech. Rep. MSR-TR-2010-82). Microsoft Research.
- Carterette, B., Allan, J., & Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. In *Proceedings of the 29th SIGIR conference on research and development in information retrieval*.
- Carterette, B., & Smucker, M. (2007). Hypothesis testing with incomplete relevance judgments. In *Proceedings of the 16th ACM conference on information and knowledge management*.
- Chapelle, O., & Chang, Y. (2011). Yahoo! Learning to rank challenge overview. In *JMLR: workshop and conference proceedings* (Vol. 14, pp. 1–24).
- Chapelle, O., Metzler, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceeding of the conference on information and knowledge management*.
- Cossock, D., & Zhang, T. (2008). Statistical analysis of Bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11), 5140–5154.
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4, 1–58.
- Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large margin rank boundaries for ordinal regression. In *Advances in large margin classifiers* (pp. 115–132).
- Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446.
- Li, P., Burges, C., & Wu, Q. (2007). Learning to rank using classification and gradient boosting. In *Advances in neural information processing systems*.
- Liu, J. (2001). *Monte Carlo strategies in scientific computing*. Berlin: Springer.
- Long, B., Chapelle, O., Zhang, Y., Chang, Y., Zheng, Z., & Tseng, B. (2010). Active learning for ranking through expected loss optimization. In *Proceedings of the 33rd SIGIR conference on research and development in information retrieval*.
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2), 109–142.
- Microsoft Research (2010). Microsoft learning to rank datasets. <http://research.microsoft.com/en-us/projects/mslr/>. Released 16 June 2010.
- Mohan, A., Chen, Z., & Weinberger, K. (2011). Web-search ranking with initialized gradient boosted regression trees. In *JMLR: workshop and conference proceedings* (Vol. 14, pp. 77–89).
- Radlinski, F., & Joachims, T. (2007). Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th SIGKDD international conference on knowledge discovery and data mining*.
- Sawade, C., Bickel, S., & Scheffer, T. (2010). Active risk estimation. In *Proceedings of the 27th international conference on machine learning*.
- Sawade, C., Landwehr, N., & Scheffer, T. (2012). Active comparison of prediction models. In *Advances in neural information processing systems*.
- Sawade, C., Bickel, S., von Oertzen, T., Scheffer, T., & Landwehr, N. (2012). Active evaluation of ranking functions based on graded relevance. In *Proceedings of the 2012 European conference on machine learning and principles and practice of knowledge discovery in databases*.
- Valizadegan, H., Jin, R., Zhang, R., & Mao, J. (2010). Learning to rank by optimizing NDCG measure. In *Advances in neural information processing systems*.
- Wasserman, L. (2004). *All of statistics: a concise course in statistical inference*. Berlin: Springer.
- Zheng, Z., Zha, H., Zhang, T., Chapelle, O., Chen, K., & Sun, G. (2007). A general boosting method and its application to learning ranking functions for web search. In *Advances in neural information processing systems*.