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

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

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:

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

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 ↔ 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

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

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

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

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

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

Proof of Theorem 6

Again there are two directions to prove: condition of Lemma 3 ↔ 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,

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

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

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

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 $\phi = \phi_{\text{sq}}$ in Section 3.2.1. For both ϕ_{cos} and ϕ_{list}, the distribution exhibiting inconsistency will be supported on two vectors

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

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

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

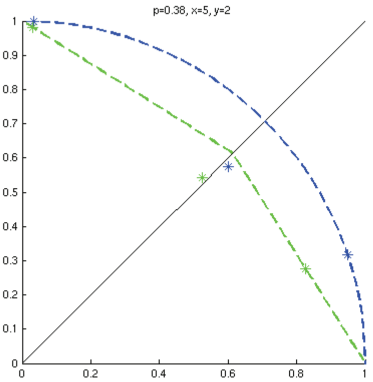


Figure 3: **Inconsistency of Cosine** The distribution is supported on $u = (1, x)$ and $v = (y, 1)$ with probability p and $1-p$ respectively. The 3 green points are $G(u)/\|G(u)\|_D$, $G(v)/\|G(v)\|_D$ and their weighted mean. The 3 blue points are $G(u)/\|G(u)\|_2$, $G(v)/\|G(v)\|_2$ and their weighted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

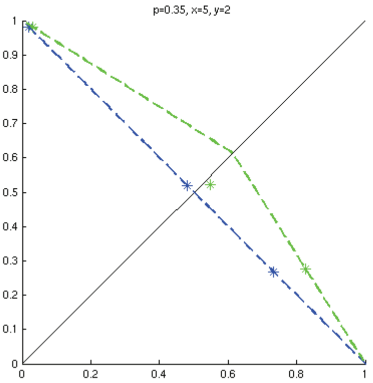


Figure 4: **Inconsistency of Cross Entropy** The distribution is supported on $u = (1, x)$ and $v = (y, 1)$ with probability p and $1-p$ respectively. The 3 green points are $G(u)/\|G(u)\|_D$, $G(v)/\|G(v)\|_D$ and their weighted mean. The 3 blue points are $\exp(u)/\mathbf{1}^\top \exp(u)$, $\exp(v)/\mathbf{1}^\top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

$$\begin{aligned}
\bar{\ell}_{\mathrm{NDCG}}(\mathbf{s};\eta) &= \mathbb{E}\left[-\frac{1}{\|G(\mathbf{r})\|_D}\sum_{j=1}^{m}\frac{G(r_j)}{F(\pi_{\mathbf{s}}(j))}\right] \\
&= \mathbb{E}\left[-\frac{1}{\|G(\mathbf{r})\|_D}\sum_{j=1}^{m}\frac{G(r_{\pi_{\mathbf{s}}^{-1}(j)})}{F(j)}\right] \qquad (11) \\
&= \mathbb{E}\left[-\sum_{j=1}^{m}\frac{(g(\mathbf{s}))_{\pi_{\mathbf{s}}^{-1}(j)}}{F(j)}\right] + T_1 \\
&\le \mathbb{E}\left[-\sum_{j=1}^{m}\frac{(g(\mathbf{s}))_{\pi^{-1}(j)}}{F(j)}\right] + T_1 \\
&= \mathbb{E}\left[-\frac{1}{\|G(\mathbf{r})\|_D}\sum_{j=1}^{m}\frac{G(r_{\pi^{-1}(j)})}{F(j)}\right] + T_2 + T_1 \\
&= \mathbb{E}\left[-\frac{1}{\|G(\mathbf{r})\|_D}\sum_{j=1}^{m}\frac{G(r_j)}{F(\pi(j))}\right] + T_2 + T_1 \\
&= \bar{\ell}_{\mathrm{NDCG}}(\pi;\eta) + T_2 + T_1 \ . \qquad (12)
\end{aligned}$$

where

$$\begin{aligned}
T_1 &:= \mathbb{E}\left[\sum_{j=1}^{m}\frac{1}{F(j)}\cdot\left((g(\mathbf{s}))_{\pi_{\mathbf{s}}^{-1}(j)} - \frac{G(r_{\pi_{\mathbf{s}}^{-1}(j)})}{\|G(\mathbf{r})\|_D}\right)\right] \ , \\
T_2 &:= \mathbb{E}\left[\sum_{j=1}^{m}\frac{1}{F(j)}\cdot\left(\frac{G(r_{\pi^{-1}(j)})}{\|G(\mathbf{r})\|_D} - (g(\mathbf{s}))_{\pi^{-1}(j)}\right)\right] \ .
\end{aligned}$$

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

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

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

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

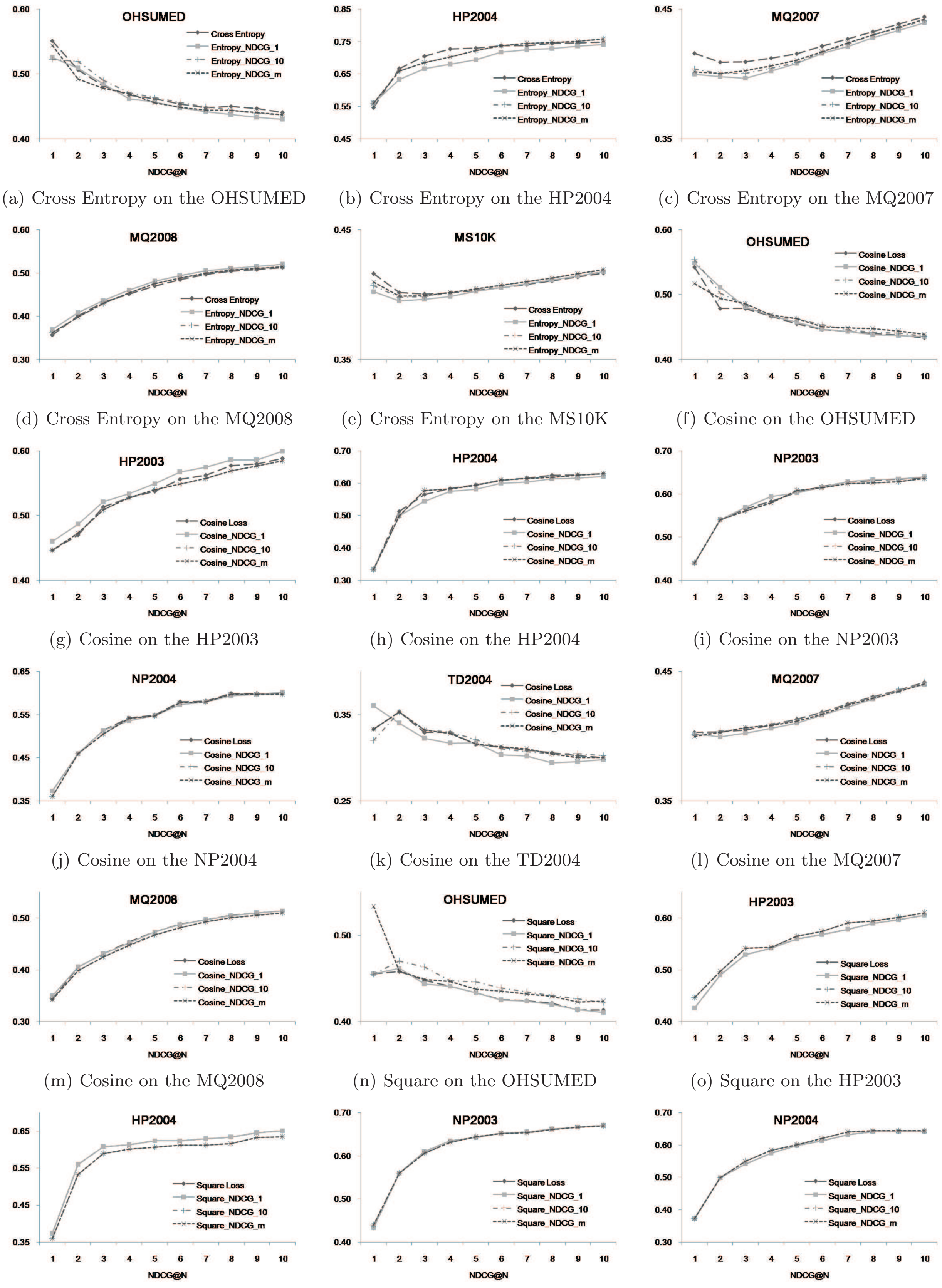
Taking maximum over π yields,

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

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

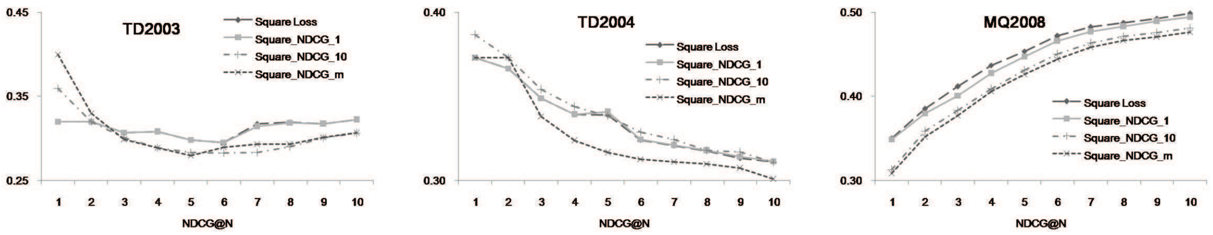
(d) Cross Entropy on the MQ2008 (e) Cross Entropy on the MS10K (f) Cosine on the OHSUMED

(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

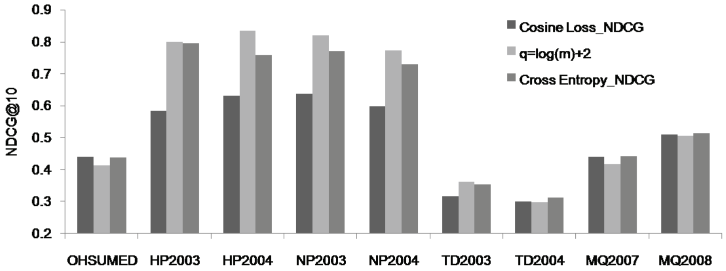


Figure 6: One example of normalized loss functions, $q = \log(m_i) + 2$ vs. existing listwise loss functions w/ recovering NDCG consistency

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%	