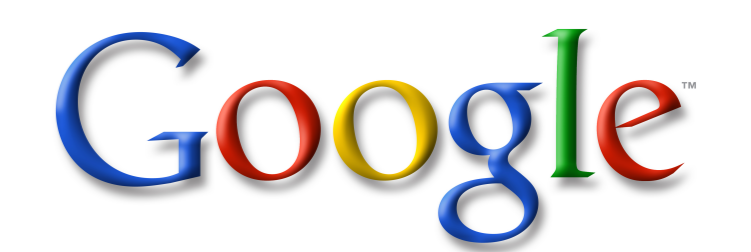


# Simple Risk Bounds for Position-Sensitive Max-Margin Ranking Algorithms



Stefan Riezler, Fabio De Bona

Google Research, Zürich

## Abstract

**R**ISK bounds for position-sensitive max-margin ranking algorithms can be derived straightforwardly from a structural result for Rademacher averages presented by [1]. We apply this result to pairwise and listwise hinge loss that are position-sensitive by virtue of rescaling the margin by a pairwise or listwise position-sensitive prediction loss. Similar bounds have recently been presented for probabilistic listwise ranking algorithms by [2]. More involved risk bounds for pairwise ranking algorithms have been presented before by [3] (using algorithmic stability), and for structured prediction by [4] (using PAC-Bayesian theory).

## 1. Notation

Notation	Meaning
$S = \{(x_q^{(i)}, y_q^{(i)})\}_{i=1}^n$	training sample
$x_q = \{x_{q1}, \dots, x_{q,n(q)}\}$	list of documents
$y_q = \{y_{q1}, \dots, y_{q,n(q)}\}$	ranking on docs
$\pi_q \in \Pi_q$	permutation of docs
$(i, j) \in \mathcal{P}_q$	pairs of docs
$\phi(x_{qi})$	feature function
$\phi(x_{qi}, \pi_q)$	partial order
$\frac{1}{\binom{n(q)}{2}} \sum_{(i,j) \in \mathcal{P}_q} \phi(x_{qi}) - \phi(x_{qj}) \operatorname{sgn}(\frac{1}{y_{qi}} - \frac{1}{y_{qj}})$	feature map
$\bar{f}(x_{qi}, x_{qj}, y_{qi}, y_{qj})$	ranking difference on doc level
$\langle w, \phi(x_{qi}, \pi_q) - \phi(x_{qj}, \pi_q) \rangle$	ranking difference on query level
$L(y_q, \pi_q) \in [0, 1]$	prediction loss

## 2. Position-Sensitive Max-Margin Ranking

**P**OSITION-sensitivity promotes high precision in the top ranks, corresponding to user studies in web search that show that users typically only look at the very top results returned by a search engine.

Position-sensitive pairwise max-margin learning accrues a penalty for misranking a pair of instances that is higher for misrankings involving higher rank positions than for misrankings in lower rank positions. Let  $m = n(q)$ ,  $(z)_+ = \max\{0, z\}$ , and  $[z] = 1$  if  $z$  is true, 0 otherwise:

**Definition 1** (Pairwise Hinge Loss).

$$\ell_P(\bar{f}; x_q, y_q) = \sum_{(i,j) \in \mathcal{P}_q} \left( \left| \frac{m}{y_{qi}} - \frac{m}{y_{qj}} \right| - \bar{f}(x_{qi}, x_{qj}, y_{qi}, y_{qj}) \right)_+$$

We use the pairwise 0-1 loss as basic loss function where  $\ell_{0-1}(\bar{f}; x_q, y_q) \leq \ell_P(\bar{f}; x_q, y_q)$  for all  $\bar{f}, x_q, y_q$ .

**Definition 2** (0-1 Loss).

$$\ell_{0-1}(\bar{f}; x_q, y_q) = \sum_{(i,j) \in \mathcal{P}_q} [\bar{f}(x_{qi}, x_{qj}, y_{qi}, y_{qj}) < 0].$$

Listwise max-margin algorithms are position-sensitive by virtue of position-sensitivity of the prediction loss  $L$ .

**Definition 3** (Listwise Hinge Loss).

$$\ell_L(\bar{f}; x_q, y_q) = \sum_{\pi_q \in \Pi_q \setminus y_q} (L(y_q, \pi_q) - \bar{f}(x_q, y_q, \pi_q))_+$$

The basic loss function for the listwise case is defined by the prediction loss  $L$  itself. For example, the prediction loss  $L_{AP}$  for AP on the query level is defined as follows with respect to binary rank labels  $y_{qj} \in \{1, 2\}$ :

**Definition 4** (AP Loss).

$$L_{AP}(y_q, \pi_q) = 1 - AP(y_q, \pi_q)$$

$$\text{where } AP(y_q, \pi_q) = \frac{\sum_{j=1}^{n(q)} \operatorname{Prec}(j) \cdot (|y_{qj} - 2|)}{\sum_{j=1}^{n(q)} (|y_{qj} - 2|)}$$

$$\text{and } \operatorname{Prec}(j) = \frac{\sum_{k: \pi_q(k) \leq \pi_q(j)} (|y_{qk} - 2|)}{\pi_q(j)}.$$

## 3. Risk Bounds and Structural Results using Rademacher Complexity

Assume the usual definitions of expected and empirical risk:

$$R_\ell(\bar{f}) = \int_Q \ell(\bar{f}; x_q, y_q) P(dx_q, dy_q).$$

$$\hat{R}_\ell(\bar{f}; S) = \frac{1}{n} \sum_{i=1}^n \ell(\bar{f}; x_q^{(i)}, y_q^{(i)}),$$

where  $S = \{(x_q^{(i)}, y_q^{(i)})\}_{i=1}^n$ .

[1]'s central theorem on risk bounds using Rademacher averages:

**Theorem 1** (cf. [1], Theorem 8). Assume loss functions  $\bar{\ell}(\bar{f}; x_q, y_q) \in [0, 1]$ ,  $\ell(\bar{f}; x_q, y_q) \in [0, 1]$  where  $\ell$  dominates  $\bar{\ell}$  s.t. for all  $\bar{f}, x_q, y_q$ ,  $\ell(\bar{f}; x_q, y_q) \leq \bar{\ell}(\bar{f}; x_q, y_q)$ . Let  $S = \{(x_q^{(i)}, y_q^{(i)})\}_{i=1}^n$  be a training set of i.i.d. instances, and  $\bar{\mathcal{F}}$  be the class of linear ranking-difference functions. Then with probability  $1 - \delta$  over samples of length  $n$ , the following holds for all  $\bar{f} \in \bar{\mathcal{F}}$ :

$$R_{\bar{\ell}}(\bar{f}) \leq \hat{R}_{\bar{\ell}}(\bar{f}; S) + \mathcal{R}_n(\ell \circ \bar{\mathcal{F}}) + \sqrt{\frac{8 \ln(2/\delta)}{n}}$$

where  $\mathcal{R}_n(\ell \circ \bar{\mathcal{F}}) = \mathbb{E}_\sigma \sup_{\bar{f} \in \bar{\mathcal{F}}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(\bar{f}; x_q^{(i)}, y_q^{(i)})$ .

Breaking down  $\mathcal{R}_n(\ell \circ \bar{\mathcal{F}})$  into a Rademacher average  $\mathcal{R}_n(\bar{\mathcal{F}})$  for the linear ranking models, and the Lipschitz constant  $L_\ell$  for the loss function  $\ell$ :

**Theorem 2** (cf. [1], Theorem 12). Let  $\ell$  be a Lipschitz continuous loss function with Lipschitz constant  $L_\ell$ , then for all  $\bar{f} \in \bar{\mathcal{F}}$ :

$$\mathcal{R}_n(\ell \circ \bar{\mathcal{F}}) \leq 2L_\ell \mathcal{R}_n(\bar{\mathcal{F}}).$$

Rademacher average for class of linear functions:

**Lemma 1** (cf. [1], Lemma 22). Let  $\bar{\mathcal{F}}$  be the class of linear ranking difference functions bounded by  $BM$ . Then for all  $\bar{f} \in \bar{\mathcal{F}}$ :

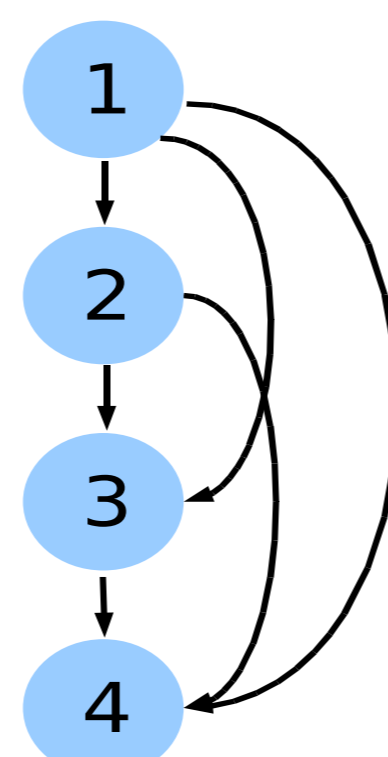
$$\mathcal{R}_n(\bar{\mathcal{F}}) = \frac{2BM}{\sqrt{n}}.$$

## 4. Risk Bound for Pairwise Hinge Loss Functions

**P**AIRWISE ranking of all pairs in a list of length  $m = n(q)$  involves  $\binom{m}{2}$  pairwise comparisons:

**Theorem 3.** Let  $\ell_{0-1}$  be the 0-1 loss defined in Definition (2) and  $\ell_P$  be the pairwise hinge loss defined in Definition (1) where for all  $\bar{f}, x_q, y_q$ ,  $\ell_{0-1}(\bar{f}; x_q, y_q) \leq \ell_P(\bar{f}; x_q, y_q)$ . Let  $S = \{(x_q^{(i)}, y_q^{(i)})\}_{i=1}^n$  be a training set of i.i.d. instances, and  $\bar{\mathcal{F}}$  be the class of linear ranking-difference functions where  $\|w\| \leq B$ ,  $\|\phi\| \leq M$ , and  $\|f\| \leq BM$  for all  $f \in \bar{\mathcal{F}}$ . Then with probability  $1 - \delta$  over samples of length  $n$ , the following holds for all  $\bar{f} \in \bar{\mathcal{F}}$ :

$$R_{\ell_{0-1}}(\bar{f}) \leq \hat{R}_{\ell_P}(\bar{f}; S) + \binom{m}{2} \frac{4BM}{\sqrt{n}} + \binom{m}{2} (m-1+2BM) \sqrt{\frac{8 \ln(2/\delta)}{n}}$$

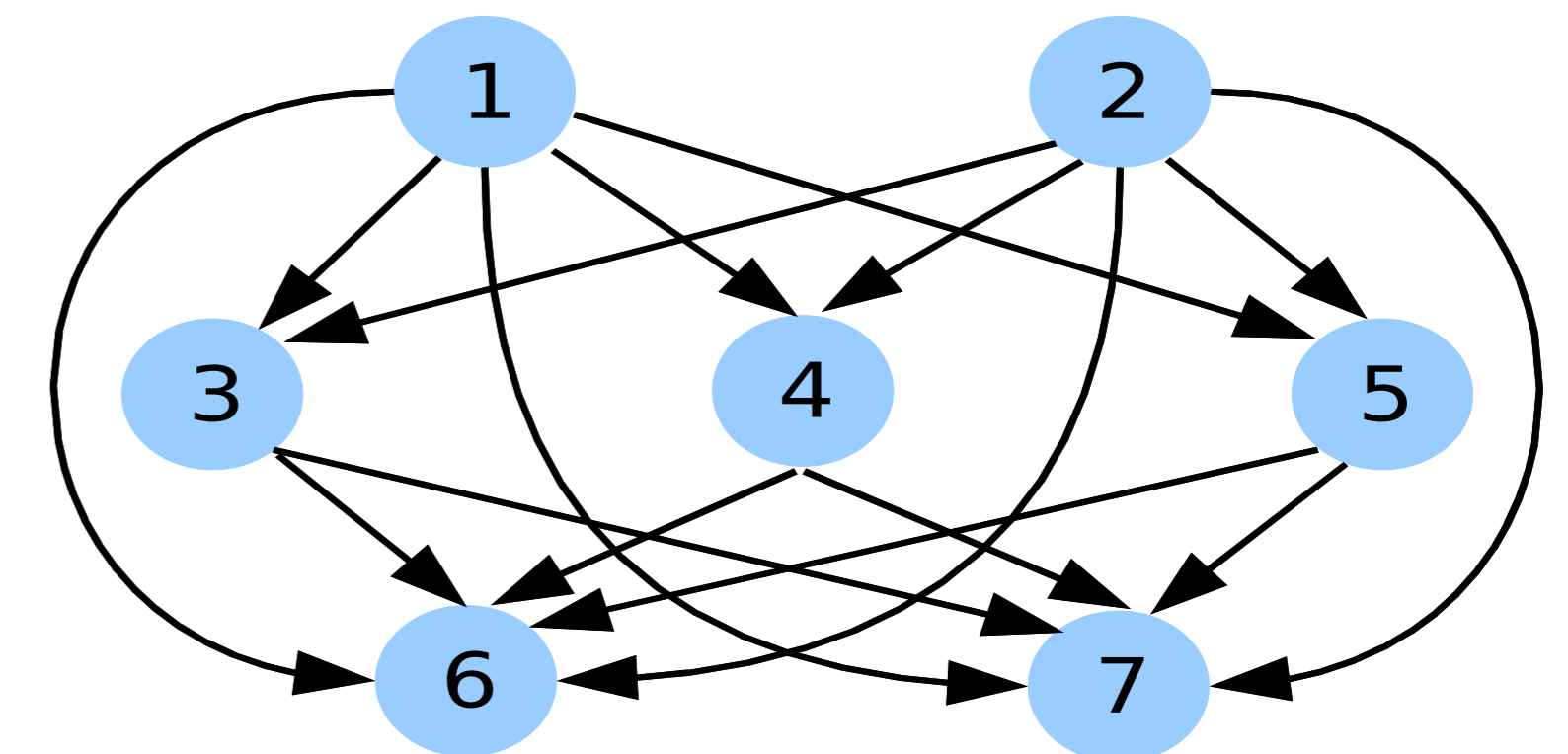


Multipartite ranking reduces the number of pairs  $|\mathcal{P}_q|$  from the set of all  $\binom{m}{2}$  pairwise comparisons to  $\sum_{i=1}^{r-1} \sum_{j=i+1}^r |l_i| |l_j|$  comparisons between documents at  $r$  relevance levels, including  $|l_i|$  documents each:

**Corollary 1.** Let  $\ell_{0-1}$  be the 0-1 loss and  $\ell_P$  be the pairwise hinge loss defined on a set of  $\sum_{i=1}^{r-1} \sum_{j=i+1}^r |l_i| |l_j|$  pairs over  $r$  relevance levels  $l_i$ . Let  $S = \{(x_q^{(i)}, y_q^{(i)})\}_{i=1}^n$  be a training set of i.i.d. instances, and  $\bar{\mathcal{F}}$  be the class of linear ranking-difference functions. Then with probability  $1 - \delta$  over samples of length  $n$ , the following holds for all  $\bar{f} \in \bar{\mathcal{F}}$ :

$$R_{\ell_{0-1}}(\bar{f}) \leq \hat{R}_{\ell_P}(\bar{f}; S) + \left( \sum_{i=1}^{r-1} \sum_{j=i+1}^r |l_i| |l_j| \right) \frac{4BM}{\sqrt{n}}$$

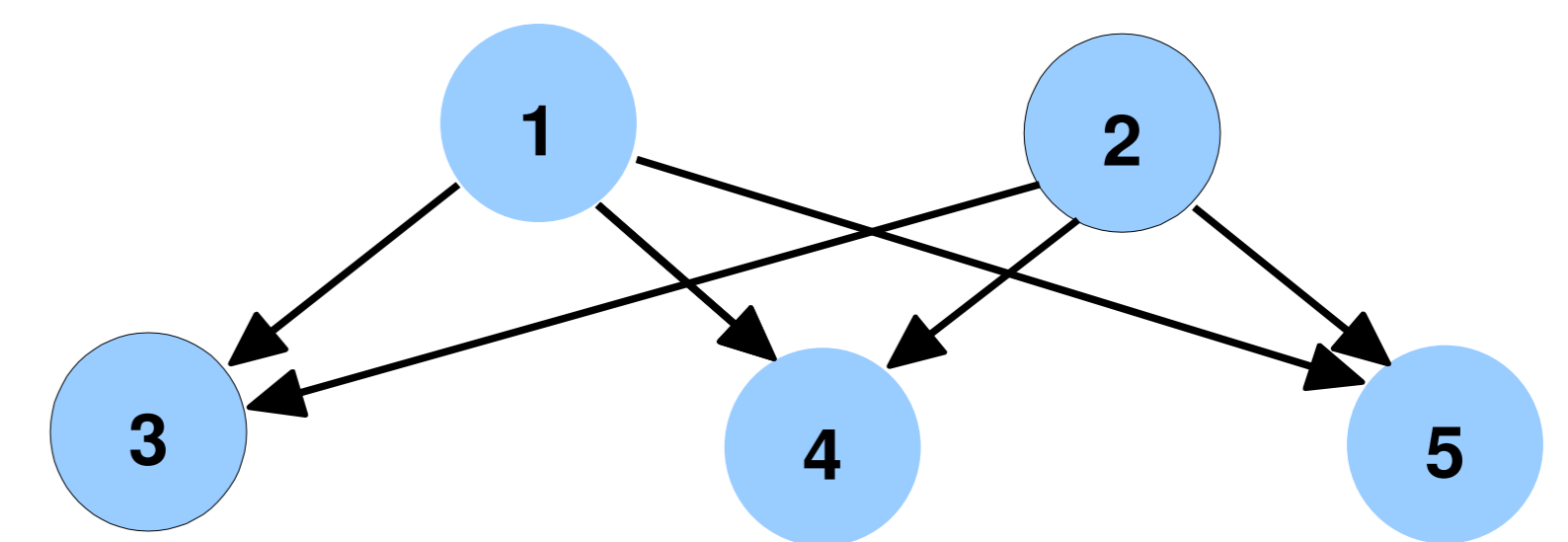
$$+ \left( \sum_{i=1}^{r-1} \sum_{j=i+1}^r |l_i| |l_j| \right) (r-1+2BM) \sqrt{\frac{8 \ln(2/\delta)}{n}}.$$



Bipartite ranking of  $g$  relevant and  $b$  non-relevant documents involves  $|\mathcal{P}_q| = g \cdot b$  pairs:

**Corollary 2.** Let  $\ell_{0-1}$  be the 0-1 loss and  $\ell_P$  be the pairwise hinge loss defined on a set of  $g \cdot b$  pairs for bipartite ranking of  $g$  relevant and  $b$  non-relevant documents. Let  $S = \{(x_q^{(i)}, y_q^{(i)})\}_{i=1}^n$  be a training set of i.i.d. instances, and  $\bar{\mathcal{F}}$  be the class of linear ranking-difference functions. Then with probability  $1 - \delta$  over samples of length  $n$ , the following holds for all  $\bar{f} \in \bar{\mathcal{F}}$ :

$$R_{\ell_{0-1}}(\bar{f}) \leq \hat{R}_{\ell_P}(\bar{f}; S) + (g \cdot b) \frac{4BM}{\sqrt{n}} + (g \cdot b) (1+2BM) \sqrt{\frac{8 \ln(2/\delta)}{n}}.$$



## 5. Risk Bound for Listwise Hinge Loss Functions

**L**ISTWISE ranking using prediction loss functions defined on lists of length  $m = n(q)$  involves  $m!$  comparisons of permutations:

**Theorem 4.** Let  $\ell_L$  be the listwise hinge loss defined in Definition (3). Let  $S = \{(x_q^{(i)}, y_q^{(i)})\}_{i=1}^n$  be a training set of i.i.d. instances, and  $\bar{\mathcal{F}}$  be the class of linear ranking-difference functions. Then with probability  $1 - \delta$  over samples of length  $n$ , the following holds for all  $\bar{f} \in \bar{\mathcal{F}}$ :

$$R_L \leq \hat{R}_{\ell_L}(\bar{f}; S) + m! \frac{4BM}{\sqrt{n}} + m! (1+2BM) \sqrt{\frac{8 \ln(2/\delta)}{n}}.$$

Specific prediction loss functions such as AP treat permutations among relevant documents or among non-relevant documents equally, and thus involve only  $|\Pi_q| = \frac{m!}{g!b!} = \binom{m}{g} = \binom{m}{b}$  permutations, where  $g$  and  $b$  are the number of relevant and non-relevant documents, respectively.

**Corollary 3.** Let  $L_{AP}$  be the AP loss defined Definition 4 and  $\ell_{LAP}$  be the listwise hinge loss using  $L_{AP}$  as prediction loss function. Let  $S = \{(x_q^{(i)}, y_q^{(i)})\}_{i=1}^n$  be a training set of i.i.d. instances, and  $\bar{\mathcal{F}}$  be the class of linear ranking-difference functions. Then with probability  $1 - \delta$  over samples of length  $n$ , the following holds for all  $\bar{f} \in \bar{\mathcal{F}}$ :

$$R_{LAP} \leq \hat{R}_{\ell_{LAP}}(\bar{f}; S) + \binom{m}{g} \frac{4BM}{\sqrt{n}} + \binom{m}{g} (1+2BM) \sqrt{\frac{8 \ln(2/\delta)}{n}}.$$

## References

- [1] Peter L. Bartlett and Sahar Mendelson. Rademacher and Gaussian complexity: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
- [2] Yanyan Lan, Tie-Yan Liu, Zhiming Ma, and Hang Li. Generalization analysis of listwise learning-to-rank algorithms. In *Proceedings of the 26th International Conference on Machine Learning (ICML'09)*, Montreal, Canada, 2009.
- [3] Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474, 2009.
- [4] David McAllester. Generalization bounds and consistency for structured labeling. In Gökhan Bakhtir, Thomas Hofmann, and Bernhard Schölkopf, editors, *Prediction Structured Data*. The MIT Press, Cambridge, MA, 2007.