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

Stefan Riezler
Google Research
Brandschenkestrasse 110
8002 Zürich, Switzerland
riezler@google.com

Fabio De Bona*
Friedrich Miescher Laboratory
of the Max Planck Society
Spemannstrasse 39
72076 Tübingen, Germany
fabio@tuebingen.mpg.de

Abstract

We present risk bounds for position-sensitive max-margin ranking algorithms that follow 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.

1 Introduction

[2] recently presented risk bounds for probabilistic listwise ranking algorithms. The presented bounds follow straightforwardly from structural results for Rademacher averages presented by [1]. These bounds are dominated by two terms: Firstly, by the empirical Rademacher average $\mathcal{R}_n(\mathcal{F})$ of the class of ranking functions $\mathcal{F}$; secondly, by a term involving the Lipschitz constant of a Lipschitz continuous loss function. For example, for a loss function defined on the space of all possible permutations over m ranks, the Lipschitz constant involves a factor $m!$. Loss functions defined over smaller spaces involve smaller factors.

Similar risk bounds can be given for max-margin ranking algorithms based on the hinge-loss function. The bounds make use of a single structural result on Rademacher averages that reflects the structure of the output space in the Lipschitz constant of the hinge-loss function. We apply the result to pairwise and listwise hinge loss functions that are both position-sensitive by virtue of rescaling the margin by a pairwise or listwise position-sensitive prediction loss. Position-sensitivity means that high precision in the top ranks is promoted, corresponding to user studies in web search that show that users typically only look at the very top results returned by the search engine [3].

The contribution of this paper is to show how simple risk bounds can be derived for max-margin ranking algorithms by a straightforward application of structural results for Rademacher averages presented by [1]. More involved risk bounds for pairwise ranking algorithms have been presented before by [4] (using algorithmic stability), and for structured prediction by [5] (using PAC-Bayesian theory).

2 Notation

Let $S = \{(x_q, y_q)\}_{q=1}^n$ be a training sample of queries, each represented by a set of documents $x_q = \{x_{q1}, \ldots, x_{q,n(q)}\}$, and set of rank labels $y_q = \{y_{q1}, \ldots, y_{q,n(q)}\}$, where $n(q)$ is the number of documents for query q. For full rankings of all documents for a query, a total order on documents

*The work presented in this paper was done while the author was visiting Google Research, Zürich.

is assumed, with rank labels taking on values $y_{qi} \in \{1, \ldots, n(q)\}$. Documents of equivalent rank can be specified by assuming a partial order on documents, where a multipartite ranking involves $r < n(q)$ relevance levels such that $y_{qi} \in \{1, \ldots, r\}$, and a bipartite ranking involves two rank values $y_{qi} \in \{1, 2\}$ with relevant documents at rank 1 and non-relevant documents at rank 2.

Let the documents in x_q be identified by the integers $\{1, 2, \ldots, n(q)\}$. Then a permutation π_q on x_q can be defined as a bijection from $\{1, 2, \ldots, n(q)\}$ onto itself. We use Π_q to denote the set of all possible permutations on x_q, and π_{qj} to denote the rank position of document x_{qj}. Furthermore, let (i, j) denote a pair of documents in x_q and let $\mathcal{P}_q$ be the set of all pairs in x_q.

A feature function $\phi(x_{qi})$ is associated with each document $i = 1, \ldots, n(q)$ for each query q. Furthermore, a partial-order feature map as used in [6, 7] is created for each document set as follows:

$$\phi(x_q, \pi_q) = \frac{1}{n(q)(n(q)-1)/2} \sum_{(i,j)\in\mathcal{P}_q} \phi(x_{qi}) - \phi(x_{qj}) \text{sgn}(\frac{1}{\pi_{qi}} - \frac{1}{\pi_{qj}}).$$

We assume linear ranking functions $f \in \mathcal{F}$ that are defined on the document level as $f(x_{qi}) = \langle w, \phi(x_{qi})\rangle$ and on the query level as $f(x_q, \pi_q) = \langle w, \phi(x_q, \pi_q)\rangle$. Note that since feature vectors on document and query level have the same size, assuming that $||w|| \leq B$, $||\phi|| \leq M$, we get $||f|| \leq BM$ for all $f \in \mathcal{F}$.

The goal of learning a ranking over the documents x_q for a query q can be achieved either by sorting the documents according to the document-level ranking function $f(x_{qi}) = \langle w, \phi(x_{qi})\rangle$, or by finding the permutation π^* that scores highest according to the query-level ranking function:

$$\pi^* = \arg\max_{\pi_q\in\Pi_q} f(x_q, \pi_q) = \arg\max_{\pi_q\in\Pi_q} \langle w, \phi(x_q, \pi_q)\rangle .$$

For convenience, let us furthermore define ranking-difference functions on the document level

$$\bar{f}(x_{qi}, x_{qj}, y_{qi}, y_{qj}) = \langle w, \phi(x_{qi}) - \phi(x_{qj})\rangle \text{ sgn}(\frac{1}{y_{qi}} - \frac{1}{y_{qj}}),$$

and on the query level

$$\bar{f}(x_q, y_q, \pi_q) = \langle w, \phi(x_q, y_q) - \phi(x_q, \pi_q)\rangle .$$

Finally, let $L(y_q, \pi_q) \in [0, 1]$ denote a prediction loss of a predicted ranking π_q compared to the ground-truth ranking y_q. [1]

3 Position-Sensitive Max-Margin Ranking Algorithms

A position-sensitive pairwise max-margin algorithm can be given by extending the magnitude-preserving pairwise hinge-loss of [4] or [8]. For a fully ranked list of instances as gold standard, the penalty term can be made position-sensitive by accruing the magnitude of the difference of inverted ranks instead of the magnitude of score differences. Thus the penalty for misranking a pair of instances is higher for misrankings involving higher rank positions than for misrankings in lower rank positions. The pairwise hinge loss is defined as follows (where $(z)_+ = \max\{0, z\}$):

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

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

We use the pairwise 0-1 error ℓ_{0-1} as basic ranking loss function for the pairwise case. Clearly, $\ell_{0-1}(\bar{f}; x_q, y_q) \leq \ell_P(\bar{f}; x_q, y_q)$ for all $\bar{f}, x_q, y_q$. The 0-1 error is defined as follows (where $[\![z]\!] = 1$ if z is true, 0 otherwise):

[1] We slightly abuse the notation y_q to denote the permutation on x_q that is induced by the rank labels. In case of full rankings, the permutation π_q corresponding to ranking y_q is unique. For multipartite and bipartite rankings, there is more than one possible permutation for a given ranking, so that we let π_q denote a permutation that is consistent with ranking 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 for the prediction loss of (Mean) Average Precision (AP) [9] and NDCG [10] have been presented by [6] and [7], respectively. These ranking algorithms are position-sensitive by virtue of position-sensitivity of the deployed prediction loss L. The listwise hinge loss for general L is defined as follows:

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)**.**

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

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

$$\textit{and } Prec(j) = \frac{\sum_{k:\pi_{qk}\leq\pi_{qj}} (|y_{qk} - 2|)}{\pi_{qj}}.$$

4 Risk Bounds

We use the usual definitions of expected and empirical risk with respect to a loss function ℓ. The expected risk is defined with respect to an unknown probability distribution $P_{\mathcal{Q}}$ where we regard pairs of documents and ranks (x, y) as random variables on the space $\mathcal{Q}$.

$$R_\ell(\bar{f}) \quad = \quad \int_{\mathcal{Q}} \ell(\bar{f}; x, y) P_{\mathcal{Q}}(dx, dy).$$

The empirical risk is defined with respect to a sample $S = \{(x_q, y_q)\}_{q=1}^n$:

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

[1]'s central theorem on risk bounds using Rademacher averages can be restated with respect the definitions above as follows:

Theorem 1 (cf. [1], Theorem 8)**.** *Assume loss functions $\tilde{\ell}(\bar{f}; x_q, y_q) \in [0, 1]$, $\ell(\bar{f}; x_q, y_q) \in [0, 1]$ where ℓ dominates $\tilde{\ell}$ s.t. for all $\bar{f}, x_q, y_q$, $\tilde{\ell}(\bar{f}; x_q, y_q) \leq \ell(\bar{f}; x_q, y_q)$. Let $S = \{(x_q, y_q)\}_{q=1}^n$ 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 $\bar{f} \in \bar{\mathcal{F}}$:*

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

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

The complexity measure of a Rademacher average $\mathcal{R}_n(F)$ on a class of functions F quantifies the extent to which some function in F can be correlated with a random noise sequence of length n. Here the Rademacher average $\mathcal{R}_n(\ell \circ \bar{\mathcal{F}})$ is defined on a class of functions that is composed of a Lipschitz continuous loss function ℓ and a linear ranking model in $\bar{\mathcal{F}}$. It can be broken down into a Rademacher average $\mathcal{R}_n(\bar{\mathcal{F}})$ for the linear ranking models, and the Lipschitz constant L_ℓ for the loss function ℓ. The following theorem makes use of the Ledoux-Talagrand concentration inequality:

Theorem 2 (cf. [1], Theorem 12)**.** *Let ℓ be a Lipschitz continuous loss function with Lipschitz constant L_ℓ, then $\mathcal{R}_n(\ell \circ \bar{\mathcal{F}}) \leq 2L_\ell \mathcal{R}_n(\bar{\mathcal{F}})$.*

Furthermore, the Rademacher average for linear functions is given by the following Lemma:

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

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

In order to apply Theorem 1, we need to normalize loss functions to map to $[0, 1]$. For full pairwise ranking, the size of the set of pairs over $m = n(q)$ ranks is $|\mathcal{P}_q| = \binom{m}{2}$. This yields a normalization constant $Z_P = \binom{m}{2}(m - 1 + 2BM)$ for pairwise hinge loss.

An application of Theorem 2 to pairwise hinge loss yields the following:

Proposition 1. *Let $\hat{\ell}_P = \frac{1}{Z_P}\ell_P$ be the normalized pairwise hinge loss. Then for all $\bar{f} \in \bar{\mathcal{F}}$:*

$$\mathcal{R}_n(\hat{\ell}_P \circ \bar{\mathcal{F}}) \leq \frac{2}{m - 1 + 2BM}\mathcal{R}_n(\bar{\mathcal{F}}).$$

Proof. Follows directly from Theorem (2) with $L_{\hat{\ell}_P} = \sup_{\bar{f}} |\hat{\ell}'_P(\bar{f})| = |\frac{\binom{m}{2}}{\binom{m}{2}(m-1+2BM)|}|$. □

Using the 0-1 loss as dominated loss, we can combine Theorem 1 and Lemma 1 with Proposition 1 to get the following result:

Theorem 3. *Let ℓ_{0-1} be the 0-1 loss defined in Definition (2) and ℓ_P be the pairwise hinge loss defined in Definition (1). Let $S = \{(x_q, y_q)\}_{q=1}^n$ 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 $\bar{f} \in \bar{\mathcal{F}}$:*

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

Proof. Combining Theorem (1) and Proposition (1) under the use of normalized loss function $\hat{\ell}_P = \frac{1}{Z_P}\ell_P$, we get

$$R_{\hat{\ell}_P}(\bar{f}) \leq \hat{R}_{\hat{\ell}_P}(\bar{f}; S) + \frac{2}{m - 1 + 2BM}\frac{2BM}{\sqrt{n}} + \sqrt{\frac{8\ln(2/\delta)}{n}}.$$

Since for $c, F, G > 0$, the inequality $F \leq G$ implies $cF \leq cG$, we can rescale the result above to achieve a bound for the original loss functions.

$$Z_P[R_{\hat{\ell}_P}(\bar{f})] \leq Z_P[\hat{R}_{\hat{\ell}_P}(\bar{f}; S) + \frac{2}{m - 1 + 2BM}\frac{2BM}{\sqrt{n}} + \sqrt{\frac{8\ln(2/\delta)}{n}}].$$

Multiplying in the normalization constant gives

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

Finally, we can bound $R_{\ell_P}(\bar{f})$ by $R_{\ell_{0-1}}(\bar{f})$ from below since $R_{\ell_{0-1}}(\bar{f}) \leq R_{\ell_P}(\bar{f})$ follows from $\ell_{0-1} \leq \ell_P$, concluding the proof. □

Interestingly, the structure of the output space is directly reflected in the risk bounds. For full pairwise ranking over all possible pairs, a penalty of $\binom{m}{2}$ has to be paid for the exploration of the full space of all pairwise comparisons. For the case of pairwise ranking of documents at r relevance levels, including $|l_i|$ documents each, pairwise comparisons between documents at the same relevance level can be ignored. Thus, in this scenario of multipartite ranking, the number of pairs $|\mathcal{P}_q|$ is reduced from the set of all $\binom{m}{2}$ pairwise comparisons to $\sum_{i=1}^{r-1}\sum_{j=i+1}^{r}|l_i||l_j|$. A risk bound for this scenario is given by the following corollary:

Corollary 1. *Let ℓ_{0-1} be the 0-1 loss and ℓ_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 $S = \{(x_q, y_q)\}_{q=1}^n$ 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 $\bar{f} \in \bar{\mathcal{F}}$:*

$$R_{\ell_{0-1}}(\bar{f}) \leq \hat{R}_{\ell_P}(\bar{f}; S) + (\sum_{i=1}^{r-1}\sum_{j=i+1}^{r}|l_i||l_j|)\frac{4BM}{\sqrt{n}} + (\sum_{i=1}^{r-1}\sum_{j=i+1}^{r}|l_i||l_j|)(r-1+2BM)\sqrt{\frac{8\ln(2/\delta)}{n}}.$$

Bipartite ranking of rel relevant and $nrel$ non-relevant documents involves even fewer pairs, namely $|\mathcal{P}_q| = rel \cdot nrel$. A risk bound for bipartite ranking can be given as follows:

Corollary 2. *Let ℓ_{0-1} be the 0-1 loss and ℓ_P be the pairwise hinge loss defined on a set of $rel \cdot nrel$ pairs for bipartite ranking of rel relevant and $nrel$ non-relevant documents. Let $S = \{(x_q, y_q)\}_{q=1}^n$ 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 $\bar{f} \in \bar{\mathcal{F}}$:*

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

For the general case of listwise hinge loss, we get the following, using a normalization constant $Z_L = m!(1+2BM)$ for a number of $|\Pi_q| = m!$ permutations over $m = n(q)$ ranks:

Proposition 2. *Let $\hat{\ell}_L = \frac{1}{Z_L}\ell_L$ be the normalized listwise hinge loss. Then for all $\bar{f} \in \bar{\mathcal{F}}$:*

$$\mathcal{R}_n(\hat{\ell}_L \circ \bar{\mathcal{F}}) \leq \frac{2}{1+2BM}\mathcal{R}_n(\bar{\mathcal{F}}).$$

Proof. Follows directly from Theorem (2) with $L_{\hat{\ell}_L} = \sup_{\bar{f}}|\hat{\ell}'_L(\bar{f})| = |\frac{m!}{m!(1+2BM)}|$. □

A risk bound for listwise prediction loss in the general case can be given as follows.

Theorem 4. *Let ℓ_L be the listwise hinge loss defined in Definition (3). Let $S = \{(x_q, y_q)\}_{q=1}^n$ 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 $\bar{f} \in \bar{\mathcal{F}}$:*

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

Proof. Similar to the proof for (3) using the fact that the hinge loss ℓ_L bounds the prediction loss L from above (see [11], Proposition 2), where $R_L = \int_Q L(y_q, \pi_q)P(dy_q, d\pi_q)$. □

Specific prediction loss functions such as AP define a specific structure on the output space which is reflected in the risk bound for structured prediction for AP loss. For AP, permutations that involve only reorderings of relevant documents with relevant documents, or reorderings of irrelevant documents with irrelevant documents, are considered equal. This means that instead of $m!$ permutations over a list of size $m = n(q)$, the number of permutations is $|\Pi_q| = \frac{m!}{rel!nrel!} = \binom{m}{rel} = \binom{m}{nrel}$, where rel and $nrel$ are the number of relevant and irrelevant documents. A risk bound for listwise ranking using AP loss can be given as follows:

Corollary 3. *Let L_{AP} be the AP loss defined Definition 4 and $\ell_{L_{AP}}$ be the listwise hinge loss using L_{AP} as prediction loss function. Let $S = \{(x_q, y_q)\}_{q=1}^n$ 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 $\bar{f} \in \bar{\mathcal{F}}$:*

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

5 Discussion

The bounds we presented were given for algorithms that compute the hinge loss by summing over all possible outputs instead of taking the arg-max to find the most violated constraint. Since $\sum x_i \geq \max_i x_i$, for all $x_i \geq 0$, the bounds still apply to approaches that take the arg-max. On the other hand, they also apply to approaches where successively adding most violated constraints is infeasible [12]. Tighter bounds may be given for arg-max and soft-max versions. This is due to future work. Furthermore, the proofs need to be extended to other listwise loss functions such as NDCG. Lastly, an empirical validation supporting the theoretical results needs to be given.

Acknowledgements

We would like to thank Olivier Bousquet for several discussions of the work presented in this paper.

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] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2), 2007.

[4] Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474, 2009.

[5] David McAllester. Generalization bounds and consistency for structured labeling. In Gökhan Bakhir, Thomas Hofmann, and Berhard Schölkopf, editors, *Prediction Structured Data*. The MIT Press, Cambridge, MA, 2007.

[6] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. A support vector method for optimizing average precision. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, Amsterdam, The Netherlands, 2007.

[7] Soumen Chakrabarti, Rajiv Khanna, Uma Sawant, and Chiru Bhattacharayya. Structured learning for non-smooth ranking losses. In *Proceedings of the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'08)*, Las Vegas, NV, 2008.

[8] Corinna Cortes, Mehryar Mohri, and Asish Rastogi. Magnitude-preserving ranking algorithms. In *Proceedings of the 24th International Conference on Machine Learning (ICML'07)*, Corvallis, OR, 2007.

[9] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, NY, 1999.

[10] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions in Information Systems*, 20(4):422–446, 2002.

[11] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 5:1453–1484, 2005.

[12] Thomas Gärtner and Shankar Vembu. On structured output training: hard cases and an efficient alternative. *Jounal of Machine Learning Research*, 76:227–242, 2009.