

ConeRANK: Ranking as Learning Generalized Inequalities

Truyen T. Tran[†] and Duc-Son Pham[‡]

[†] Center for Pattern Recognition and Data Analytics (PRaDA),
Deakin University, Geelong, VIC, Australia

[‡] Department of Computing, Curtin University, Western Australia
Email: truyen@vietlabs.com dspham@ieee.org

June 20, 2012

Abstract

We propose a new data mining approach in ranking documents based on the concept of cone-based generalized inequalities between vectors. A partial ordering between two vectors is made with respect to a proper cone and thus learning the preferences is formulated as learning proper cones. A pairwise learning-to-rank algorithm (ConeRank) is proposed to learn a non-negative subspace, formulated as a polyhedral cone, over document-pair differences. The algorithm is regularized by controlling the ‘volume’ of the cone. The experimental studies on the latest and largest ranking dataset LETOR 4.0 shows that ConeRank is competitive against other recent ranking approaches.

1 Introduction

Learning to rank in information retrieval (IR) is an emerging subject [7, 11, 9, 4, 5] with great promise to improve the retrieval results by applying machine learning techniques to learn the document relevance with respect to a query. Typically, the user submits a query and the system returns a list of related documents. We would like to learn a ranking function that outputs the position of each returned document in the decreasing order of relevance.

Generally, the problem can be studied in the supervised learning setting, in that for each query-document pair, there is an extracted feature vector and a position label in the ranking. The feature can be either *query-specific* (e.g. the number of matched keywords in the document title) or *query-independent* (e.g. the PageRank score of the document, number of in-links and out-links, document length, or the URL domain). In training data, we have a groundtruth ranking per query, which can be in the form of a relevance score assigned to each document, or an ordered list in decreasing level of relevance.

The learning-to-rank problem has been approached from different angles, either treating the ranking problem as ordinal regression [10, 6], in which an ordinal label is assigned to a document, as pairwise preference classification [11, 9, 4] or as a listwise permutation problem [14, 5].

We focus on the pairwise approach, in that ordered pairs of document per query will be treated as training instances, and in testing, predicted pairwise orders within a query will be combined to make a final ranking. The advantage of this approach is that many existing powerful binary classifiers that can be adapted with minimal changes - SVM [11], boosting [9], or logistic regression [4] are some choices.

We introduce an entirely new perspective based on the concept of cone-based *generalized inequality*. More specifically, the inequality between two multidimensional vectors is defined with respect to a cone. Recall that a cone is a geometrical object in that if two vectors belong to the cone, then any non-negative linear combination of the two vectors also belongs to the cone. Translated into the framework of our problem, this means that given a cone \mathcal{K} , when document l is ranked higher than document m , the feature vector \mathbf{x}_l is ‘greater’ than the feature vector \mathbf{x}_m with respect to \mathcal{K} if $\mathbf{x}_l - \mathbf{x}_m \in \mathcal{K}$. Thus, given a cone, we can find the correct order of preference for any given document pair. However, since the cone \mathcal{K} is not known in advance, it needs to be estimated from the data. Thus, in our paper, we consider polyhedral cones constructed from basis vectors and propose a method for learning the cones via the estimation of this set of basis vectors.

This paper makes the following contributions:

- A novel formulation of the learning to rank problem, termed as ConeRank, from the angle of cone learning and generalized inequalities;
- A study on the generalization bounds of the proposed method;
- Efficient online cone learning algorithms, scalable with large datasets; and,
- An evaluation of the algorithms on the latest LETOR 4.0 benchmark dataset ¹.

2 Previous Work

Learning-to-rank is an active topic in machine learning, although ranking and permutations have been studied widely in statistics. One of the earliest paper in machine learning is perhaps [7]. The seminal paper [11] stimulates much subsequent research. Machine learning methods extended to ranking can be divided into:

Pointwise approaches, that include methods such as ordinal regression [10, 6]. Each query-document pair is assigned a ordinal label, e.g. from the set $\{0, 1, 2, \dots, L\}$. This simplifies the problem as we do not need to worry about the exponential number of permutations. The complexity is therefore linear in the number of query-document pairs. The drawback is that the ordering relation between documents is not explicitly modelled.

Pairwise approaches, that span preference to binary classification [11, 9, 4] methods, where the goal is to learn a classifier that can separate two documents (per query). This casts the ranking problem into a standard classification framework, wherein many algorithms are readily available. The complexity is quadratic in number of documents per query and linear in number of queries.

Listwise approaches, modelling the distribution of permutations [5]. The ultimate goal is to model a full distribution of all permutations, and the prediction phase outputs the most probable permutation. In the statistics community, this problem has been long addressed [14], from a different angle. The main difficulty is that the number of permutations is exponential and thus approximate inference is often used.

However, in IR, often the evaluation criteria is different from those employed in learning. So there is a trend to optimize the (approximate or bound) IR metrics [8].

¹Available at: <http://research.microsoft.com/en-us/um/beijing/projects/letor/letor4dataset.aspx>

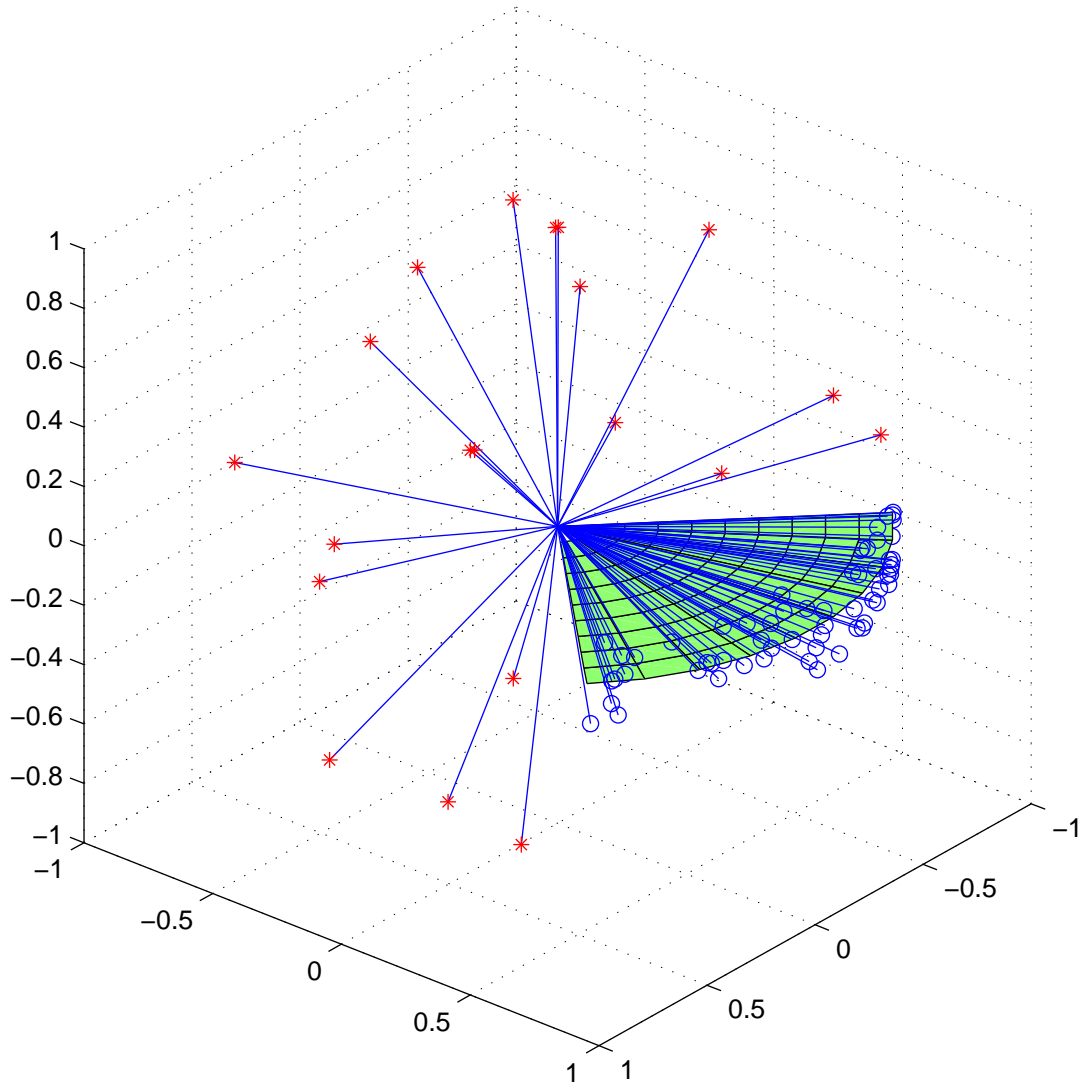


Figure 1: Illustration of ConeRank. Here the pairwise differences are distributed in 3-dimensional space, most of which however lie only on a surface and can be captured most effectively by a ‘minimum’ cone plotted in green. Red stars denotes noisy samples.

3 Proposed Method

3.1 Problem Settings

We consider a training set of P queries q_1, q_2, \dots, q_P randomly sampled from a query space \mathcal{Q} according to some distribution $P_{\mathcal{Q}}$. Associated with each query q is a set of documents represented as pre-processed feature vectors $\{\mathbf{x}_1^q, \mathbf{x}_2^q \dots\}, \mathbf{x}_l^q \in \mathbb{R}^N$ with relevance scores r_1^q, r_2^q, \dots from which ranking over documents can be based. We note that the values of the feature vectors may be query-specific and thus the same document can have different feature vectors according to different queries. Document \mathbf{x}_l^q is said to be more preferred than document \mathbf{x}_m^q for a given query q if $r_l^q > r_m^q$ and vice versa. In the *pairwise* approach, pursued in this paper, equivalently we learn a ranking function f that takes input as a pair of different documents $\mathbf{x}_l^q, \mathbf{x}_m^q$ for a given query q and returns a value $y \in \{+1, -1\}$ where $+1$ corresponds to the case where \mathbf{x}_l^q is ranked above \mathbf{x}_m^q and vice versa. For notational simplicity, we may drop the superscript q where there is no confusion.

3.2 Ranking as Learning Generalized Inequalities

In this work, we consider the ranking problem from the viewpoint of generalized inequalities. In convex optimization theory [3, p.34], a generalized inequality $\succ_{\mathcal{K}}$ denotes a partial ordering induced by a proper cone \mathcal{K} , which is convex, closed, solid, and pointed:

$$\mathbf{x}_l \succ_{\mathcal{K}} \mathbf{x}_m \iff \mathbf{x}_l - \mathbf{x}_m \in \mathcal{K}.$$

Generalized inequalities satisfy many properties such as preservation under addition, transitivity, preservation under non-negative scaling, reflexivity, anti-symmetry, and preservation under limit.

We propose to learn a generalized inequality or, equivalently, a proper cone \mathcal{K} that best describes the training data (see Fig. 1 for an illustration). Our important assumption is that this proper cone, which induces the generalized inequality, is not query-specific and thus prediction can be used for unseen queries and document pairs coming from the same distributions.

From a fundamental property of convex cones, if $\mathbf{z} \in \mathcal{K}$ then $w\mathbf{z} \in \mathcal{K}$ for all $w > 0$, and any non-negative combination of the cone elements also belongs to the cone, i.e. if $\mathbf{u}_k \in \mathcal{K}$ then $\sum_k w_k \mathbf{u}_k \in \mathcal{K}, \forall w_k > 0$.

In this work, we restrict our attention to *polyhedral* cones for the learning of generalized inequalities. A polyhedral cone is a polyhedron and a cone. A polyhedral cone can be defined as sum of rays or intersection of halfspaces. We construct the polyhedral cone \mathcal{K} from ‘basis’ vectors $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K]$. They are the extreme vectors lying on the intersection of hyperplanes that define the halfspaces. Thus, the cone \mathcal{K} is a conic hull of the basis vectors and is completely specified if the basis vectors are known. A polyhedral cone with K basis vectors is said to have an order K if one basis vector cannot be expressed as a conic combination of the others. It can be verified that under these regular conditions, a polyhedral cone is a proper cone and thus can induce a generalized inequality. We thus propose to learn the basis vectors $\mathbf{u}_k, k = 1, \dots, K$ for the characterization of \mathcal{K} .

A projection of \mathbf{z} onto the cone \mathcal{K} , denoted by $P_{\mathcal{K}}(\mathbf{z})$, is generally defined as some $\mathbf{z}' \in \mathcal{K}$ such that a certain criterion on the distance between \mathbf{z} and \mathbf{z}' is met. As $\mathbf{z}' \in \mathcal{K}$, it follows that it admits a conic representation $\mathbf{z}' = \sum_{k=1}^K w_k \mathbf{u}_k = \mathbf{U}\mathbf{w}$, $w_k \geq 0$. By restricting the order $K \leq N$, it can be shown that when \mathbf{U} is full-rank then the conic representation is unique.

Define an ordered document-pair (l, m) difference as $\mathbf{z} = \mathbf{x}_l - \mathbf{x}_m$ where, without loss of generality, we assume that $r_l \geq r_m$. The linear representation of $\mathbf{z}' \in \mathcal{K}$ can be found from

$$\min_{\mathbf{w}} \quad \|\mathbf{z} - \mathbf{U}\mathbf{w}\|_2^2, \quad \mathbf{w} \geq \mathbf{0} \quad (1)$$

where the inequality constraint is element-wise. It can be seen that $\mathbf{P}_{\mathcal{K}}(\mathbf{z}) = \mathbf{z}, \forall \mathbf{z} \in \mathcal{K}$. Otherwise, if $\mathbf{z} \notin \mathcal{K}$ then it can be easily proved by contradiction that the solution \mathbf{w} is such that $\mathbf{U}\mathbf{w}$ lies on a facet of \mathcal{K} . Let \mathcal{K}^- be the cone with the basis $-\mathbf{U}$ then it can be easily shown that if $\mathbf{z} \in \mathcal{K}^-$ then $\mathbf{P}_{\mathcal{K}}(\mathbf{z}) = \mathbf{0}$.

Returning to the ranking problem, we need to find a K -degree polyhedral cone \mathcal{K} that captures most of the training data. Define the ℓ_2 distance from \mathbf{z} to \mathcal{K} as $d_{\mathcal{K}}(\mathbf{z}) = \|\mathbf{z} - \mathbf{P}_{\mathcal{K}}(\mathbf{z})\|_2$ then we define the document-pair-level loss as

$$l(\mathcal{K}; \mathbf{z}, y) = d_{\mathcal{K}}(\mathbf{z})^2. \quad (2)$$

Suppose that for a query q , a set of document pair differences $S_q = \{\mathbf{z}_1^q, \dots, \mathbf{z}_{n_q}^q\}$ with relevance differences $\phi_1^q, \dots, \phi_{n_q}^q, \phi_j^q > 0$ can be obtained. Following [13], we define the empirical query-level loss as

$$\hat{L}(\mathcal{K}; q, S_q) = \frac{1}{n_q} \sum_{j=1}^{n_q} l(\mathcal{K}; \mathbf{z}_j^q, y_j^q). \quad (3)$$

For a full training set of P queries and $S = \{S_{q_1}, \dots, S_{q_P}\}$ samples, we define the query-level empirical risk as

$$\hat{R}(\mathcal{K}; S) = \frac{1}{P} \sum_{i=1}^P \hat{L}(\mathcal{K}; q_i, S_{q_i}). \quad (4)$$

Thus, the polyhedral cone \mathcal{K} can be found from minimizing this query-level empirical risk. Note that even though other performance measures such as mean average precision (MAP) or normalized discounted cumulative gain (NDCG) is the ultimate assessment, it is observed that good empirical risk often leads to good MAP/NDCG and simplifies the learning. We next discuss some additional constraints for the algorithm to achieve good generalization ability.

3.3 Modification

Normalization. Using the proposed approach, the direction of the vector \mathbf{z} is more important than its magnitude. However, at the same time, if the magnitude of \mathbf{z} is small it is desirable to suppress its contribution to the objective function. We thus propose the normalization of input document-pair differences as follows

$$\mathbf{z} \leftarrow \rho \mathbf{z} / (\alpha + \|\mathbf{z}\|_2), \quad \alpha, \rho > 0. \quad (5)$$

The constant ρ is simply the scaling factor whilst α is to suppress the noise when $\|\mathbf{z}\|_2$ is too small. With this normalization, we note that

$$\|\mathbf{z}\|_2 \leq \rho. \quad (6)$$

Relevance weighting. In the current setting, we consider all ordered document-pairs equally important. This is however a disadvantage because the cost of the mismatch between the two vectors which are close in rank is less than the cost between those distant in rank. To address this issue, we propose an extension of (2)

$$l(\mathcal{K}; \mathbf{z}, y) = \phi d_{\mathcal{K}}(\mathbf{z})^2. \quad (7)$$

where $\phi > 0$ is the corresponding ordered relevance difference.

Conic regularization. From statistical learning theory [15, ch.4], it is known that in order to obtain good generalization bounds, it is important to restrict the hypothesis space from which

the learned function is to be found. Otherwise, the direct solution from an unconstrained empirical risk minimization problem is likely to overfit and introduces large variance (uncertainty). In many cases, this translates to controlling the complexity of the learning function. In the case of support vector machines (SVMs), this has the intuitive interpretation of maximizing the margin, which is the inverse of the norm of the learning function in the Hilbert space.

In our problem, we seek a cone which captures most of the training examples, i.e. the cone that encloses the conic hull of most training samples. In the SVM case, there are many possible hyperplanes that separates the samples without a controlled margin. Similarly, there is also a large number of polyhedral cones that can capture the training samples without further constraints. In fact, minimizing the empirical risk will tend to select the cone with larger solid angle so that the training examples will have small loss (see Fig. 2). In our case, the complexity is translated roughly to the size (volume) of the cone. The bigger cone will likely overfit (enclose) the noisy training samples and thus reduces generalization. Thus, we propose the following constraint to *indirectly* regularize the size of the cone

$$0 \leq \lambda_l \leq \|\mathbf{w}\|_1 \leq \lambda_u, \quad \mathbf{w} \geq \mathbf{0} \quad (8)$$

where \mathbf{w} is the coefficients defined as in (1) and for simplicity we set $\lambda_l = 1$. To see how this effectively controls \mathcal{K} , consider a 2D toy example in Fig. 2. If $\lambda_u = 1$, the solution is the cone \mathcal{K}_1 . In this case, the loss of the positive training examples (within the cone) is the distance from them to the simplex define over the basis vectors $\mathbf{u}_1, \mathbf{u}_2$ (i.e. $\{\mathbf{z} : \mathbf{z} = \lambda\mathbf{u}_1 + (1 - \lambda)\mathbf{u}_2, 0 \leq \lambda \leq 1\}$) and the loss of the negative training example is the distance to the cone. With the same training examples, if we let $\lambda_u > 1$ then there exists a cone solution \mathcal{K}_2 such that all the losses are effectively zero. In particular, for each training example, there exists a corresponding $\|\mathbf{w}\|_1 = \lambda$ such that the corresponding simplex $\{\mathbf{z} : \mathbf{z} = w_1\mathbf{u}_1 + w_2\mathbf{u}_2, w_1 + w_2 = \lambda\}$, passes all positive training examples.

Finally, we note that as the product $\mathbf{U}\mathbf{w}_j^{q_i}$ appears in the objective function and that both \mathbf{U} and $\mathbf{w}_j^{q_i}$ are variables then there is a scaling ambiguity in the formulation. We suggest to address this scale ambiguity by considering the norm constraint $\|\mathbf{u}_k\|_2 = c > 0$ on the basis vectors.

In summary, the proposed formulation can be explicitly written as

$$\begin{aligned} \min_{\mathbf{U}} \left\{ \frac{1}{P} \sum_{i=1}^P \frac{1}{n_{q_i}} \left(\sum_{j=1}^{n_{q_i}} \min_{\mathbf{w}_j^{q_i}} \phi_j^{q_i} \|\mathbf{z}_j^{q_i} - \mathbf{U}\mathbf{w}_j^{q_i}\|_2^2 \right) \right\} \\ \text{s.t. } \|\mathbf{u}_k\|_2 = c, \mathbf{w}_j^{q_i} \geq \mathbf{0}, 0 < \lambda_l \leq \|\mathbf{w}_j^{q_i}\|_1 \leq \lambda_u. \end{aligned} \quad (9)$$

3.4 Generalization bound

We restrict our study on generalization bound from an algorithmic stability viewpoint, which is initially introduced in [2] and based on the concentration property of random variables. In the ranking context, generalization bounds for point-wise ranking / ordinal regression have been obtained [1, 8]. Recently, [13] show that the generalization bound result in [2] still holds in the ranking context. More specifically, we would like to study the variation of the expected query-level risk, defined as

$$R(\mathcal{K}) = \int_{\mathcal{Q} \times \mathcal{Y}} L(\mathcal{K}; q) P_{\mathcal{Q}}(dq). \quad (10)$$

where $L(\mathcal{K}; q)$ denotes the expected query-level loss defined as

$$L(\mathcal{K}; q) = \int_{\mathcal{Z}} l(\mathcal{K}; \mathbf{z}^q, y^q) P_{\mathcal{Z}}(d\mathbf{z}^q) \quad (11)$$

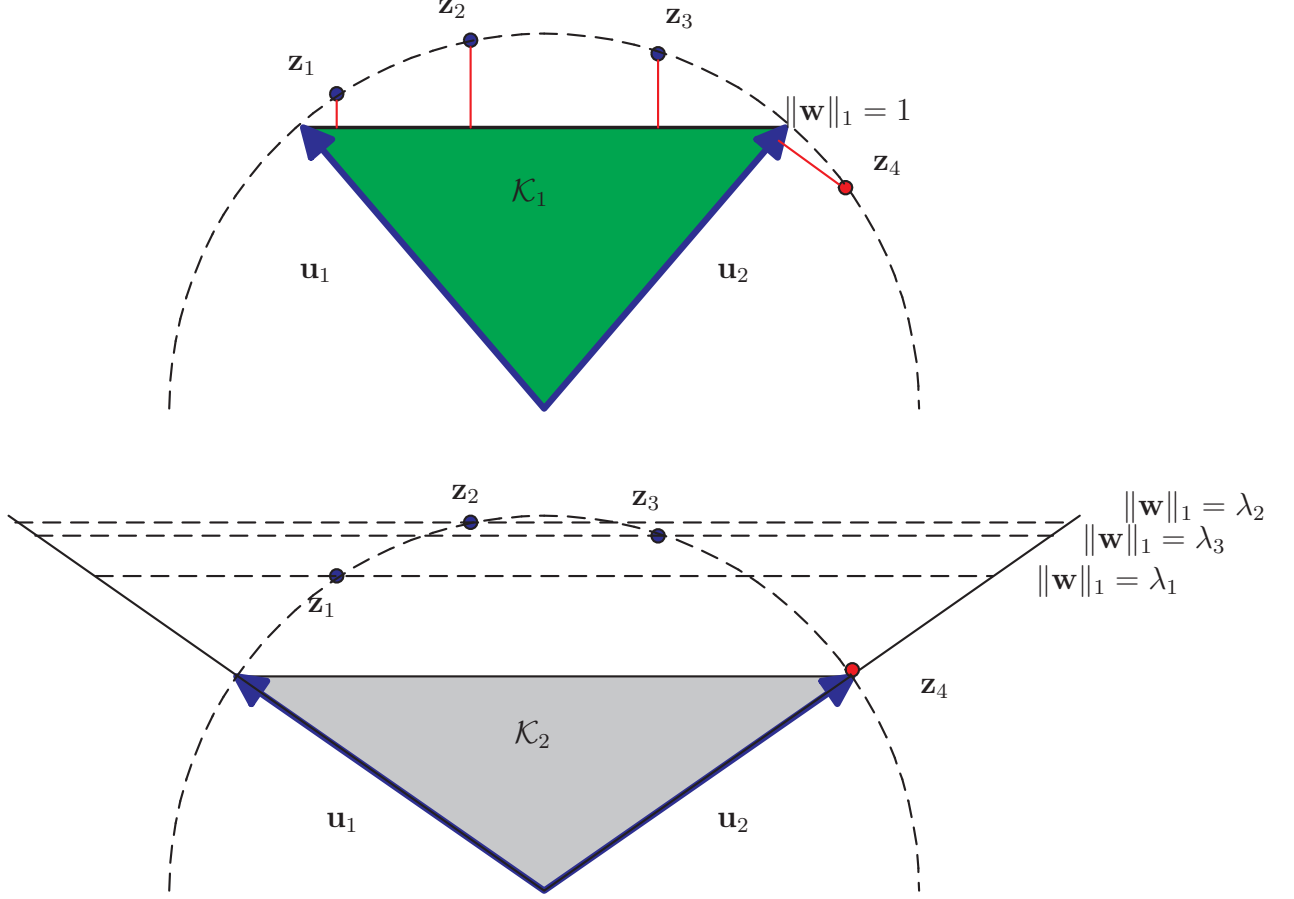


Figure 2: Illustration of different cone solutions. For simplicity, we plot for the case $c = 1$ and $\|\mathbf{z}\|_2 \approx 1$.

and $P_{\mathbf{Z}}$ denotes the probability distribution of the (ordered) document differences.

Following [2] and [13] we define the uniform leave-one-query-out document-pair-level stability as

$$\beta = \sup_{q \in \mathcal{Q}, i \in [1, \dots, P]} |l(\mathcal{K}_S; \mathbf{z}^q, y^q) - l(\mathcal{K}_{S-i}; \mathbf{z}^q, y^q)| \quad (12)$$

where \mathcal{K}_S and \mathcal{K}_{S-i} are respectively the polyhedral cones learned from the full training set and that without the i th query. As stated in [13], it can be easily shown the following query-level stability bounds by integration or average sum of the term on the left hand side in the above definition

$$|L(\mathcal{K}_S; q) - L(\mathcal{K}_{S-i}; q)| \leq \beta, \forall i \quad (13)$$

$$|\hat{L}(\mathcal{K}_S; q) - \hat{L}(\mathcal{K}_{S-i}; q)| \leq \beta, \forall i. \quad (14)$$

Using the above query-level stability results and by considering S_{q_i} as query-level samples, one can directly apply the result in [2] (see also [13]) to obtain the following generalization bound

Theorem 1 *For the proposed ConeRank algorithm with uniform leave-one-query-out document-pair-level stability β , with probability of at least $1 - \varepsilon$ it holds*

$$R(\mathcal{K}_S) \leq \hat{R}(\mathcal{K}_S) + 2\beta + (4P\beta + \gamma) \sqrt{\frac{\ln(1/\varepsilon)}{2P}}, \quad (15)$$

where $\gamma = \sup_{q \in \mathcal{Q}} l(\mathcal{K}_S; \mathbf{z}^q, y^q)$ and $\varepsilon \in [0, 1]$.

As can be seen, the bound on the expected query-level risk depends on the stability. It is of practical interest to study the stability β for the proposed algorithm. The following result shows that the change in the cone due to leaving one query out can provide an effective upper bound on the uniform stability β . For notational simplicity, we only consider the non-weighted version of the loss, as the weighted version is simply a scale of the bound by the maximum weight.

Theorem 2 Denote as \mathbf{U} and \mathbf{U}^{-i} the ‘basis’ vectors of the polyhedral cones \mathcal{K}_S and \mathcal{K}_{S-i} respectively. For a ConeRank algorithm with non-weighted loss, we have

$$\beta \leq 2s_{\max}\lambda_u(\rho + \sqrt{K}c\lambda_u) + s_{\max}^2\lambda_u^2, \quad (16)$$

where $s_{\max} = \max_i \|\mathbf{U} - \mathbf{U}^{-i}\|$, $\|\bullet\|$ denotes the spectral norm, and ρ is the normalizing factor of \mathbf{z} (c.f. (6)).

Proof. Following the proposed algorithm, we equivalently study the bound of

$$\beta = \sup_{\substack{q \in \mathcal{Q} \\ \|\mathbf{z}^q\|_2 \leq \rho}} \left| \min_{\mathbf{w} \in \mathcal{C}} \|\mathbf{z}^q - \mathbf{U}\mathbf{w}\|_2^2 - \min_{\mathbf{w} \in \mathcal{C}} \|\mathbf{z}^q - \mathbf{U}^{-i}\mathbf{w}\|_2^2 \right|$$

where the constraint set $\mathcal{C} = \{\mathbf{w} : \mathbf{w} \geq \mathbf{0}, \lambda_l \leq \|\mathbf{w}\|_1 \leq \lambda_u\}$. Without loss of generality, we can assume that

$$\min_{\mathbf{w} \in \mathcal{C}} \|\mathbf{z}^q - \mathbf{U}\mathbf{w}\|_2^2 > \min_{\mathbf{w} \in \mathcal{C}} \|\mathbf{z}^q - \mathbf{U}^{-i}\mathbf{w}\|_2^2$$

and the minima are attained at \mathbf{w} and \mathbf{w}^{-i} respectively. Due to the definition, it follows that

$$\beta \leq \sup_{\substack{q \in \mathcal{Q} \\ \|\mathbf{z}^q\|_2 \leq \rho}} (\|\mathbf{z}^q - \mathbf{U}\mathbf{w}^{-i}\|_2^2 - \|\mathbf{z}^q - \mathbf{U}^{-i}\mathbf{w}^{-i}\|_2^2).$$

Expanding the term on the left, and using matrix norm inequalities, one obtains

$$\begin{aligned} \beta \leq \sup_{q \in \mathcal{Q}} & (2\|\mathbf{U}\|\|\Delta\| + \|\Delta\|^2)\|\mathbf{w}^{-i}\|_2^2 \\ & + 2\|\mathbf{z}^q\|_2\|\Delta\|\|\mathbf{w}^{-i}\|_2) \end{aligned} \quad (17)$$

where $\Delta = \mathbf{U} - \mathbf{U}^{-i}$. The proof follows by the following facts

- $\|\mathbf{U}\| \leq \sqrt{K}c$ due to each $\|\mathbf{u}_k\|_2 \leq c$ and that $\|\mathbf{U}\| \leq \|\mathbf{U}\|_F$ where $\|\bullet\|_F$ denotes the Frobenius norm.
- $\|\mathbf{w}\|_2^2 \leq \|\mathbf{w}\|_1^2$ for $\mathbf{w} \geq \mathbf{0}$
- $\|\mathbf{z}^q\|_2 \leq \rho$ due to the normalization

and that $\|\Delta\| \leq s_{\max}$ by definition.

It is more interesting to study the bound on s_{\max} . We conjecture that this will depend on the sample size as well as the nature of the proposed conic regularization. However, this is still an open question and such an analysis is beyond the scope of the current work.

We note importantly that as the stability bound can be made small by lowering λ_u . Doing so definitely improves stability at the cost of making the empirical risk large and hence the bias becomes significantly undesirable. In practice, it is important to select proper values of the parameters to provide optimal bias-variance trade-off. Next, we turn the discussion on practical implementation of the ideas, taking into account the large-scale nature of the problem.

4 Implementation

In the original formulation (9), the scaling ambiguity is resolved by placing a norm constraint on \mathbf{u}_k . However, a direct implementation seems difficult. In what follows, we propose an alternative implementation by resolving the ambiguity on \mathbf{w} instead. We fix $\|\mathbf{w}\|_1 = 1$ and consider the norm inequality constraint on \mathbf{u}_k as $\|\mathbf{u}_k\|_2 \leq c$ (i.e. convex relaxation on equality constraint) where c is a constant of $\mathcal{O}(\|\mathbf{z}^q\|_2)$. This leads to an *approximate* formulation

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{w}_j^{q_i}} & \left\{ \frac{1}{P} \sum_{i=1}^P \frac{1}{n_{q_i}} \left(\sum_{j=1}^{n_{q_i}} \phi_j^{q_i} \|\mathbf{z}_j^{q_i} - \mathbf{U} \mathbf{w}_j^{q_i}\|_2^2 \right) \right\} \\ \text{s.t. } & \|\mathbf{u}_k\|_2 \leq c, \mathbf{w}_j^{q_i} \geq \mathbf{0}, \|\mathbf{w}_j^{q_i}\|_1 = 1. \end{aligned} \quad (18)$$

The advantage of this approximation is that the optimization problem is now convex with respect to each \mathbf{u}_k and still convex with respect to each $\mathbf{w}_j^{q_i}$. This suggests an alternating and iterative algorithm, where we only vary a subset of variables and fix the rest. The objective function should then always decrease. As the problem is not strictly convex, there is no guarantee of a global solution. Nevertheless, a locally optimal solution can be obtained. The additional advantage of the formulation is that gradient-based methods can be used for each sub-problem and this is very important in large-scale problems.

Algorithm 1 Stochastic Gradient Descent

Input: queries q_i and pair differences $\mathbf{z}_j^{q_i}$.
Randomly initialize \mathbf{u}_k , $\forall k \leq K$; set $\mu > 0$
repeat
 1. The *folding-in* step (fixed \mathbf{U}):
 Randomly initialize $\mathbf{w}_j^{q_i} : \mathbf{w}_j^{q_i} \geq \mathbf{0}; \|\mathbf{w}_j^{q_i}\|_1 = 1$;
 repeat
 1a. Compute $\mathbf{w}_j^{q_i} \leftarrow \mathbf{w}_j^{q_i} - \mu \partial \hat{R}(\mathbf{w}_j^{q_i}) / \partial \mathbf{w}_j^{q_i}$
 1b. Set $\mathbf{w}_j^{q_i} \leftarrow \max\{\mathbf{w}_j^{q_i}, \mathbf{0}\}$ (element-wise)
 1c. Normalize $\mathbf{w}_j^{q_i} \leftarrow \mathbf{w}_j^{q_i} / \|\mathbf{w}_j^{q_i}\|_1$
 until converged
 2. The *basis-update* step (fixed \mathbf{w}):
 for $k = 1$ **to** K **do**
 2a. Update $\mathbf{u}_k \leftarrow \mathbf{u}_k - \mu \partial \hat{R}(\mathbf{u}_k) / \partial \mathbf{u}_k$
 2b. Normalize \mathbf{u}_k to norm c if violated.
 end for
until converged

4.1 Stochastic Gradient

Since the number of pairs may be large for typically real datasets, we do not want to store every \mathbf{w}_j^q . Instead, for each iteration, we perform a *folding-in* operation, in that we fix the basis \mathbf{U} , and estimate the coefficients \mathbf{w}_j^q . Since this is a convex problem, it is possible to apply the stochastic gradient (SG) method as shown in Algorithm 1. Note that we express the empirical risk as the function of *only* variable of interest when other variables are fixed for notational simplicity. In practice, we also need to check if the cone is proper and we find this is always satisfied.

4.2 Exponentiated Gradient

Exponentiated Gradient (EG) [12] is an algorithm for estimating distribution-like parameters. Thus, Step 1a can be replaced by

$$\mathbf{w}_j^{q_i} \leftarrow \mathbf{w}_j^{q_i} \exp \left\{ -\mu \partial \hat{R}(\mathbf{w}_j^{q_i}) / \partial \mathbf{w}_j^{q_i} \right\} \text{ (element-wise).}$$

For faster numerical computation (by avoiding the exponential), as shown in [12], this step can be approximated by

$$(\mathbf{w}_j^{q_i})_k \leftarrow (\mathbf{w}_j^{q_i})_k \left(1 - \mu \left(\partial \hat{R}(\hat{\mathbf{z}}_j^{q_i}) / \partial \hat{\mathbf{z}}_j^{q_i} \right)^\top (\mathbf{u}_k - \hat{\mathbf{z}}_j^{q_i}) \right)$$

where the empirical risk \hat{R} is parameterized in terms of $\hat{\mathbf{z}}_j^{q_i} = \mathbf{U} \mathbf{w}_j^{q_i}$. When the learning rate μ is sufficiently small, this update readily ensures the normalization of $\mathbf{w}_j^{q_i}$. The main difference between SG and EG is that, update in SG is *additive*, while it is *multiplicative* in EG.

Algorithm 2 Query-level Prediction

Input: New query q with pair differences $\{\mathbf{z}_j^q\}_{j=1}^{n_q}$

Maintain a scoring array A of all pre-computed feature vectors, initialize $A_l = 0$ for all l .

Set $\phi_j^q = 1, \forall j \leq n_q$.

for $j = 1$ **to** n_q **do**

 Perform *folding-in* to estimate the coefficients without the non-negativity constraints.

 Check if the sum of the coefficients is positive, then $A_l \leftarrow A_l + 1$; otherwise $A_m \leftarrow A_m + 1$

end for

Output the ranking based on the scoring array A .

4.3 Prediction

Assume that the basis $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K)$ has been learned during training. In testing, for each query, we are also given a set of feature vectors, and we need to compute a ranking function that outputs the appropriate positions of the vectors in the list.

Unlike the training data where the order of the pair (l, m) is given, now this order information is missing. This breaks down the conic assumption, in that the difference of the two vectors is the non-negative combination of the basis vectors. Since the either preference orders can potentially be incorrect, we relax the constraint of the non-negative coefficients. The idea is that, if the order is correct, then the coefficients are mostly positive. On the other hand, if the order is incorrect, we should expect that the coefficients are mostly negative. The query-level prediction is proposed as shown in Algorithm 2. As this query-level prediction is performed over a query, it can address the shortcoming of logical discrepancy of document-level prediction in the pairwise approach.

5 Discussion

RankSVM [11] defines the following loss function over ordered pair differences

$$L(\mathbf{u}) = \frac{1}{P} \sum_j \max(0, 1 - \mathbf{u}^\top \mathbf{z}_j) + \frac{C}{2} \|\mathbf{u}\|_2^2$$

where $\mathbf{u} \in \mathbb{R}^N$ is the parameter vector, $C > 0$ is the penalty constant and P is the number of data pairs.

Being a pairwise approach, RankNet instead uses

$$L(\mathbf{u}) = \frac{1}{P} \sum_j \log(1 + \exp\{-\mathbf{u}^\top \mathbf{z}_j\}) + \frac{C}{2} \|\mathbf{u}\|_2^2.$$

This is essentially the 1-class SVM applied over the ordered pair differences. The quadratic regularization term tends to push the separating hyperplane away from the origin, i.e. maximizing the 1-class margin.

It can be seen that the RankSVM solution is the special case when the cone approaches a halfspace. In the original RankSVM algorithm, there is no intention to learn a non-negative subspace where ordinal information is to be found like in the case of ConeRank. This could potentially give ConeRank more analytical power to trace the origin of preferences.

6 Experiments

6.1 Data and Settings

We run the proposed algorithm on the latest and largest benchmark data LETOR 4.0. This has two data sets for supervised learning, namely MQ2007 (1700 queries) and MQ2008 (800 queries). Each returned document is assigned a integer-valued relevance score of $\{0, 1, 2\}$ where 0 means that the document is irrelevant with respect to the query. For each query-document pair, a vector of 46 features is pre-extracted, and available in the datasets. Example features include the term-frequency and the inverse document frequency in the body text, the title or the anchor text, as well as link-specific like the PageRank and the number of in-links. The data is split into a training set, a validation set and a test set. We normalize these features so that they are roughly distributed as Gaussian with zero means and unit standard deviations. During the folding-in step, the parameters \mathbf{w}_j^q corresponding to pair j th of query q are randomly initialized from the non-negative uniform distribution and then normalized so that $\|\mathbf{w}_j^q\|_1 = 1$. The basis vectors \mathbf{u}_k are randomly initialized to satisfy the relaxed norm constraint. The learning rate is $\mu = 0.001$ for the SG and $\mu = 0.005$ for the EG. For normalization, we select $\alpha = 1$ and $\rho = \sqrt{N}$ where N is the number of features, and we set $c = 2\rho$.

6.2 Results

The two widely-used evaluation metrics employed are the Mean Average Precision (MAP) and the Normalized Discounted Cumulative Gain (NDCG). We use the evaluation scripts distributed with LETOR 4.0.

In the first experiment, we investigate the performance of the proposed method with respect to the number of basis vectors K . The result of this experiment on the MQ2007 dataset is shown in Fig. 3. We note an interesting observation that the performance is highest at about $K = 10$ out of 46 dimensions of the original feature space. This seems to suggest that the idea of capturing an informative subspace using the cone makes sense on this dataset. Furthermore, the study on the eigenvalue distribution of the non-centralized ordered pairwise differences on the MQ2007 dataset, as shown in Fig. 4, also reveals that this is about the dimension that can capture most of the data energy.

We then compare the proposed and recent base-line methods² in the literature and the results on the MQ2007 and MQ2008 datasets are shown in Table 1. The proposed ConeRank is studied with $K = 10$ due to the previous experiment. We note that all methods tend to perform better

²from <http://research.microsoft.com/en-us/um/beijing/projects/letor/letor4baseline.aspx>

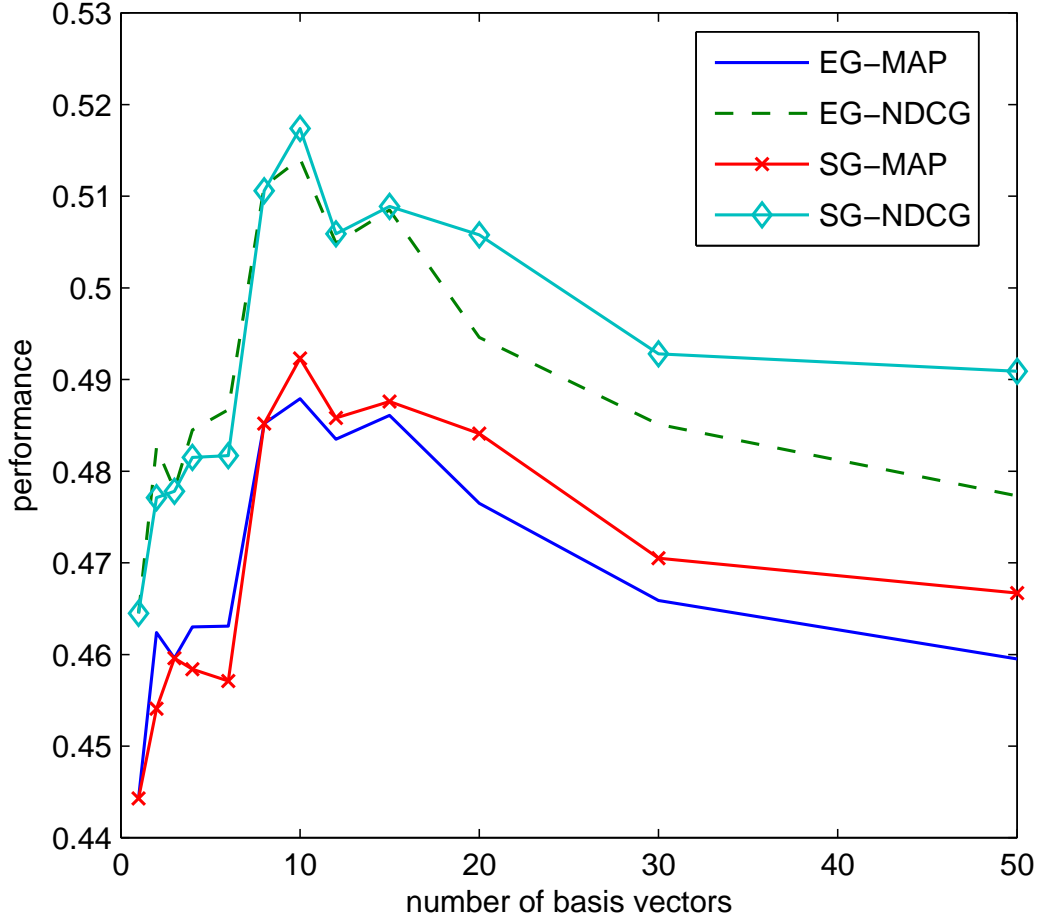


Figure 3: Performance versus basis number

on MQ2007 than MQ2008, which can be explained by the fact that the MQ2007 dataset is much larger than the other, and hence provides better training.

On the MQ2007 dataset, ConeRank compares favourably with other methods. For example, ConeRank-SG achieves the highest MAP score, whilst its NDCG score differs only less than 2% when compared with the best (RankSVM-struct). On the MQ2008 dataset, ConeRank still maintains within the 3% margin of the best methods on both MAP and NDCG metrics.

Table 1: Results on LETOR 4.0.

ALGORITHMS	MQ2007		MQ2008	
	MAP	NDCG	MAP	NDCG
ADARANK-MAP	0.482	0.518	0.463	0.480
ADARANK-NDCG	0.486	0.517	0.464	0.477
LISTNET	0.488	0.524	0.450	0.469
RANKBOOST	0.489	0.527	0.467	0.480
RANKSVM-STRUCT	0.489	0.528	0.450	0.458
CONERANK-EG	0.488	0.514	0.444	0.456
CONERANK-SG	0.492	0.517	0.454	0.464

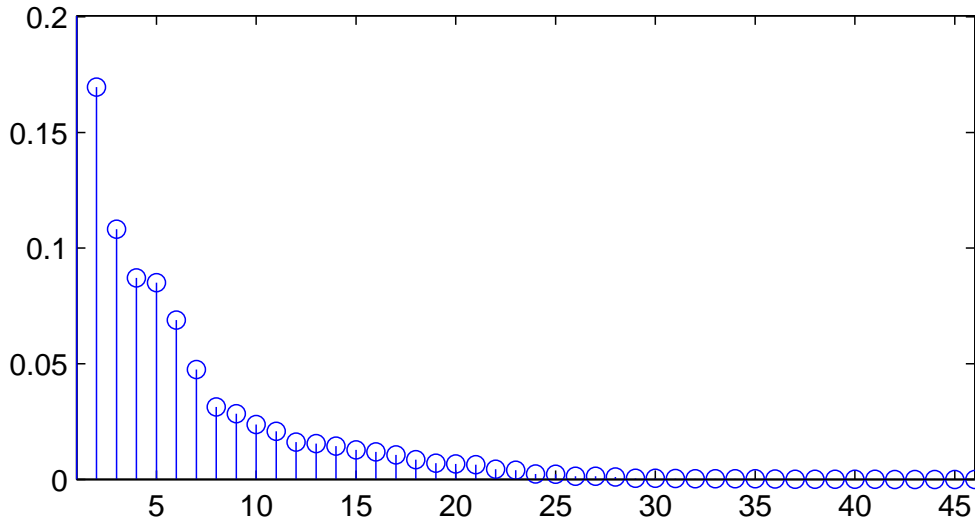


Figure 4: Eigenvalue distribution on the MQ2007 dataset.

7 Conclusion

We have presented a new view on the learning to rank problem from a generalized inequalities perspective. We formulate the problem as learning a polyhedral cone that uncovers the non-negative subspace where ordinal information is found. A practical implementation of the method is suggested which is then observed to achieve comparable performance to state-of-the-art methods on the LETOR 4.0 benchmark data.

There are some directions that require further research, including a more rigorous study on the bound of the spectral norm of the leave-one-query-out basis vector difference matrix, a better optimization scheme that solves the original formulation without relaxation, and a study on the informative dimensionality of the ranking problem.

References

- [1] S. Agarwal, *Lecture notes in artificial intelligence*. Springer-Verlag, 2008, ch. Generalization bounds for some ordinal regression algorithms, pp. 7–21.
- [2] O. Bousquet and A. Elisseeff, “Stability and generalization,” *Journal of Machine Learning Research*, pp. 499–526, 2002.
- [3] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [4] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, “Learning to rank using gradient descent,” in *Proc. ICML*, 2005.
- [5] Z. Cao, T. Qin, T. Liu, M. Tsai, and H. Li, “Learning to rank: from pairwise approach to listwise approach,” in *Proc. ICML*, 2007.
- [6] W. Chu and Z. Ghahramani, “Gaussian processes for ordinal regression,” *Journal of Machine Learning Research*, vol. 6, no. 1, p. 1019, 2006.
- [7] W. Cohen, R. Schapire, and Y. Singer, “Learning to order things,” *J Artif Intell Res*, vol. 10, pp. 243–270, 1999.

- [8] D. Cossock and T. Zhang, “Statistical analysis of Bayes optimal subset ranking,” *IEEE Transactions on Information Theory*, vol. 54, pp. 5140–5154, 2008.
- [9] Y. Freund, R. Iyer, R. Schapire, and Y. Singer, “An efficient boosting algorithm for combining preferences,” *Journal of Machine Learning Research*, vol. 4, no. 6, pp. 933–969, 2004.
- [10] R. Herbrich, T. Graepel, and K. Obermayer, “Large margin rank boundaries for ordinal regression,” in *Proc. KDD*, 2000.
- [11] T. Joachims, “Optimizing search engines using clickthrough data,” in *Proc. KDD*, 2002.
- [12] J. Kivinen and M. Warmuth, “Exponentiated gradient versus gradient descent for linear predictors,” *Information and Computation*, 1997.
- [13] Y. Lan, T.-Y. Liu, T. Quin, Z. Ma, and H. Li, “Query-level stability and generalization in learning to rank,” in *Proc. ICML*, 2008.
- [14] R. Plackett, “The analysis of permutations,” *Applied Statistics*, pp. 193–202, 1975.
- [15] B. Schölkopf and Smola, *Learning with kernels*. MIT Press, 2002.