Generalization Analysis for Ranking Using Integral Operator

Yong Liu,[1] **Shizhong Liao,**[2] **Hailun Lin,**[1] **Yinliang Yue,**[1*] **Weiping Wang**[1]
[1]Institute of Information Engineering, CAS
[2]School of Computer Science and Technology, Tianjin University
{liuyong,linhailun,yueyinliang,wangweiping}@iie.ac.cn, {szliao,yongliu}@tju.edu.cn

Abstract

The study on generalization performance of ranking algorithms is one of the fundamental issues in ranking learning theory. Although several generalization bounds have been proposed based on different measures, the convergence rates of the existing bounds are usually at most $O\big(\frac{1}{\sqrt{n}}\big)$, where n is the size of data set. In this paper, we derive novel generalization bounds for the regularized ranking in reproducing kernel Hilbert space via integral operator of kernel function. We prove that the rates of our bounds are much faster than $O\big(\frac{1}{\sqrt{n}}\big)$. Specifically, we first introduce a notion of local Rademacher complexity for ranking, called local ranking Rademacher complexity, which is used to measure the complexity of the space of loss functions of the ranking. Then, we use the local ranking Rademacher complexity to obtain a basic generalization bound. Finally, we establish the relationship between the local Rademacher complexity and the eigenvalues of integral operator, and further derive sharp generalization bounds of faster convergence rate.

Introduction

Ranking is an important problem in various applications, such as information retrieval, natural language processing, computational biology and social sciences (Freun et al. 2003; Cortes, Mohri, and Rastogi 2007). In ranking, one learns a real-valued function that assigns scores to instances, but the scores themselves do not matter; instead, what is important is the relative ranking of instances induced by those scores. From different perspectives, various ranking algorithms have been proposed including Ranking SVM (Herbrich, Graepel, and Obermayer 1999), PRanking(Crammer and Singer 2001), RankNet (Burges et al. 2005; Burges, Ragno, and Le 2007), RankBoost (Freun et al. 2003), P-norm push ranking (Rudin 2009), subset ranking (Cossock and Zhang 2008), ListNet (Cao et al. 2007), ListMLE (Xia et al. 2008), kernel-based regularized ranking (Agarwal and Niyogi 2009), etc.

To understand existing ranking algorithms and guide the development of new ones, people have to investigate the generalization ability of ranking algorithms. A sharper generalization bound usually implies more consistent performances on the training set and the test set. In recent years, some generalization bounds have been developed to estimate the ability of ranking algorithms based on different measures, such as VC-dimension (Freun et al. 2003), cover number (Rudin and Schapire 2009; Rejchel 2012), algorithmic stability (Agarwal and Niyogi 2009), Rademacher complexity (Clémençon, Lugosi, and Vayatis 2005; 2008; Lan et al. 2008; 2009; Chen, Liu, and Ma 2010), etc. Although there have been several recent advances in the studying of generalization bounds of ranking algorithms, the orders of the convergence rates of the existing generalization bounds are usually at most $O\big(\frac{1}{\sqrt{n}}\big)$.

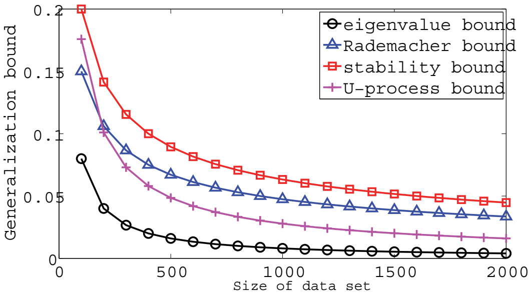

Figure 1: Our eigenvalue-based bound in Theorem 4, the standard stability-based bound (Agarwal and Niyogi 2009) and Rademacher bound (Clémençon, Lugosi, and Vayatis 2005), and U-process-based bound (Rejchel 2012).

In this paper, we consider the regularized ranking algorithms in reproducing kernel Hilbert space (RKHS), and derive sharper generalization bounds with convergence rates that are much faster than $O\big(\frac{1}{\sqrt{n}}\big)$. Specifically, we first introduce a notion of local ranking Rademacher complexity for ranking, which is an extension of the traditional local Rademacher complexity for classification or regression problems. We then use this notion as a tool to perform generalization analysis for ranking, and derive sharper generalization bounds with the eigenvalues of the integral operator associated with kernel function. We finally conduct experiments to empirically analyze the performance of our pro-

[*]Corresponding author

posed bounds. Our bound, the standard algorithmic stability bound (Agarwal and Niyogi 2009), Rademacher bound (Clémençon, Lugosi, and Vayatis 2005) and the state-of-the-art U-process bound (Rejchel 2012) are plotted in Figure 1 (see Section 4 in detail). The plot shows that our bound is sharper than the stability bound, Rademacher bound and U-process bound, which demonstrates the effectiveness of using the eigenvalues of integral operator to estimate the generalization error for ranking.

To our knowledge, this is the first attempt to use the notion of local Rademacher complexity and the eigenvalues of integral operator to derive generalization error bounds for ranking. Major contributions of the paper include:

1) A novel extension version of local Rademacher complexity of the space of loss functions for ranking is proposed.
2) The proof of theorem that gives the generalization error bound on the basis of the local ranking Rademacher complexity is given.
3) The relationship between the local ranking Rademacher complexity and the eigenvalues of integral operator is established.
4) The generalization bounds of fast convergence rates based on the eigenvalues of integral operator are derived.

Related Work

In this subsection, we introduce the related work about the generalization bounds of raking algorithms and the local Rademacher complexity.

Generalization Bounds of Ranking Algorithms Freund et al. (2003) gave a generalization bound of RankBoost in the bipartite setting. Their bound was expressed in terms of the VC-dimension of a class of binary classification functions. Agarwal and Niyogi (2005) used a different tool, namely that of algorithmic stability (Bousquet and Elisseeff 2002), to obtain generalization bounds for bipartite ranking algorithms. Agarwal and Niyogi (2009) generalized the above results for general ranking problems. Cortes, Mohri, and Rastogi (2007) considered a different setting of the ranking problem and derived stability-based generalization bounds for algorithms in this setting. Rudin and Schapire (2009) derived a margin-based bound in terms of the covering numbers, and described a new algorithm based on the derived bound. Lan et al. (2008) explored the query-level generalization ability of ranking algorithms using query-level stability. Lan et al. (2009) generalized the above work, and gave a generalization bound for the listwise ranking algorithms based on the basis of Rademacher complexity of the class of compound functions. Chen, Liu, and Ma (2010) proposed a novel theoretical framework to perform generalization analysis of the listwise ranking algorithms under the assumption of two-layer sampling. Rejchel (2012) investigated the generalization properties of convex risk minimizers based on the empirical and U-process theory (Pakes and Pollard 1989), and proposed a generalization bound for the regularized ranking algorithms with Gaussian kernel on the hinge loss. Although various generalization bounds have been developed based on different measures, the convergence rates of the existing bounds are usually at most $O\big(\frac{1}{\sqrt{n}}\big)$. In this paper, based on the notion of local Rademacher complexity and eigenvalues of integral operator, we will derive bounds with rates that are much faster than $O\big(\frac{1}{\sqrt{n}}\big)$.

Local Rademacher Complexity One of the useful data-dependent complexity measures used in the generalization analysis for traditional classification or regression problems is the notion of Rademacher complexity (Bartlett and Mendelson 2002; Koltchinskii and Panchenko 2002). However, it provides global estimates of the complexity of the function class, that is, it does not reflect the fact that the algorithm will likely pick functions that have a small error. In recent years, several authors have applied $local$ Rademacher complexity to obtain better generalization error bounds for traditional classification or regression (Bartlett, Bousquet, and Mendelson 2005; Koltchinskii 2006; Liu and Liao 2015). The local Rademacher complexity considers Rademacher averages of a smaller subset of the hypothesis set, so it is always smaller than the corresponding global one. Unfortunately, the empirical loss of ranking is usually a non-sum-of-i.i.d. pairwise loss, therefore the techniques for proving tight generalization error bounds for traditional classification or regression can not be applied to ranking. To address this problem, in this paper, we define a new version of local Rademacher complexity for ranking, and use the permutations to convert the non-sum-of-i.i.d. pairwise loss to a sum-of-i.i.d. form for deriving tight bounds.

The rest of the paper is organized as follows. In Section 2, we introduce some notations and preliminaries. In Section 3, we give the definition of local ranking Rademacher complexity, and then use it to perform generalization analysis. In Section 4, we bound the local Rademacher complexity using the eigenvalues of integral operator, and further derive tight generalization error bounds of fast convergence rate. We end in Section 5 with conclusion. Due to limited space, we only give the sketch of proofs in main body, the details are given in supplementary material.

Notations and Preliminaries

Let $S = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n$ be a sample set of size n drawn identically and independently from a fixed, but unknown probability distribution P on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes the input space and $\mathcal{Y} \subset \mathbb{R}$ denotes the output domain. In the problem of ranking, the goal is to learn a real-valued function $f : \mathcal{X} \to \mathbb{R}$ that induces a ranking or ordering over an instance space. Let $\ell : \mathcal{Y}^{\mathcal{X}} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+ \cup \{0\}$ be a loss function that assigns to each function f and $z, z' \in \mathcal{Z}$ to a non-negative real number $\ell(f, z, z')$, interpreted as the penalty or loss of f in its relative ranking of $\mathbf{x}$ and $\mathbf{x}'$ given corresponding labels y and y'.

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel, that is, K is symmetric and for any finite set of instances, its corresponding kernel matrix is positive semidefinite (Aronszajn 1950). The reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with K is defined to be the completion of the linear span of the set of functions $\{K(\cdot, \mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ with the inner product denoted as $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{x}) \rangle = K(\mathbf{x}, \mathbf{x}')$.

In this paper, we study the regularized ranking algorithms in RKHS:

$$f_S = \arg\min_{f\in\mathcal{H}} \frac{1}{n(n-1)} \sum_{i\neq j} \ell(f, z_i, z_j) + \lambda \|f\|_K^2,$$

where $\mathcal{H}$ is the RKHS with respect to K, λ is the regularization parameter, and ℓ is a loss function. Common loss including:

- 0-1 loss

  $$\ell(f, z, z') = I(\mathrm{sgn}(y - y') \cdot (f(\mathbf{x}) - f(\mathbf{x}')) < 0),$$

  $I(\cdot)$ is the indicator function, $\mathrm{sgn}(u) = 1$ if $u > 0$, $\mathrm{sgn}(u) = 0$ if $u = 0$, otherwise -1.

- Hinge loss

  $$\ell(f, z, z') = \left(|y - y'| - (f(\mathbf{x}) - f(\mathbf{x}')) \cdot \mathrm{sgn}(y - y')\right)_+ .$$

- Least squares loss

  $$\ell(f, z, z') = \Big(|y - y'| - \mathrm{sgn}(y - y') \cdot (f(\mathbf{x}) - f(\mathbf{x}'))\Big)^2 .$$

- γ loss

  $$\ell(f, z, z') = \begin{cases} |y - y'| & \text{if } t \leq 0, \\ |y - y'| - \dfrac{t}{\gamma} & \text{if } 0 < \dfrac{t}{\gamma} < |y - y'|, \\ 0 & \text{if } \dfrac{t}{\gamma} \geq |y - y'|, \end{cases}$$

  where $t = (f(\mathbf{x}) - f(\mathbf{x}'))\mathrm{sgn}(y - y')$.

The quality of a ranking function f can be measured by its generalization error,

$$L(f) = \mathbb{E}_{z,z'}\left[\ell(f, z, z')\right],$$

where, the expectation is taken over the random choice of the samples z, z' according to P. Since the distribution P is unknown, the generalization error of a ranking function f must be estimated from an empirically observable quantity, and its empirical ranking error is defined as follows:

$$\hat{L}_n(f) = \frac{1}{n(n-1)} \sum_{i\neq j} \ell(f, z_i, z_j).$$

In the following, we assume that $\kappa = \sup_{\mathbf{x}\in\mathcal{X}} K(\mathbf{x}, \mathbf{x}) < \infty$, and $\ell : \mathcal{Y}^{\mathcal{X}} \times \mathcal{Z} \times \mathcal{Z} \to [0, M]$, $M > 0$ is a constant.

A Basic Generalization Bound

In this section, we will first introduce a novel notion of local Rademacher complexity for ranking, and then use it to derive a basic generalization bound.

The Rademacher complexity is usually defined over the space of functions $\mathcal{H}$. In this paper, we generalize this notion to a more general ranking learning framework. For this purpose, we switch from the space of functions $\mathcal{H}$ to the space of loss functions.

Definition 1. *Given a space of functions $\mathcal{H}$ with its associated loss function ℓ, the space of loss functions $\mathcal{L}$ is defined as:*

$$\mathcal{L} = \left\{\ell_f := \ell(f, \cdot, \cdot) | f \in \mathcal{H}\right\}, \tag{1}$$

where $\ell_f : \mathcal{Z} \times \mathcal{Z} \to [0, M]$

$$\ell_f(z, z') = \ell(f, z, z').$$

Then, the generalization error and the empirical error can be rewritten in terms of the space of loss functions:

$$L(f) \equiv L(\ell_f) = \mathbb{E}_{z,z'}\left[\ell(f, z, z')\right],$$

$$\hat{L}_n(f) \equiv \hat{L}_n(\ell_f) = \frac{1}{n(n-1)} \sum_{i\neq j} \ell(f, z_i, z_j).$$

Since we do not know in advance which $f \in \mathcal{H}$ will be chosen during the learning phase, in order to estimate $L(\ell_f)$, we have to study the behavior of the difference between the generalization error and the empirical error. To this end, we introduce the notion of uniform deviation of $\mathcal{L}$, denoted as

$$\hat{U}_n(\mathcal{L}) = \sup_{\ell_f\in\mathcal{L}} \left\{L(\ell_f) - \hat{L}_n(\ell_f)\right\}. \tag{2}$$

Note that $\left\{L(\ell_f) - \hat{L}_n(\ell_f)\right\}_{\ell_f\in\mathcal{L}} \leq \hat{U}_n(\mathcal{L})$, so we have

$$L(\ell_f) \leq \hat{L}_n(\ell_f) + \hat{U}_n(\mathcal{L}), \forall \ell_f \in \mathcal{L}.$$

$\hat{U}_n(\mathcal{L})$ is not computable, but we can bound its value via the ranking Rademacher complexity defined as follows:

Definition 2. *Assume $\mathcal{L}$ is a space of loss functions as defined in Definition 1. Then the empirical **ranking Rademacher complexity** of $\mathcal{L}$ is:*

$$\hat{R}_n(\mathcal{L}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\ell_f\in\mathcal{L}} \left|\frac{2}{\lfloor n/2\rfloor} \sum_{i=1}^{\lfloor n/2\rfloor} \sigma_i \ell(f, z_i, z_{\lfloor n/2\rfloor+i})\right|\right],$$

where $\sigma_1, \sigma_2, \ldots, \sigma_{\lfloor n/2\rfloor}$ is an i.i.d. family of Rademacher variables taking values -1 and 1 with equal probability independent of the sample $S = (z_1, \ldots, z_n)$, and $\lfloor n/2\rfloor$ is the largest integer no greater than $\frac{n}{2}$. The ranking Rademacher complexity of $\mathcal{L}$ is:

$$R_n(\mathcal{L}) = \mathbb{E}_S \hat{R}_n(\mathcal{L}).$$

One can see that the ranking Rademacher complexity is defined on the space of loss functions of ranking, which is an extension of the traditional Rademacher complexity defined on $\mathcal{H}$. Thus, we can use it as a tool to perform generalization analysis of ranking.

Although we can use the ranking Rademacher complexity to bound generalization error for ranking, it does not take into consideration the fact that, typically, the hypotheses selected by a learning algorithm have a better performance than in the worst case and belong to a more favorable subfamily of the set of all hypotheses (Cortes, Kloft, and Mohri 2013). Therefore, to derive tight generalization bound, we consider the use of the local Rademacher complexity in this paper. To this end, let $\mathcal{L}^r$ be a star-shaped space of $\mathcal{L}$ with respect to $r > 0$,

$$\mathcal{L}^r = \left\{a\ell_f \middle| a \in [0, 1], \ell_f \in \mathcal{L}, L[(a\ell_f)^2] \leq r\right\},$$

where $L(\ell_f^2) = \mathbb{E}_{z,z'}\left[\ell^2(f, z, z')\right]$.

Definition 3. *For any $r > 0$, the* ***local ranking Rademacher complexity*** *of $\mathcal{L}$ is defined as*

$$R_n(\mathcal{L}^r) := R_n\left(\left\{a\ell \middle| a \in [0,1], \ell \in \mathcal{L}, L[(a\ell)^2] \le r\right\}\right).$$

The key idea to obtain sharp generalization error bound is to choose a much smaller class $\mathcal{L}^r \subseteq \mathcal{L}$ with as small a variance as possible, while requiring that the f_S is still in $\{f | f \in \mathcal{H}, \ell_f \in \mathcal{L}^r\}$.

The generalization error with ranking local Rademacher complexity is given as follows:

Theorem 1. *Assume that r^* is the fixed point of $R_n(\mathcal{L}^r)$, that is, r^* is the solution of $R_n(\mathcal{L}^r) = r$ with respect to r. Then, $\forall k > \max\left(1, \frac{\sqrt{2}}{2M}\right)$, with probability $1 - \delta$:*

$$L(f_S) \le \max\left\{\frac{k}{k-1}\hat{L}_n(f_S), \hat{L}_n(f_S) + c_1 r^* + \frac{c_2}{n-1}\right\},$$

where $c_1 = 8kM$ and $c_2 = 8k\ln\delta + 6$.

Proof. We sketch the proof here. We first prove that the generalization error can be bounded through an assumption over the uniform deviation: if $\hat{U}_n(\bar{\mathcal{L}}) \le \frac{r}{Mk}$, then $\forall f \in \mathcal{H}$,

$$L(f) \le \max\left\{\left(\frac{k}{k-1}\hat{L}_n(f)\right), \left(\hat{L}_n(f) + \frac{r}{Mk}\right)\right\}.$$

where $\bar{\mathcal{L}} = \left\{\frac{r}{\max(L(\ell_f^2), r)}\ell_f \middle| \ell_f \in \mathcal{L}\right\}$. Then, we propose the upper bounded of $\hat{U}_n(\bar{\mathcal{L}})$ with $R_n(\mathcal{L}^r)$: $\hat{U}_n(\bar{\mathcal{L}}) \le 2R_n(\mathcal{L}^r) + \sqrt{\frac{2r\ln\delta}{\lfloor n/2\rfloor}} + \frac{4\ln\delta}{3\lfloor n/2\rfloor}$. The above results show that we can choose a suitable r to satisfy the assumption $\hat{U}_n(\bar{\mathcal{L}}) \le \frac{r}{Mk}$ to accomplish this theorem. Finally, we show that the suitable r can be chosen with the fixed point of $R_n(\mathcal{L}^r)$. □

The $R_n(\mathcal{L}^r)$ is a sub-root function (see Lemma 3 in the supplementary material), so the fixed point r^* of $R_n(\mathcal{L}^r)$ is existing and unique.

Remark 1. *The generalization error bounds with the local Rademacher complexity for traditional classification or regression problems have been given in (Bartlett, Bousquet, and Mendelson 2005; Koltchinskii 2006). However, since the empirical ranking error $\hat{L}_n(f) = \frac{1}{n(n-1)}\sum_{i\neq j}\ell(f, z_i, z_j)$ is a non-sum-of-i.i.d. pairwise loss, the techniques for proving the generalization bounds in (Bartlett, Bousquet, and Mendelson 2005; Koltchinskii 2006) for classification and regression can not be applied to ranking. To address this problem, inspired by (Clémençon, Lugosi, and Vayatis 2005), we introduce permutations to convert the non-sum-of-i.i.d. pairwise loss to a sum-of-i.i.d. form. Thus, Theorem 1 is a non-trivial extension of (Bartlett, Bousquet, and Mendelson 2005; Koltchinskii 2006) for ranking.*

Tight Generalization Bounds

In the section, we first establish the relationship between the local ranking Rademacher complexity and the eigenvalues of integral operator of associated with kernel function, and further derive the sharp generalization bounds with fast convergence rates.

Note that K is a Mercer kernel, thus K can be written as $K(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty}\lambda_i\psi_i(\mathbf{x})\psi_i(\mathbf{x}')$, where $(\lambda_i)_{i=1}^{\infty}$ and $(\psi_i)_{i=1}^{\infty}$ are the sequence of eigenvalues and eigenfunctions of the integral operator $L_K : \mathcal{H} \to \mathcal{H}$

$$L_K(f) = \int_{\mathcal{X}} K(\cdot, \mathbf{x}')f(\mathbf{x}')\mathrm{d}P_{\mathcal{X}}(\mathbf{x}'), \tag{3}$$

where $P_{\mathcal{X}}$ is the marginal distribution of P on $\mathcal{X}$.

We say a loss function ℓ is an L-Lipschitz loss function, if the following inequality holds: $\forall f_1, f_2 \in \mathcal{H}$ and $\forall z, z' \in \mathcal{Z}$,

$$\begin{aligned}&|\ell(f_1, z, z') - \ell(f_2, z, z')|\\ &\le L\left(|f_1(\mathbf{x}) - f_2(\mathbf{x})| + |f_1(\mathbf{x}') - f_2(\mathbf{x}')|\right).\end{aligned}$$

Note that both the popular hinge loss and 0-1 loss are 1-lipschitz, the γ loss is $\frac{1}{\gamma}$-Lipschitz. Moreover, the square loss ranking is $(2F + \frac{4\kappa F}{\sqrt{\lambda}})$-Lipschitz (Agarwal and Niyogi 2009) if $\forall y \in \mathcal{Y}, |y| \le F$. Therefore, it is natural to assume that ℓ is an L-Lipschitz loss function.

The relationship between the eigenvalues of integral operator and $R_n(\mathcal{L}^r)$ is given as follows:

Theorem 2. *If the loss ℓ is an L-Lipschitz loss function, then*

$$R_n(\mathcal{L}^r) \le 2L\sqrt{\frac{2}{\lfloor n/2\rfloor}\min_{\theta\ge 0}\left(\frac{r}{4L^2}\theta + \sum_{i>\theta}\lambda_i\right)}.$$

Proof. We sketch the proof here. By exploiting the contraction inequality (Theorem 2.2 of Koltchinskii (2011)), we first establish the relationship between the local ranking Rademacher complexity and the traditional local Rademacher complexity:

$$R_n(\mathcal{L}^r) \le 2L\mathbb{E}_{D,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{H}_D}\left|\frac{2}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\sigma_i f(\mathbf{x}_i)\right|\right],$$

where $\mathcal{H}_D = \{f | f \in \mathcal{H}, \|f\|^2 \le \frac{r}{4L^2}\}$, and $D = \{z_i, \ldots, z_{\lfloor n/2\rfloor}\} \in Z^{\lfloor n/2\rfloor}$. According to Theorem 6.5 of (Bartlett, Bousquet, and Mendelson 2005), we can bound the traditional local Rademacher complexity with eigenvalues of integral operator, which finishes the proof. □

According to the above theorem, we know that we can use the tail eigenvalues $\sum_{i>\theta}\lambda_i$ to bound $R_n(\mathcal{L}^r)$. Therefore, we can derive tight generalization bounds under some mild assumptions over the eigenvalues of integral operator.

Theorem 3. *If the loss ℓ is an L-Lipschitz loss function, and kernel function K satisfies the assumption of algebraically decreasing eigenvalues, that is*

$$\exists\alpha > 1, c > 0 : \lambda_i \le ci^{-\alpha},$$

where $(\lambda_i)_{i=1}^{\infty}$ are the sequence of eigenvalues (arranged in descending order) of integral operator L_K defined in (3). Then, with probability $1 - \delta$, $\forall k \ge \max(1, \frac{\sqrt{2}}{2M})$, $L(f_S)$

$$\le \max\left\{\hat{L}_n(f_S) + c_1 k\left[\frac{1}{n-1}\right]^{\frac{\alpha}{\alpha+1}} + \frac{c_2 k + 6}{n-1}, \frac{k\hat{L}_n(f_S)}{k-1}\right\},$$

where $c_1 = \frac{32Mc}{\alpha-1}(4L^2)^{\alpha}$ and $c_2 = 8\ln\delta$.

Proof. We sketch the proof here. Based on the assumption of algebraically decreasing eigenvalues, we first prove that $\min_\theta \left(\frac{\theta r}{4L^2} + \sum_{j>\theta} \lambda_j\right) \leq \frac{2c}{\alpha-1}\left(\frac{r}{4L^2}\right)^{\frac{\alpha-1}{\alpha}}$. Substituting the above inequality into Theorem 2, we then obtain that $R_n(\mathcal{L}^r) \leq 2L\sqrt{\frac{1}{n-1}\frac{4c}{\alpha-1}\left(\frac{r}{4L^2}\right)^{\frac{\alpha-1}{\alpha}}}$. Finally, we estimate the fixed point of r^* of $R_n(\mathcal{L}^r)$ to complete the proof. □

The assumption of algebraically decreasing eigenvalues of kernel function is a common assumption, for example, met for the popular shift invariant kernel, finite rank kernels and convolution kernels (Williamson, Smola, and Scholkopf 2001).

Note that $f_S = \arg\min_{f\in\mathcal{H}} \hat{L}_n(f) + \lambda\|f\|_K^2$, so $\hat{L}_n(f_S)$ is dependent with n, which is usually assumed that $\hat{L}_n(f_S) = O(n^{-\beta})$, $\beta \geq \frac{1}{2}$ (Eberts and Steinwart 2011; Steinwart, Hush, and Scovel 2009). Thus, under this assumption, by Theorem 3, we have

$$\begin{aligned} L(f_S) - \hat{L}_n(f_S) &\leq \frac{n^{-\beta}}{k-1} + c_1 k\big[\frac{1}{n-1}\big]^{\frac{\alpha}{\alpha+1}} + \frac{c_2 k + 6}{n-1} \\ &= O\Big(\frac{n^{-\beta}}{k} + n^{-\frac{\alpha}{\alpha+1}} + \frac{k}{n}\Big). \end{aligned}$$

If we set $k = \Omega(\sqrt{n^{-\beta}} n^{\frac{\alpha}{2(\alpha+1)}})$, we can obtain that

$$L(f_S) - \hat{L}_n(f_S) = O(n^{-\frac{\alpha}{2(\alpha+1)} - \frac{\beta}{2}} + n^{-\beta - \frac{\alpha+2}{2(\alpha+1)}})$$

Note that $\alpha > 1$, $\beta \geq \frac{1}{2}$, so the rate is faster than $O(\frac{1}{\sqrt{n}})$.

If a slight stronger assumption on eigenvalues of integral operator is satisfied, the much faster rate of generalization bound can be obtained:

Theorem 4. *If the loss ℓ is an L-Lipschitz loss function, and kernel function K satisfies the assumption of factorial decreasing eigenvalues, that is*

$$\exists c > 0, \lambda_i \leq c\frac{1}{i!}.$$

Then, with probability at least $1-\delta$, $\forall k \geq \max(1, \frac{\sqrt{2}}{2M})$,

$$L(f_S) \leq \max\left\{\frac{k}{k-1}\hat{L}_n(f_S), \hat{L}_n(f_S) + \frac{c_1 k + 6}{n-1}\right\},$$

where $c_1 = 32M\theta^* + 8\ln\delta$ *and* θ^* *is the solution of* $\frac{\theta}{4L^2} = c\exp(-\theta)$ *with respect to* θ.

Proof. The proof is similar with that of Theorem 3. □

Although the assumption of the factorial decreasing eigenvalues is stronger than that of algebraically decreasing, it is also a mild condition met for the popular Gaussian kernel and Laplace kernel (Shi, Belkin, and Yu 2008; Cortes, Kloft, and Mohri 2013).

From Theorem 4, we have $L(f_S) - \hat{L}_n(f_S) \leq \frac{\hat{L}_n(f_S)}{k-1} + \frac{c_1 k+6}{n-1}$. Thus, $L(f_S) - \hat{L}_n(f_S) = O\left(\frac{\hat{L}_n(f_S)}{k} + \frac{k}{n}\right)$. Under the mild assumption that $\hat{L}_n(f_S) = O(n^{-\beta})$, $\beta \geq \frac{1}{2}$ (Eberts and Steinwart 2011; Steinwart, Hush, and Scovel 2009), when setting $k = \Omega(\sqrt{n\hat{L}_n(f_S)})$, we can obtain that

$$L(f_S) - \hat{L}_n(f_S) = O\Big(\sqrt{\hat{L}_n(f_S)/n}\Big) = O\Big(n^{-\frac{\beta+1}{2}}\Big).$$

Note that $\beta \geq \frac{1}{2}$, thus the order is faster than $O\big(\frac{1}{n^{3/4}}\big)$, which is much faster than $O\big(\frac{1}{\sqrt{n}}\big)$. When $\beta \geq 1$, the order is even faster than $O\big(\frac{1}{n}\big)$.

Comparison with Related Work

Generalization bounds based on the notion of algorithmic stability and Rademacher complexity are standard. (Agarwal and Niyogi 2009) showed that the regularized ranking algorithms in RKHS had good stability properties, and further derived a stability-based generalization bound for the regularized ranking algorithms: with probability at least $1-\delta$,

$$L(f_S) \leq \hat{L}_n(f_S) + \frac{32\kappa L}{\lambda n} + \left(\frac{16\kappa L}{\lambda} + M\right)\sqrt{\frac{2\log(1/\delta)}{n}}.$$

The convergence rate of the stability-based generalization bound is $O\big(\frac{1}{\sqrt{n}}\big)$.

In (Clémençon, Lugosi, and Vayatis 2008; 2005), they used the Rademacher average for deriving generalization bounds:

$$L(f_S) - \hat{L}_n(f_S) \leq MR_n(\mathcal{L}) + \sqrt{\frac{M^2\log(1/\delta)}{n-1}},$$

where $\mathfrak{R}_n(\mathcal{L})$ is Rademacher average over $\mathcal{L}$, which is in the order $O\big(\frac{1}{\sqrt{n}}\big)$ for various kernel classes in practice. Thus, this bound converges with rate $O\big(\frac{1}{\sqrt{n}}\big)$. In (Lan et al. 2008; 2009; Chen, Liu, and Ma 2010), they used the Rademacher average for deriving the "query-level generalization bounds" for the listwise ranking algorithms. These generalization bounds are established to study the generalization performance of listwise ranking algorithms, but not suitable for the regularized ranking algorithms in RKHS. Moreover, the convergence rate is also $O\big(\frac{1}{\sqrt{n}}\big)$.

Based on the VC dimension, (Freun et al. 2003) gave a generalization bound for RankBoost on the pairwise 0-1 loss: Assuming the function class $\mathcal{H}$ has a finite VC dimension, and the size of the positive and negative samples are equal, then with probability $1-\delta$:

$$L(f_S) \leq \hat{L}_n(f_S) + 4\sqrt{\frac{V'(\log\frac{2n}{V'} + 1) + \log\frac{18}{\delta}}{n}},$$

where $V' = 2(V+1)(T+1)\log(e(T+1))$ (refer to (Freun et al. 2003) for detail). One can see that if $\mathcal{H}$ has a finite VC dimension, the convergence rate can reach $O\big(\frac{1}{\sqrt{n}}\big)$. However, for the popular kernel function, such as Gaussian kernel, the VC dimension of its RKHS is infinite. Moreover, this bound is established for analyzing the generalization performance of bipartite document ranking.

The margin-based generalization bound based on the covering numbers of the hypothesis space is given in (Rudin and

Schapire 2009): with probability $1-\delta$

$$L(f_S) - \hat{L}_n(f_S) \leq \sqrt{\frac{4}{nE^2}\left[\frac{8\log|\mathcal{H}|}{\theta^2}\log\frac{4nE^2\theta^2}{\log|\mathcal{H}|} + 2\log\frac{2}{\delta}\right]},$$

where θ is the margin and E is a constant (refer to (Rudin and Schapire 2009) for detail). This bound converges with rate $O\left(\frac{1}{\sqrt{n}}\right)$.

Based on the empirical and U-process theory (Pakes and Pollard 1989), (Rejchel 2012) investigated the statistical properties of convex risk minimizers. Under the assumption on the covering number of the hypothesis space, $N(t, \mathcal{H}, \rho) \leq D_i t^{-V_i}$, and some other assumptions (see in (Rejchel 2012)), they proposed a generalization bound for the regularized ranking algorithms on the hinge loss (see (Rejchel 2012) in detail):

$$L(f_S) - \inf_{f\in\mathcal{H}} L(f) \leq c_1 \max\left(\frac{\log n}{n}, \frac{1}{n^\gamma}\right) + c_2\frac{c_3 + \log\delta}{n},$$

where $\gamma \in \left(\frac{2}{3}, 1\right)$. The rate of this bound is at most $O\left(\frac{1}{n^\gamma}\right)$. The assumption on the covering number of the hypothesis space is satisfied only for some special kernels, such as Gaussian kernel. For Gaussian kernel, the rate of our bound can reach $O(n^{-\frac{1+\beta}{2}})$, $\frac{1+\beta}{2} \geq \frac{3}{4}$.

The above theoretical analysis indicates that it is a good choice to use the integral operator to analyze the generalization ability of ranking.

Simulation Studies

In this subsection, we will compare our bounds with the standard stability-based bound (Agarwal and Niyogi 2009), Rademacher bound (Clémençon, Lugosi, and Vayatis 2005) and the state-of-the-art bound (called U-process bound) (Rejchel 2012) on simulated datasets. For all experiments, we simulate data from the true ranking model $y = f^*(x) + \varepsilon$ for $x \in [0,1]$, where $f^*(x) = x^2$, the noise variables $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ are normally distributed with variance $\sigma^2 = 0.1$, and the samples $x_i \sim \mathrm{Uni}[0,1]$. We use the the regularized ranking algorithms on Hinge loss. The regularization parameter is set to be $\lambda = 0.01$ for all experiments[1].

In the first experiment, we use the Fourier kernel $K(x,x') = \sum_{i=1}^{20}\frac{1}{i^3}\psi_i(x)\psi_i(x')$, where ψ_i is the Fourier orthonormal basis. Note that this Fourier kernel satisfies the assumption of algebraically decreasing eigenvalues, and the eigenvalue decay rate $\alpha = 3$. Our eigenvalue-based bound in Theorem 3, the stability-based bound (Agarwal and Niyogi 2009) and the Rademacher-based bound (Clémençon, Lugosi, and Vayatis 2005) with Fourier kernel are plotted in Figure 2 [2]. The plot shows that the convergence rate of our bound is faster than that of the stability-based and Rademacher-based bounds, which conforms to our theoretical analysis.

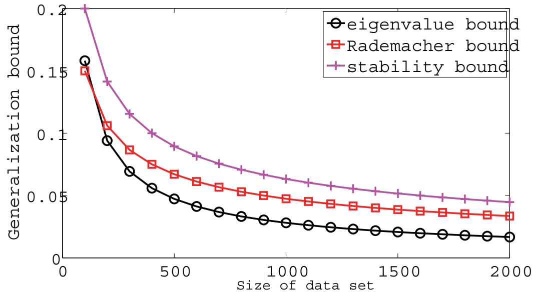

Figure 2: Our eigenvalue-based bound of Theorem 3, the stability-based bound (Agarwal and Niyogi 2009) and Rademacher bound (Clémençon, Lugosi, and Vayatis 2005). The kernel used in this figure is the Fourier kernel $K(x,x') = \sum_{i=1}^{20}\frac{1}{i^3}\psi_i(x)\psi_i(x')$.

In the second experiment, we using the Gaussian kernel: $K(x,x') = \exp\left(-\frac{1}{2}|x-x'|^2\right)$, which satisfies the assumption of factorial decreasing eigenvalues. Note that $\inf_{f\in\mathcal{H}} L(f)$ in the U-process bound is not computable, in this experiment[3], we set $\inf_{f\in\mathcal{H}} L(f) = 0$. Our eigenvalue-based bound (in Theorem 4), stability-based bound, Rademacher-based bound and U-process bound (Rejchel 2012) with Gaussian kernel are plotted in Figure 1. We can find that the convergence rate of our bound is the fastest.

The results agree with our theoretical findings, and also demonstrate the effectiveness of using eigenvalues of integral operator for deriving generalization bounds.

Conclusion

In this paper, we propose sharp generalization bounds via the local Rademacher complexity and integral operator for the regularized ranking algorithms in reproducing kernel Hilbert space (RKHS). The order of the proposed bound is much faster than $O\left(\frac{1}{\sqrt{n}}\right)$, while for most of existing bounds, the order are at most $O\left(\frac{1}{\sqrt{n}}\right)$.

In future work, we will design the novel ranking algorithms of fast convergence rate based on the theoretical results of this paper.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under grant No. 61602467, No. 61303056 and Youth Innovation Promotion Association, CAS, No.2016146.

References

Agarwal, S., and Niyogi, P. 2005. Stability and generalization of bipartite ranking algorithms. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT 2005)*, 32–47.

[1] According to the setting in the experiments, we know that $L = 1$, $M = 1$. In our bounds, we set $k = \log(n)$ for all experiments.

[2] Since the U-process bound is satisfied only for Gaussian kernel, so we only compare our bound with the stability bound and Rademacher bound for Fourier kernel in this experiment.

[3] In this experiment, we set $\gamma = 1$ for U-process bound and set $\beta = 1$ for our bound.

Agarwal, S., and Niyogi, P. 2009. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research* 10:441–474.

Aronszajn, N. 1950. Theory of reproducing kernels. *Transactions of the American Mathematical Society* 68:337–404.

Bartlett, P. L., and Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3:463–482.

Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2005. Local Rademacher complexities. *The Annals of Statistics* 33(4):1497–1537.

Bousquet, O., and Elisseeff, A. 2002. Stability and generalization. *Journal of Machine Learning Research* 2:499–526.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, 89–96. ACM.

Burges, C. J.; Ragno, R.; and Le, Q. V. 2007. Learning to rank with nonsmooth cost functions. *Advances in Neural Information Processing Systems 20 (NIPS 2007)* 19:193–200.

Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; and Li, H. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine learning (ICML 2007)*, 129–136. ACM.

Chen, W.; Liu, T.-Y.; and Ma, Z. 2010. Two-layer generalization analysis for ranking using rademacher average. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, 370–378.

Clémençon, S.; Lugosi, G.; and Vayatis, N. 2005. Ranking and scoring using empirical risk minimization. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT, 2005)*, 1–15. Springer.

Clémençon, S.; Lugosi, G.; and Vayatis, N. 2008. Ranking and empirical minimization of U-statistics. *The Annals of Statistics* 36:844–874.

Cortes, C.; Kloft, M.; and Mohri, M. 2013. Learning kernels using local Rademacher complexity. In *Advances in Neural Information Processing Systems 25 (NIPS 2013)*, 2760–2768.

Cortes, C.; Mohri, M.; and Rastogi, A. 2007. Magnitude-preserving ranking algorithms. In *Proceedings of the 24th International Conference on Machine learning (ICML 2007)*, 169–176. ACM.

Cossock, D., and Zhang, T. 2008. Statistical analysis of bayes optimal subset ranking. *IEEE Transactions on Information Theory* 54(11):5140–5154.

Crammer, K., and Singer, Y. 2001. Pranking with ranking. In *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, 641–647.

Eberts, M., and Steinwart, I. 2011. Optimal learning rates for least squares svms using gaussian kernels. In *Advances in Neural Information Processing Systems 25 (NIPS 2011)*, 1539–1547.

Freun, Y.; Iyer, R.; Schapire, R. E.; and Singer, Y. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research* 4:933–969.

Herbrich, R.; Graepel, T.; and Obermayer, K. 1999. Large margin rank boundaries for ordinal regression. In *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, 115–132.

Koltchinskii, V., and Panchenko, D. 2002. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics* 30:1–50.

Koltchinskii, V. 2006. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics* 34(6):2593–2656.

Koltchinskii, V. 2011. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Lan, Y.; Liu, T.-Y.; Qin, T.; Ma, Z.; and Li, H. 2008. Query-level stability and generalization in learning to rank. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, 512–519.

Lan, Y.; Liu, T.-Y.; Ma, Z.; and Li, H. 2009. Generalization analysis of listwise learning-to-rank algorithms. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, 577–584. ACM.

Liu, Y., and Liao, S. 2015. Eigenvalues ratio for kernel selection of kernel methods. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*, 2814–2820.

Pakes, A., and Pollard, D. 1989. Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society* 57:1027–1057.

Rejchel, W. 2012. On ranking and generalization bounds. *Journal of Machine Learning Research* 13(1):1373–1392.

Rudin, C., and Schapire, R. E. 2009. Margin-based ranking and an equivalence between adaboost and rankboost. *Journal of Machine Learning Research* 10:2193–2232.

Rudin, C. 2009. The P-Norm Push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research* 10:2233–2271.

Shi, T.; Belkin, M.; and Yu, B. 2008. Data spectroscopy: Learning mixture models using eigenspaces of convolution operators. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 936–943. ACM.

Steinwart, I.; Hush, D.; and Scovel, C. 2009. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Conference on Learning Theory (COLT 2009)*, 79–93.

Williamson, R.; Smola, A.; and Scholkopf, B. 2001. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory* 47(6):2156–2132.

Xia, F.; Liu, T.-Y.; Wang, J.; Zhang, W.; and Li, H. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th International Conference on Machine learning (ICML 2008)*, 1192–1199. ACM.