Generalization Performance of Regularized Ranking With Multiscale Kernels

Yicong Zhou, *Senior Member, IEEE*, Hong Chen, Rushi Lan, and Zhibin Pan

***Abstract*— The regularized kernel method for the ranking problem has attracted increasing attentions in machine learning. The previous regularized ranking algorithms are usually based on reproducing kernel Hilbert spaces with a single kernel. In this paper, we go beyond this framework by investigating the generalization performance of the regularized ranking with multiscale kernels. A novel ranking algorithm with multiscale kernels is proposed and its representer theorem is proved. We establish the upper bound of the generalization error in terms of the complexity of hypothesis spaces. It shows that the multiscale ranking algorithm can achieve satisfactory learning rates under mild conditions. Experiments demonstrate the effectiveness of the proposed method for drug discovery and recommendation tasks.**

***Index Terms*— Drug discovery, generalization performance, multiscale kernel, ranking, recommendation tasks, reproducing kernel Hilbert space (RKHS).**

I. Introduction

MACHINE learning methods for ranking have attracted more and more attentions in information retrieval and search engines. From the viewpoint of machine learning, the goal of ranking is to search a score function such that the predicted orders of two inputs are consistent as possible as the true relations. Following this purpose, many ranking algorithms have been proposed from different perspectives. Examples include ranking support vector machine (RankSVM) [20], [22], RankNet [4], [5], RankBoost [18], MPRank [12], [13], RankRLS [26], [27], the gradient descent ranking algorithms [8], [28], the P-norm push ranking [31], and the multiparte ranking [23].

Among these algorithms, the regularized ranking scheme with least square ranking loss has been used widely in various ranking tasks [1], [6], [8], [12], [13]. Similar to the least square regularized regression in [14], [16], and [41], these regularized ranking algorithms depend heavily on the reproducing kernel Hilbert spaces (RKHSs) associated with Mercer kernels. It is well known that the flat nonlinear functions can be well approximated by RKHSs with Gaussian kernels. However, as shown in [41] and [45], it is unsuitable to use the single Gaussian kernel to approximate the nonflat functions including the smooth and steep variations. Under this setting, the learned function is difficult to approximate the smooth and steep portions simultaneously.

Multikernel methods have been used successfully in many fields of learning systems [17], [34], [40]. In the regression problem, there are extensive studies on experiments and theory for the nonflat function approximation [7], [41], [42], [45]. These studies show that the multiscale kernel methods are more efficient than the corresponding algorithms with a single kernel. The candidates of multiscale kernels could be Gaussian kernels with different widths, frame-based kernels, and wavelet-based kernels. For the ranking task, the multiscale kernel can achieve better performance than the single kernel when the intrinsic optimal predictor is a nonflat function. However, for the regularized ranking problem in RKHSs, as we know, there is no work on this multiscale theme for algorithm design and generalization analysis.

It is well known that generalization performance is an important measure to evaluate the learning ability of machine learning algorithms [16], [24], [25], [35], [36], [43]. Recently, generalization error analysis also attracts increasing attentions in the ranking problem. The techniques of error analysis for ranking problem mainly include stability analysis in [1] and [11]–[13], uniform convergence estimate based on the capacity of hypothesis spaces [10], [19], [29], [30], [44], and approximation estimate based on the operator approximation in [6] and [8]. Although these progresses on generalization analysis have been made, the theoretical results for these regularized kernel methods only focus on ranking in a single-kernel-based RKHS [1], [6], [8], [12], [13].

Actually, the target ranking function with high- and low-frequency components can be approximated by small- and large-scale kernels, respectively. Inspired by the studies of multiscale kernels in [41], we propose a multiscale least square regularized ranking (MLSRRank) algorithm to realize efficient ranking, which extends the regularized ranking in [12] to the multikernel setting. The estimates of the generalization error are established based on the covering numbers of multiple-kernel-based RKHSs. In particular, the analysis of learning rates is given for the RKHSs with Gaussian kernels.

Manuscript received June 26, 2013; revised January 7, 2015 and May 7, 2015; accepted May 16, 2015. Date of publication June 5, 2015; date of current version April 15, 2016. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2662015PY046 and Grant 2014PY025, in part by NSFC under Grant 11001092 and Grant 61300143, in part by the Macau Science and Technology Development Fund under Grant FDCT/106/2013/A3, and in part by the Research Committee at University of Macau under Grant MYRG2014-00003-FST, Grant MRG017/ZYC/2014/FST, Grant MYRG113(Y1-L3)-FST12-ZYC, and Grant MRG001/ZYC/2013/FST. *(Yicong Zhou and Hong Chen equally contributed to this work.) (Corresponding author: Hong Chen.)*

Y. Zhou and R. Lan are with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@umac.mo; rslan@umac.mo).

H. Chen and Z. Pan are with the Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan 430070, China (e-mail: chenh@mail.hzau.edu.cn; zhibinpan2008@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2015.2434887

The experimental studies on several data sets demonstrate the effectiveness of our proposed algorithm. To the best of our knowledge, this paper is among the first endeavors on generalization performance analysis for the multiscale ranking.

In summary, the main contributions of this paper are listed as follows.

1) A novel multiscale framework for ranking is proposed, and its representer theorem is proved. This theorem provides a simple and fast way to implement the proposed algorithm.
2) Generalization error analysis of MLSRRank is established in terms of the capacity of hypothesis spaces. From our error analysis, we know that the MLSRRank can reach satisfactory learning rates under mild conditions.
3) Experimental evaluations on benchmark data sets demonstrate the effectiveness of MLSRRank. This extends the effectiveness analysis of the multiscale kernel for regression [41] to the ranking setting.

The rest of this paper is organized as follows. In Section II, we introduce some necessary background of ranking, and propose MLSRRank. The representer theorem is also given. In Section III, the theoretical results on generalization performance are stated. In Section IV, we present the experimental results on the real-world data sets. Finally, we close this paper with a brief conclusion in Section V.

II. Multiscale Regularized Ranking Algorithm

Now, we recall some basic concepts of ranking (see [1] and references therein for details). Let $\mathcal{X} \in \mathbb{R}^d$ be a compact metric space and $\mathcal{Y} = [0, M]$ for some $M > 0$. The relation between the input $x \in \mathcal{X}$ and the output $y \in \mathcal{Y}$ is described by a probability distribution $\rho(x, y)$ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Given samples $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in \mathcal{Z}^m$ independently drawn according to ρ, the ranking problem aims at finding a function $f : \mathcal{X} \to \mathbb{R}$ that ranks future input instances with larger labels higher than those with smaller labels. x is to be ranked preferred over x' if $y - y' > 0$ and lower than x' if $y - y' < 0$. In particular, $y - y' = 0$ indicates no ranking preference between the two input instances.

For any $z = (x, y)$, $z' = (x', y')$ in $\mathcal{Z}$, we consider the least square ranking loss

$$\ell(f, z, z') = (y - y' - (f(x) - f(x')))^2. \tag{1}$$

The expected risk of a ranking function f is defined as

$$\mathcal{E}(f) = \int_{\mathcal{Z}} \int_{\mathcal{Z}} (y - y' - (f(x) - f(x')))^2 d\rho(x, y) d\rho(x', y'). \tag{2}$$

The algorithms discussed in this paper are based on a Tikhonov regularization scheme associated with a Mercer kernel. We usually call a symmetric and positive semidefinite continuous function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a Mercer kernel. The RKHS $\mathcal{H}_K$ associated with the kernel K is defined (see [3]) to be the closure of the linear span of the set of functions $\{K(x, \cdot) : x \in \mathcal{X}\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ given by $\langle K(x, \cdot), K(x', \cdot) \rangle_K = K(x, x')$. The reproducing property of RKHS tells us that $f(x) = \langle f, K(x, \cdot) \rangle_K$.

A. Least Square Regularized Ranking Algorithm

Given $\mathbf{z}$ and a ranking function f, the empirical ranking risk is defined as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m^2} \sum_{i,j=1}^{m} (y_i - y_j - (f(x_i) - f(x_j)))^2. \tag{3}$$

Observe that $\mathcal{E}_{\mathbf{z}}(f)$ in (3) can be considered as the discrete version of $(m - 1/m)\mathcal{E}(f)$. That is to say $E\mathcal{E}_{\mathbf{z}}(f) = (m - 1/m)\mathcal{E}(f)$ according to definitions (2) and (3).

The LSRRank algorithm is defined as the minimizer in $\mathcal{H}_K$ [1], [6], [8]

$$f_{\mathbf{z}} := f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_K^2 \right\} \tag{4}$$

where $\lambda > 0$ is the regularization parameter.

In fact, LSRRank is a special case of MPRank in [12] with the least square ranking loss defined in (1). To highlight the feature of this loss function, we call algorithm (4) LSRRank. In addition, the difference between LSRRank and RankRLS is that the former considers the empirical risk associated with all input pairs, while the latter only includes the input pairs that are relevant to the task in the question [27].

Denote $D = mI - \mathbf{1}\mathbf{1}^T$, where I is an m-order unit matrix and $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^m$. Meanwhile, denote $Y = (y_i)_{i=1}^m = (y_1, \ldots, y_m)^T$ and denote $\mathbf{K}$ as an m-order matrix whose (i, j) entry is $K(x_i, x_j)$.

By means of the properties of RKHSs, we can get the representer theorem for the minimizer $f_{\mathbf{z}}$ in (4). A similar result can be found in [6]. For completeness, we present the proof in Appendix A.

Lemma 1: The minimizer $f_{\mathbf{z}}$ in (4) can be represented as

$$f_{\mathbf{z}}(x) = \sum_{i=1}^{m} \alpha_{\mathbf{z},i} K(x_i, x)$$

where $\alpha_{\mathbf{z}} = (\alpha_{\mathbf{z},1}, \ldots, \alpha_{\mathbf{z},m})^T \in \mathbb{R}^m$ is the unique solution of the linear system

$$\left(D\mathbf{K} + \frac{m^2}{2}\lambda I \right) \alpha = DY. \tag{5}$$

Set $\tilde{y}_i = 2y_i - (2/m)\sum_{k=1}^m y_k$ and $\tilde{K}(x_i, x_j) = 2K(x_i, x_j) - (2/m)\sum_{k=1}^m K(x_i, x_k)$. From linear system (5) in Lemma 1, we can get

$$(\tilde{\mathbf{K}} + m\lambda I)\alpha = \tilde{Y}.$$

An interesting observation is that this equation is externally similar to the least square regularized regression in [41]. In fact, some relationships are also given in [21] for the expected risks between ranking and regression.

B. Multiscale Least Square Regularized Ranking Algorithm

Although the ranking algorithm (4) is used widely in the ranking problem, it might be difficult to search the target function with the high- and low-frequency components simultaneously. The same difficulty for regression algorithms is overcome by learning the frameworks with multiscale kernels [41], [42]. In this paper, we use the idea of multiscale kernels to propose an algorithm for learning a ranking function.

The MLSRRank algorithm is implemented under a regularized framework in a sum space of RKHSs. Define the sum space of $\mathcal{H}_{K_t}$ for $t = 1, \ldots, l$ as

$$\mathcal{H}_{\oplus} = \mathcal{H}_{K_1} \oplus \cdots \oplus \mathcal{H}_{K_l} = \left\{ f : f = \sum_{t=1}^{l} f_t,\ f_t \in \mathcal{H}_{K_t} \right\}$$

where K_t is a Mercer kernel. For $f \in \mathcal{H}_{\oplus}$ and $\upsilon \in V = \{(\upsilon_1, \ldots, \upsilon_l)^T \,|\, \min\{\upsilon_t, t = 1, \ldots, l\} = 1\}$, we define a norm $\|\cdot\|_{\oplus,\upsilon}$ on $\mathcal{H}_{\oplus}$ as

$$\|f\|_{\oplus,\upsilon}^2 = \min \left\{ \sum_{t=1}^{l} \upsilon_t \|f_t\|_{K_t}^2 : f = \sum_{t=1}^{l} f_t,\ f_t \in \mathcal{H}_{K_t} \right\}.$$

For the given sample $\mathbf{z}$ and the regularization parameter λ, the MLSRRank is defined as the minimizer of the following regularized framework:

$$f_{\mathbf{z},\oplus} := f_{\mathbf{z},\lambda,\oplus,\upsilon} = \operatorname*{arg\,min}_{f \in \mathcal{H}_{\oplus}} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_{\oplus,\upsilon}^2 \right\}. \qquad (6)$$

Denote

$$A = \begin{pmatrix} m\lambda\upsilon_1 I + \frac{2}{m} D\mathbf{K}_1 & \ldots & \frac{2}{m} D\mathbf{K}_l \\ \ldots & \ldots & \ldots \\ \frac{2}{m} D\mathbf{K}_1 & \ldots & m\lambda\upsilon_l I + \frac{2}{m} D\mathbf{K}_l \end{pmatrix}$$

where $\mathbf{K}_t$ is an m-order matrix whose (i, j) entry is $K_t(x_i, x_j)$.

In terms of the properties of RKHSs, we can get the expression of $f_{\mathbf{z},\oplus}$ as follows. The proof can be found in Appendix B.

Theorem 1: The minimizer $f_{\mathbf{z},\oplus}$ in (6) can be represented as $f_{\mathbf{z},\oplus} = \sum_{t=1}^{l} f_t$, where

$$f_t = \sum_{i=1}^{m} \alpha_{\mathbf{z},i}^t K_t(x_i, x), \quad t = 1, \ldots, l$$

and $\alpha_{\mathbf{z}}^t = (\alpha_{\mathbf{z},1}^t, \ldots, \alpha_{\mathbf{z},m}^t)^T \in \mathbb{R}^m$ is the unique solution of the linear system

$$A \begin{pmatrix} \alpha^1 \\ \ldots \\ \alpha^l \end{pmatrix} = \begin{pmatrix} \frac{2}{m} DY \\ \ldots \\ \frac{2}{m} DY \end{pmatrix}. \qquad (7)$$

From Theorem 1, we know that MLSRRank can be implemented easily through the above linear system. The effects of different scale kernels are adjusted by parameters υ_t, $t = 1, \ldots, l$.

III. Generalization Error Analysis

In this section, we will estimate the upper bound of the generalization error for $f_{\mathbf{z},\oplus}$. In order to adapt to more general ranking tasks, first, we present the generalization error analysis for the general convex losses. Then, we apply the derived result to obtain the generalization analysis of MLSRRank.

For analysis, we introduce the following condition of the loss function used in [1] and [44].

Definition 1: Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$ and let $L > 0$. We say that the ranking loss ℓ is L-admissible with respect to $\mathcal{F}$ if for all $g_1, g_2 \in \mathcal{F}, z, z' \in \mathcal{Z}$

$$|\ell(g_1, z, z') - \ell(g_2, z, z')| \le L(|g_1(x) - g_2(x)| + |g_1(x') - g_2(x')|).$$

It is worth noting that the least square ranking loss $(y - y' - (f(x) - f(x')))^2$ is $(2M + 4r)$ admissible with respect to the uniform-bounded function set $\{f \in \mathcal{F} : \|f\|_\infty \le r\}$.

For the general loss function ℓ, we denote the expected risk as

$$\mathcal{E}^\ell(f) = \int_{\mathcal{Z}} \int_{\mathcal{Z}} \ell(f, z, z') d\rho d\rho$$

and denote the corresponding empirical risk as

$$\mathcal{E}_{\mathbf{z}}^\ell(f) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \ell(f, z_i, z_j).$$

Based on the MLSRRank (6), the multiscale algorithm with the general loss ℓ can be written as

$$f_{\mathbf{z},\oplus}^\ell := f_{\mathbf{z},\gamma,\oplus,\upsilon}^\ell = \operatorname*{arg\,min}_{f \in \mathcal{H}_{\oplus}} \left\{ \mathcal{E}_{\mathbf{z}}^\ell(f) + \gamma \|f\|_{\oplus,\upsilon}^2 \right\} \qquad (8)$$

where $\gamma > 0$ is a regularization parameter.

Remark 1: For $\ell(f, z, z') = (y - y' - (f(x) - f(x')))^2$, we can see that $\mathcal{E}_{\mathbf{z}}(f) = (m - 1/m)\mathcal{E}_{\mathbf{z}}^\ell(f)$ and $\mathcal{E}(f) = \mathcal{E}^\ell(f)$. Then, $f_{\mathbf{z},\oplus} = f_{\mathbf{z},\gamma,\oplus,\upsilon}^\ell$ for $\gamma = (m/m - 1)\lambda$. Hence, the analysis result of $\mathcal{E}^\ell(f_{\mathbf{z},\oplus}^\ell)$ can be used to estimate $\mathcal{E}(f_{\mathbf{z},\oplus})$.

We estimate the excess error $\mathcal{E}^\ell(f_{\mathbf{z},\oplus}^\ell) - \mathcal{E}^\ell(f^*)$, where f^* is the minimizer of $\mathcal{E}^\ell(f)$ over all measurable functions.

Now, some necessary notations are introduced. Denote the regularization functions in $\mathcal{H}_{K_t}$ $(t = 1, 2 \ldots l,)$ and $\mathcal{H}_{\oplus}$ as

$$f_{\gamma,K_t}^\ell = \operatorname*{arg\,min}_{f \in \mathcal{H}_{K_t}} \left\{ \mathcal{E}^\ell(f) + \gamma \|f\|_{K_t}^2 \right\}$$

and

$$f_{\gamma,\oplus}^\ell = \operatorname*{arg\,min}_{f \in \mathcal{H}_{\oplus}} \left\{ \mathcal{E}^\ell(f) + \gamma \|f\|_{\oplus,\upsilon}^2 \right\}.$$

Denote the corresponding regularization errors as

$$D_{K_t}(\gamma) = \min_{f \in \mathcal{H}_{K_t}} \left\{ \mathcal{E}^\ell(f) - \mathcal{E}^\ell(f^*) + \gamma \|f\|_{K_t}^2 \right\}$$

and

$$D_{\oplus,\upsilon}(\gamma) = \min_{f \in \mathcal{H}_{\oplus}} \left\{ \mathcal{E}^\ell(f) - \mathcal{E}^\ell(f^*) + \gamma \|f\|_{\oplus,\upsilon}^2 \right\}.$$

The following error decomposition for the excess error can be obtained in terms of the definitions of $f_{\mathbf{z},\oplus}^\ell$ and $f_{\gamma,\oplus}^\ell$.

Proposition 1: Following the definitions of $f_{\mathbf{z},\oplus}^\ell$ and f^*, we have:

$$\mathcal{E}^\ell(f_{\mathbf{z},\oplus}^\ell) - \mathcal{E}^\ell(f^*) \le \left\{ \mathcal{E}^\ell(f_{\mathbf{z},\oplus}^\ell) - \mathcal{E}_{\mathbf{z}}^\ell(f_{\mathbf{z},\oplus}^\ell) \right\} + \left\{ \mathcal{E}_{\mathbf{z}}^\ell(f_{\gamma,\oplus}^\ell) - \mathcal{E}^\ell(f_{\gamma,\oplus}^\ell) \right\} + D_{\oplus,\upsilon}(\gamma). \qquad (9)$$

In the remainder of this section, we will focus on the estimates of the first and second terms on the right-hand side of (9).

With the same technique used in [44], we can obtain the estimate of $\mathcal{E}^{\ell}_{\mathbf{z}}(f^{\ell}_{\gamma,\oplus}) - \mathcal{E}^{\ell}(f^{\ell}_{\gamma,\oplus})$ by applying the McDiarmid's inequality. The proof is presented in Appendix C. In fact, the similar result also can be derived by the Hoeffding's inequality of U-statistic [10].

Proposition 2: Let $\kappa = \sup_{x,x' \in \mathcal{X}, t \in \{1,\ldots,l\}} K_t(x,x')$ and $\upsilon \in V = \{(\upsilon_1,\ldots,\upsilon_l)^T \mid \min\{\upsilon_t, t=1,\ldots,l\} = 1\}$. For any $0 < \delta \leq 1$, there holds with confidence at least $1-\delta$

$$\mathcal{E}^{\ell}_{\mathbf{z}}(f^{\ell}_{\gamma,\oplus}) - \mathcal{E}^{\ell}(f^{\ell}_{\gamma,\oplus}) \leq M_1 \sqrt{\frac{2\ln(1/\delta)}{m}}$$

where $M_1 = \sup_{z,z' \in Z, \|f\|_\infty \leq \kappa (l D_{\oplus,\upsilon}(\gamma)/\gamma)^{1/2}} |\ell(f,z,z')|$.

In order to derive the estimate of $\mathcal{E}^{\ell}(f^{\ell}_{\mathbf{z},\oplus}) - \mathcal{E}^{\ell}_{\mathbf{z}}(f^{\ell}_{\mathbf{z},\oplus})$, we need to give the complexity measure of $\mathcal{H}_\oplus$. Here, we use the covering number (see [14], [16], [46]) to measure the capacity of the sum space.

Definition 2: For $\epsilon > 0$, the covering number $\mathcal{N}(\mathcal{H},\epsilon)$ is defined to be the smallest integer $l \in \mathbb{N}$ such that there exist l disks in $C(\mathcal{X})$ with the radius ϵ and centers in $\mathcal{H}$ covering the set $\mathcal{H}$.

Denote $B_{K_t,R} = \{f : f \in \mathcal{H}_{K_t}, \|f\|_{K_t} \leq R\}$ and $B_{\oplus,\upsilon,R} = \{f : f \in \mathcal{H}_\oplus, \|f\|_{\oplus,\upsilon} \leq R\}$.

The following concentration estimation on $\mathcal{H}_\oplus$ will be proved in Appendix D.

Proposition 3: For $R = (M/\sqrt{\gamma})$, denote $M_2 = \sup_{z,z' \in \mathcal{Z}, f \in B_{\oplus,\upsilon,R}} |\ell(f,z,z')|$. For all $\varepsilon > 0$

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m}\left\{\mathcal{E}^{\ell}(f^{\ell}_{\mathbf{z},\oplus}) - \mathcal{E}^{\ell}_{\mathbf{z}}(f^{\ell}_{\mathbf{z},\oplus}) > \varepsilon\right\} \leq \mathcal{N}\left(B_{\oplus,\upsilon,R}, \frac{\varepsilon}{4L}\right)\exp\left\{-\frac{m\varepsilon^2}{8M_2^2}\right\}$$

where $L = 2M + 4\kappa R$.

It is the position to present the error estimate of $f^{\ell}_{\mathbf{z},\oplus}$. The proof can be found in Appendix E.

Theorem 2: Let $\mathcal{H}_\oplus$ be the sum space of RKHSs with Gaussian kernels $K_t(x,x') = \exp\{-\|x - x'\|/\mu_t^2\}$ for $t = 1,\ldots,l$ and $\mu_1 < \mu_2 < \cdots < \mu_l$. For any $0 < \delta < 1$, $\mathcal{E}^{\ell}(f^{\ell}_{\mathbf{z},\oplus}) - \mathcal{E}^{\ell}(f^*)$ can be bounded by

$$M_1\sqrt{\frac{2\ln(1/\delta)}{m}} + \max\left\{2M_2\sqrt{\frac{C_d l \mu_1^{-2(d+1)} + \ln(1/\delta)}{m}} + \left(4l\kappa L \frac{M}{\sqrt{\gamma}}\right)^{\frac{s}{2+s}} \left(\frac{4C_d l M_2^2}{m}\right)^{\frac{1}{2+s}}\right\} + D_{\oplus,\upsilon}(\gamma)$$

with confidence at least $1-\delta$. Here, C_d is a parameter depending on the dimension d, and $s > 0$ is a parameter that can be close to 0.

Based on Remark 1, we can get the following convergence analysis from Theorem 2 directly.

Theorem 3: Let $\mathcal{H}_\oplus$ be the sum space of RKHSs with Gaussian kernels $K_t(x,x') = \exp\{-\|x - x'\|/\mu_t^2\}$ for $t = 1,\ldots,l$ and $\mu_1 < \mu_2 < \cdots < \mu_l$. Assume that $D_{\oplus,\upsilon}((m - 1/m\lambda) \leq c_\beta \lambda^\beta$ for some $0 < \beta < 1$. Choosing $\lambda = m^{-\epsilon}$, for any $0 < \delta < 1$, we have

$$\mathcal{E}(f_{\mathbf{z},\oplus}) - \mathcal{E}(f^*) \leq C m^{-\min\left\{\beta\epsilon, \frac{1}{2} - (1-\beta)\epsilon, \frac{(1-2\epsilon)}{2}, \frac{1}{2+s} - \frac{8+3s}{4+2s}\epsilon\right\}}$$

with confidence at least $1-\delta$, where the constant C is independent of m and $s > 0$ is a parameter that can be arbitrarily small.

Remark 2: MLSRRank can achieve the convergence rate $O(m^{-\min\{\epsilon,(1/2+s)-(8+3s/4+2s)\epsilon\}})$ when $\beta \to 1$. In particular, choosing $\epsilon = (2/12 + 7s)$, and $s = 1$, the convergence rate $O(m^{-(1/6)})$ can be reached. The polynomial decay rate is satisfactory for the machine learning algorithms.

Recently, a nice generalization bound is obtained for the kernel-based regularized ranking [29, Example 3]. Although our analysis cannot reach this fast convergence rate, it is based on a different loss function. Moreover, our result does not require the variance-expectation bound condition for the loss function and distribution in [29].

Remark 3: Denote by $\mathscr{F}$ the measurable function space, and define $\mathscr{G} = \{f^* \in \mathscr{F} : f^* = \arg\min_{f \in \mathscr{F}} \mathcal{E}(f)\}$ as the target function set. We can observe that the target function f^* is not unique and the regression function $f_\rho \in \mathscr{G}$ (see [8], [21]), where

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), \quad x \in \mathcal{X}.$$

It is easy to deduce that

$$\begin{aligned} D_{\oplus,\upsilon}(\lambda) &= \min_{f \in \mathcal{H}_\oplus}\left\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|^2_{\oplus,\upsilon}\right\} \\ &\leq 4 \min_{f \in \mathcal{H}_\oplus}\left\{\int (f(x) - f_\rho(x))^2 d\rho + \lambda\|f\|^2_{\oplus,\upsilon}\right\}. \end{aligned}$$

Hence, the assumption of $D_{\oplus,\upsilon}(\lambda)$ in Theorem 2 is consistent with the previous approximation assumption in the least square regression [7], [16], [37], [38], [41], [42]. Moreover, the derived learning rate here can be improved by the iterative technique in [7], [32], [37], and [38].

Remark 4: The error analysis established here is inspired from the theoretical studies in [41] for the least square regularized regression. There is a key difference between our work and that in [41]. In the ranking problem, the empirical risk is measured on pairs of samples and cannot be expressed as a sum of independent random variables. Hence, the concentration inequalities used in [41] for the regression problem cannot be used to characterize the convergence of ranking directly. The theoretical analysis of ranking is more complicated than the regression setting.

IV. Experiments

In this section, we evaluate MLSRRank on several benchmark data sets on the drug discovery and recommendation tasks. The experimental results show that the MLSRRank can achieve competitive performance compared with several state-of-the-art ranking algorithms.

A. Algorithm and Parameter Selection

From Theorem 1, we know that the explicit computation steps of MLSRRank can be summarized in Algorithm 1.

In the experiments, we choose the Gaussian kernels with different widths as the candidate multiscale kernels. Here, we denote the width of a Gaussian kernel by σ and denote the regularization parameter by λ. The width parameter σ

```
Algorithm 1 Multiscale Least Square Regularized Ranking Algorithm (MLSRRank)
Require:
    Training set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, kernel functions $K_t$, a regularization parameter $\lambda > 0$, and $\upsilon \in V = \left\{(\upsilon_1, \ldots, \upsilon_l)^T \mid \min\{\upsilon_t, t = 1, \ldots, l\} = 1\right\}$.
Computing the necessary matrices:
    $D = mI - \mathbf{1}\mathbf{1}^T;\ DY; \mathbf{K}_t = (K_t(x_i, x_j))_{i,j=1}^m, D\mathbf{K}_t.$
Solving the linear system (7) to derive $\alpha_{\mathbf{z}}^t = (\alpha_{\mathbf{z},1}^t, \ldots, \alpha_{\mathbf{z},m}^t)^T \in \mathbb{R}^m, t = 1, \ldots, l$.
return A ranking function
    $f_{\mathbf{z},\oplus}(x) = \sum_{t=1}^{l} \sum_{i=1}^{m} \alpha_{\mathbf{z},i}^t K_t(x_i, x).$
```

in the Gaussian kernel is selected from $\{4^{-3}, 4^{-2}, \ldots, 4^3\}$, and the regularization parameter λ is selected from $\{10^{-5}, 10^{-4}, \ldots, 10^1\}$. For simplicity, we choose $l = 2$ in all experiments. Given parameters $\lambda, \sigma_1, \sigma_2, \upsilon_1$, and υ_2, we solve linear system (7) to obtain the predictive function f and then evaluate the predictive performance on the test samples. For the data sets of drug discovery, we select the parameter by the cross validation from the training samples to illustrate the effectiveness of MLSRRank compared with other algorithms. Moreover, for the data sets of the recommendation system, we select the parameters as those resulting in the best predictive performance for LSRRank and MLSRRank.

B. Empirical Evaluation for QSAR Analysis

The experiments are based on two quantitative structure-activity relationship (QSAR) data sets, including inhibitors of dihydrofolate reductase (DHFR) and cyclooxygenase-2 (COX2), which correspond to the biological activities represented as pIC50 values [2], [33].

1) Data Sets and Evaluation Procedures: The DHFR inhibitor data set contains 361 compounds, with pIC50 values belonging to (3.3, 9.8); the COX2 inhibitor data set contains 282 compounds with pIC50 values belonging to (4, 9). For the DHFR data set, 237 out of 361 compounds are used as the training set and the remaining compounds are assigned to the test set. For the COX2 data set, 188 of 292 compounds are used as the training set and the rest of the compounds form the test set. Each compound in two data sets is represented by the 2.5-D chemical descriptors in [33]. Details of the data can be found in [33] and the references therein.

The DHFR inhibitor data set contains 70 real-valued descriptors, and the COX2 inhibitor data set contains 74 real-valued descriptors. In our experiments, each of these descriptors is scaled to (0, 1). In our experiments, we follow the same setup in [2].

2) Methods for Comparison: Here, we implement the algorithms of MLSRRank and compare the derived results with the results of RankSVM and support vector regression (SVR) in [2].

1) *RankSVM [2], [22]:* RankSVM for real-valued labels is a regularized ranking model. The dual formulation of this model leads to a convex quadratic program, and can be solved by a standard quadratic program solver or a gradient projection algorithm [2].
2) *SVR-Based Ranking:* SVR [36] can be used to learn a prediction function directly, and then rank the instances in a decreasing order of the predicted values. The selections of the kernel function and the regularization parameter are necessary to implement this algorithm.

3) Performance Measures: Denote $T = \{(x_i, y_i)\}_{i=1}^{m'}$ as a test set and f as a predictive function. The following measures are used to evaluate the ranking performance [2].

1) *Ranking Error:* The ranking error is defined as
$$\frac{1}{|P|} \sum_{(i,j) \in P} (y_i - y_j) \left(\mathbf{1}_{f(x_i) < f(x_j)} + \frac{1}{2} \mathbf{1}_{f(x_i) = f(x_j)} \right).$$
where $P = \{(i, j) | y_i > y_j\}$ denotes the set of preference pairs in T.
2) *Pearson Correlation Coefficient:* Let μ_f and σ_f be the mean and standard deviation of f, respectively. Let μ_y and σ_y be the mean and standard deviation of y, respectively. The correlation coefficient is defined as
$$\frac{1}{m' - 1} \sum_{i=1}^{m'} \frac{(f(x_i) - \mu_f)(y_i - \mu_y)}{\sigma_f \sigma_y}.$$
3) *Kendall's τ:* Kendall's τ is defined as the ratio of the number of concordant pairs subtracted by the number of discordant pairs in the total number of all pairs. The Kendall's τ can be given by
$$\frac{2}{|P|} \sum_{(i,j) \in P} \left(\mathbf{1}_{f(x_i) < f(x_j)} + \frac{1}{2} \mathbf{1}_{f(x_i) = f(x_j)} \right) - 1$$
if ties in the learned ranking are broken uniformly at random.
4) *Spearman's ρ:* Let $\beta_f(i)$ be the rank of x_i in the ranking returned by f and let $\beta_y(i)$ be the rank of x_i in the ranking based on the actual activities. The Spearman's ρ coefficient is defined as
$$\frac{1}{m' - 1} \sum_{i=1}^{m'} \frac{\left(\beta_f(i) - \mu'_f\right)\left(\beta_y(i) - \mu'_y\right)}{\sigma'_f \sigma'_y}.$$
In fact, it is the Pearson correlation between the vectors $\beta_f = (f(1), \ldots, f(m'))$ and $\beta_y = (y(1), \ldots, y(m'))$.

The Pearson correlation coefficient, the Kendall's τ, and Spearman's ρ belong to $(-1, 1)$, where 1 represents perfect agreement and -1 represents perfect disagreement. For simplicity, we denote these four measures as error, corr, τ, and ρ.

4) Experimental Results: The experimental procedures follow exactly the same setup in [2]. The results on ranking performance are summarized in Tables I and II. Those results of RankSVM and SVR come from [2] according to the above-mentioned performance measures. For the DHFR data set, the optimal parameters $\sigma_1 = 16, \sigma_2 = 64, \lambda = 0.001, \upsilon_1 = 1$, and $\upsilon_2 = 0.5$ are obtained by the cross validation. For the

TABLE I
QSAR RANKING RESULTS ON ORIGINAL SPLITS IN THE DHFR DATA SET

DHFR	SVR	RankSVM	SGDRank	LSRRank	MLSRRank
error	0.1837	**0.1726**	0.2916	0.1972	0.1782
corr	0.7519	**0.7618**	0.6019	0.7319	0.7508
τ	0.5571	**0.5747**	0.4351	0.5457	0.5654
ρ	0.8540	**0.8632**	0.6168	0.7476	0.7445

TABLE II
QSAR RANKING RESULTS ON ORIGINAL SPLITS IN THE COX2 DATA SET

COX2	SVR	RankSVM	SGDRank	LSRRank	MLSRRank
error	0.3138	0.3173	0.3110	0.2976	**0.2953**
corr	0.5836	0.5703	0.5664	0.6119	**0.6147**
τ	0.4351	0.4346	0.4270	0.4484	**0.4523**
ρ	0.6100	0.6174	0.5972	0.6280	**0.6340**

COX2 data set, we choose $\sigma_1 = 64, \sigma_2 = 64, \lambda = 0.01$, and $\upsilon_1 = \upsilon_2 = 2$ by the cross validation for prediction.

From these experimental results, we can see that MLSRRank has the best performance on the COX2 data set and the RankSVM reaches the best performance on the DHFR data set. Meanwhile, MLSRRank has shown better performance than LSRRank, especially for the DHFR data set. In fact, only the optimal predictor is a nonflat function, and the multiscale method can improve the learning performance of the single-kernel methods. Hence, the effectiveness of MLSRRank is depended on the characteristics of data sets. The results also indicate that the hinge loss is more suitable for the DHFR data set than the least square loss. However, MLSRRank can achieve better performance than SVR and SGDRank associated with least square losses.

In order to better understand the predictive performance of MLSRRank, we further present the experimental results with different numbers of training samples. For simplicity, we use the parameters selected in the previous experiment. Ten samples are selected as the labeled data at random from the original training set. The remaining training data is considered as the unlabeled data. Then, we test it on the original test set. All the performance evaluations are recorded. We then train MLSRRank on 20 training samples, 30 training samples, and so on. Each time, we add ten training samples. We repeat the whole process 30 times, and report the mean results in Fig. 1.

From Fig. 1, we can see that the MLSRRank can reach better predictive performance with larger numbers of training samples. This phenomenon is consistent with the theoretical analysis, and shows that the proposed algorithm can exploit the data information sufficiently.

C. Empirical Evaluation for the Recommendation Task

To better compare the generalization ability of LSRRank and MLSRRank, we evaluate their best performance on the data sets for the recommendation task. The recommendation task is aiming to provide a given user a list of unseen movies/jokes/books ordered by the predicted preference. The experimental setup used here is the same with [12]. All data sets are available at: http://www.grouplens.org/taxonomy/term/14.

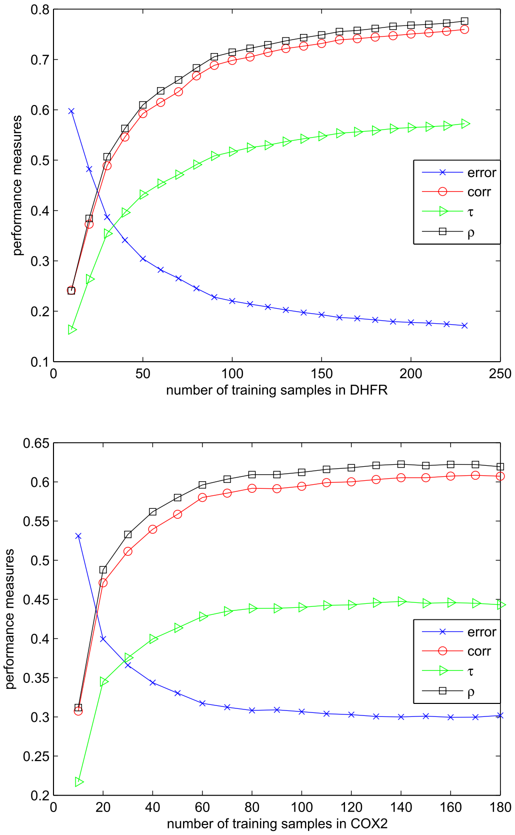


Fig. 1. Performance of MLSRRank versus the number of training samples.

1) Data Sets and Evaluation Procedures: The MovieLens data set contains 1 000 209 anonymous ratings of 3883 movies made by 6040 MovieLens users. Rating is an integer belonging to $\{1, \ldots, 5\}$. Only part of the movies is rated. The Jester Joke Recommender System data set contains 4.1 million continuous ratings ranging from -10 to 10 of 100 jokes from 73 421 users. The book-crossing data set contains 278 858 users and 1 149 780 ratings for 271 379 books.

For the MovieLens data set, the reviewers are grouped into 20–40 movies, 40–60 movies, and 60–80 movies based on the number of movies they had reviewed. The users, reviewed between 50 and 300 movies, are selected as the test reviewers used in our experiment. For a given test reviewer, we chose 300 reference reviewers randomly from one of these three groups and use their ratings to form the input vectors. Half of the test reviewer's movie ratings are used for training and the other half is used for testing; 300 different test reviewers are selected at random, and we recorded the average performance. For stability, we repeated the experiments ten times. The mean values and standard deviations are reported for each of the three groups after ten repeated experiments. For the Jester Joke Recommender System data set, we also establish the evaluation procedure with the same way as above.

For the book-crossing data set, we only select those users who have reviewed at least 200 books, and then only consider books with at least ten reviews. We finally obtain a data

TABLE III
COMPARISON OF LSRRANK AND MLSRRANK (MEAN AND STANDARD DEVIATIONS)

Method	MovieLens 20-40	MovieLens 40-60	MovieLens 60-80	Jester 20-40	Jester 40-60	Jester 60-80	Books
LSRRank (MSD)	1.92± 0.07	1.92± 0.04	1.88± 0.06	41.39± 1.12	41.24± 1.92	41.29± 1.67	2.18± 2.85
MLSRRank (MSD)	**1.90± 0.04**	**1.88± 0.05**	**1.85± 0.03**	**39.54± 1.61**	**39.54± 1.61**	**40.42± 1.46**	**1.80± 2.32**
LSRRank (M1D)	1.05± 0.02	1.05± 0.01	1.04± 0.02	4.76± 0.08	4.76± 0.11	4.75± 0.10	0.78± 0.69
MLSRRank (M1D)	**1.01± 0.02**	**1.02± 0.01**	**1.02± 0.01**	**4.63± 0.09**	**4.73± 0.08**	**4.71± 0.07**	**0.72± 0.64**

set of 87 books and 130 reviewers. For this data set, we choose one of 129 reviewers as a test reviewer each time, and use other 129 reviewers as reference reviewers. We report the mean values and standard deviations based on 130 leave-one-out experiments.

2) Performance Measures: In order to evaluate the magnitude-preserving algorithms (LSRRank and MLSRRank), we introduce the following measures used in [12]. Here, we set $\{(x_i, y_i)\}_{i=1}^{m'}$ as a test set and set f as a predictive function.

1) *Mean Squared Difference (MSD):* The MSD is defined as

$$\frac{1}{m'^2}\sum_{i,j=1}^{m'}((y_i - y_j) - (f(x_i) - f(x_j)))^2.$$

2) *Mean 1-Norm Difference (M1D):* The M1D over all pairs is defined as

$$\frac{1}{m'^2}\sum_{i,j=1}^{m'}|(y_i - y_j) - (f(x_i) - f(x_j))|.$$

3) Experimental Results: The experiments are used to describe the learning ability of the MLSRRank with respect to the magnitude-preserving measures MSD and M1D. The experimental results are reported in Table III. We can observe that in terms of MSD and M1D, MLSRRank outperforms LSRRank. This demonstrates that MLSRRank usually has better generalization performance than the corresponding algorithm with a single kernel for the recommendation tasks.

V. CONCLUSION

We have introduced an algorithm that learns ranking functions from the samples by the multiscale-kernel-based regularization framework. The implementation of the algorithm is simple and its representer theorem has been provided. In terms of the covering numbers of sum spaces, we presented the upper bound of the generalization error. Experiments performed on public data sets have demonstrated the effectiveness of the proposed algorithm.

Along the line of this paper, some improvements are necessary for future study that we discuss in the following.

1) *Design the Algorithm for Parameter Selection:* For the proposed multiscale ranking, five different parameters need to be selected based on the training data. Hence, a selection algorithm should be given to reduce the computation complexity. This is closely related to the recent studies in [40] and [41].
2) *Design the Sampling Method for Ranking:* In this paper, it is assumed that the training data are sampled in an independent identically distributed (i.i.d.) manner. Usually, the sampling technique in machine learning can draw important samples to training and reduce the requirement of the number of training data. Hence, it is crucial to find a reasonable way to realize efficient sampling for ranking, and extend the generalization analysis to non-i.i.d. samples. Recently, there are some advances along this line for fisher linear discriminant in [47] and online SVM in [39].

APPENDIX A
PROOF OF LEMMA 1

Proof: Define the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$ associated with a discrete subset $\mathbf{x} = \{x_i\}_{i=1}^m$ of $\mathscr{X}$ by

$$S_{\mathbf{x}}(f) = (f(x_i))_{i=1}^m = (f(x_1), \ldots, f(x_m))^T.$$

The adjoint of the sampling operator, $S_{\mathbf{x}}^T : \mathbb{R}^m \to \mathcal{H}_K$, is given by

$$S_{\mathbf{x}}^T c = \sum_{i=1}^m c_i K(x_i, \cdot), \quad c = (c_i)_{i=1}^m = (c_1, \ldots, c_m)^T \in \mathbb{R}^m.$$

By means of the reproducing property $f(x) = \langle f, K(x, \cdot)\rangle_K$, we know

$$\frac{\partial\left(\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right)}{\partial f} = 4\left(\frac{1}{m^2}S_{\mathbf{x}}^T D S_{\mathbf{x}} + \frac{\lambda}{2}I\right)f - \frac{4}{m^2}S_{\mathbf{x}}^T DY.$$

Then, $f_{\mathbf{z}}$ is given by the solution $(\partial(\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2)/\partial f) = 0$. That is to say $f_{\mathbf{z},\lambda}$ satisfies

$$\left(\frac{1}{m^2}S_{\mathbf{x}}^T D S_{\mathbf{x}} + \frac{\lambda}{2}I\right) f_{\mathbf{z}} = \frac{1}{m^2}S_{\mathbf{x}}^T DY. \tag{10}$$

By the properties of RKHS, we know the solution of optimization problem (4) can be denoted by a linear span of the kernel functions $\{K(x_i, \cdot)\}_{i=1}^m$. That is to say $f_{\mathbf{z}} = \sum_{i=1}^m \alpha_{\mathbf{z},i} K(x_i, \cdot) = S_{\mathbf{x}}^T \alpha_{\mathbf{z}}$ for some $\alpha_{\mathbf{z}} = (\alpha_{\mathbf{z},1}, \ldots, \alpha_{\mathbf{z},m})^T \in \mathbb{R}^m$. Then, from (10), we have

$$\left(D S_{\mathbf{x}} S_{\mathbf{x}}^T + \frac{\lambda m^2}{2}I\right)\alpha_{\mathbf{z}} = DY.$$

The desired result follows from $S_{\mathbf{x}}S_{\mathbf{x}}^T\alpha = \mathbf{K}\alpha$. ■

Appendix B
Proof of Theorem 1

Proof: First, recall some basic properties for the Mercer Kernel. Let $L_K : L^2_{\rho_{\mathcal{X}}} \to L^2_{\rho_{\mathcal{X}}}$ be an operator associated with a Mercer kernel K defined as

$$L_K f(x) = \int_{\mathcal{X}} K(x,u) f(u) d\rho_X(u), \quad x \in \mathcal{X}.$$

Denote $\{\phi_k^t\}_{k\geq 1}$ as the orthonormal basis of $L^2_{\rho_X}$ consisting of eigenfunctions of L_{K_t} and denote $\{\lambda_k^t\}_{k\geq 1}$ as their corresponding eigenvalues. Note that, for any $f_t \in \mathcal{H}_{K_t}$, $f_t = \sum_{k=1}^{\infty} c_k^t \phi_k^t$ and $\|f\|^2_{K_t} = \sum_{k=1}^{\infty}((c_k^t)^2/\lambda_k^t)$.

For each $s \geq 1$ and $f = \sum_{t=1}^{l} f_t$, we have

$$\begin{aligned}\frac{\partial\left(\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|^2_{\mathcal{H}_{\oplus,\upsilon}}\right)}{\partial c_k^t} &\\ = -\frac{2}{m^2}\sum_{i,j=1}^{m}&(y_i - y_j - (f(x_i) - f(x_j)))\\ &\cdot(\phi_k^t(x_i) - \phi_k^t(x_j)) + \frac{2\lambda \upsilon_t c_k^t}{\lambda_k^t}.\end{aligned}$$

By setting the above equation equal to 0, we obtain

$$\begin{aligned}c_k^t = \frac{\lambda_k^t}{m^2\lambda\upsilon_t}\sum_{i,j=1}^{m}&(y_i - y_j - (f(x_i) - f(x_j)))\\ &\cdot\big(\phi_k^t(x_i) - \phi_k^t(x_j)\big).\end{aligned}$$

Then

$$\begin{aligned}f_t(x) &= \sum_{k=1}^{\infty}\frac{\lambda_k^t}{m^2\lambda\upsilon_t}\sum_{i,j=1}^{m}(y_i - y_j - (f(x_i) - f(x_j)))\\ &\quad\cdot\big(\phi_k^t(x_i) - \phi_k^t(x_j)\big)\phi_k^t\\ &= \frac{1}{m^2\lambda\upsilon_t}\sum_{i=1}^{m}K_t(x_i,x)\\ &\quad\times\Bigg(y_i - \frac{1}{m}\sum_{j=1}^{m}y_j - \sum_{s=1}^{l}\sum_{k=1}^{m}a_k^t\\ &\quad\times\Bigg(K_s(x_k,x_i) - \frac{1}{m}\sum_{j=1}^{m}K_s(x_k,x_j)\Bigg)\Bigg).\end{aligned}$$

From the properties of RKHS, we know $f_t = \sum_{i=1}^{m}\alpha_i^t K_t(x_i,x)$. Hence

$$\begin{aligned}\alpha_i^t = \frac{1}{m^2\lambda\upsilon_t}\Bigg(y_i - \frac{1}{m}\sum_{j=1}^{m}y_j - \sum_{s=1}^{l}\sum_{k=1}^{m}a_k^t\Bigg(K_s(x_k,x_i)&\\ - \frac{1}{m}\sum_{j=1}^{m}K_s(x_k,x_j)\Bigg)\Bigg).&\end{aligned}$$

Then, the desired result follows by denoting $\alpha^t = (\alpha_1^t,\ldots,\alpha_m^t)^T$. ■

Appendix C
Proof of Proposition 2

The main tool used here is the McDiarmid's inequality.

Lemma 2 (McDiarmid's Inequality): Let $x_1,\ldots,x_n$ be independent random variables taking values in a set $\mathscr{A}$, and assume that $\phi : \mathscr{A}^n \to \mathbb{R}$ satisfies

$$\sup_{x_1,\ldots,x_n,\tilde{x}_i\in\mathscr{A}} |\phi(x_1,\ldots,x_n) - \phi(x_1,\ldots,\tilde{x}_i,\ldots x_n)| \leq b_i$$

for every $1 \leq i \leq n$. Then, for every $\varepsilon > 0$

$$\mathrm{Prob}\{\phi(x_1,\ldots,x_n) - E\phi \geq \varepsilon\} \leq \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^{n}b_i^2}\right\}.$$

Now, we present the proof of Proposition 2.

Proof: Let $\mathbf{z} = \{z_i\}_{i=1}^m \in Z^m$ and $\mathbf{z}^k = (z_1,\ldots,z_{k-1},z_k',z_{k+1},\ldots,z_m)$. Denote $\phi(\mathbf{z}) = \mathcal{E}_{\mathbf{z}}^{\ell}(f_{\gamma,\oplus}^{\ell})$, then $E\phi(\mathbf{z}) = \mathcal{E}^{\ell}(f_{\gamma,\oplus}^{\ell})$. From the definition of $f_{\gamma,\oplus}^{\ell}$, we get $\|f_{\gamma,\oplus}^{\ell}\|_\infty \leq \kappa\sqrt{lD_{\oplus,\upsilon}(\gamma)/\gamma}$. Then, for any $1 \leq k \leq m$, we have

$$\begin{aligned}&\phi(\mathbf{z}) - \phi(\mathbf{z}^k)|\\ &\quad\leq \frac{2}{m(m-1)}\sum_{i\neq k}\left|\ell\big(f_{\gamma,\oplus}^{\ell},z_i,z_j\big) - \ell\big(f_{\gamma,\oplus}^{\ell},z_k,z_j\big)\right| \leq \frac{2M_1}{m}.\end{aligned}$$

According to the McDiarmid inequality, for any $\varepsilon > 0$, we obtain

$$\mathop{\mathrm{Prob}}_{\mathbf{z}\in Z^m}\left\{\mathcal{E}_{\mathbf{z}}^{\ell}\big(f_{\gamma,\oplus}^{\ell}\big) - \mathcal{E}^{\ell}\big(f_{\gamma,\oplus}^{\ell}\big) \geq \varepsilon\right\} \leq \exp\left\{\frac{-m\varepsilon^2}{2M_1^2}\right\}.$$

The desired result follows by setting $\delta = \exp\{(-m\varepsilon^2/2M_1^2)\}$. ■

Appendix D
Proof of Proposition 3

Proof: With the same fashion in the proof of Proposition 2, we can get for any $f \in B_{\oplus,\upsilon,R}$

$$\mathrm{Prob}\left\{\mathcal{E}^{\ell}(f) - \mathcal{E}_{\mathbf{z}}^{\ell}(f) \geq \varepsilon\right\} \leq \exp\left\{-\frac{m\varepsilon^2}{2M_2^2}\right\}.$$

Now, we turn to the technique in [14] to obtain the uniform convergence estimate. Let $J = \mathcal{N}(B_{\oplus,\upsilon,R},(\varepsilon/4L))$ and f_i, $1 \leq i \leq J$ be the centers of disks D_j such that $B_{\oplus,\upsilon,R} \subset \cup_{i=1}^{J} D_i$. Note that, for all $f \in D_j$ and $\mathbf{z} \in Z^m$

$$\left|\mathcal{E}^{\ell}(f) - \mathcal{E}_{\mathbf{z}}^{\ell}(f) - \big(\mathcal{E}(f_j) - \mathcal{E}_{\mathbf{z}}^{\ell}(f_j)\big)\right| \leq 2L\|f - f_j\|_\infty \leq \frac{\varepsilon}{2}.$$

Then

$$\sup_{f\in D_j}\big(\mathcal{E}^{\ell}(f) - \mathcal{E}_{\mathbf{z}}^{\ell}(f)\big) \geq \varepsilon \quad\Rightarrow\quad \mathcal{E}(f_j) - \mathcal{E}_{\mathbf{z}}^{\ell}(f_j) \geq \frac{\varepsilon}{2}.$$

That is to say

$$\begin{aligned}&\mathop{\mathrm{Prob}}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in D_j}\big(\mathcal{E}^{\ell}(f) - \mathcal{E}_{\mathbf{z}}^{\ell}(f)\big) \geq \varepsilon\right\}\\ &\qquad\leq \mathop{\mathrm{Prob}}_{\mathbf{z}\in Z^m}\left\{\mathcal{E}(f_j) - \mathcal{E}_{\mathbf{z}}^{\ell}(f_j) \geq \frac{\varepsilon}{2}\right\} \leq \exp\left\{-\frac{m\varepsilon^2}{8M_2^2}\right\}.\end{aligned}$$

Note that

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_{\oplus,\upsilon,R}}\left(\mathcal{E}^{\ell}(f)-\mathcal{E}^{\ell}_{\mathbf{z}}(f)\right)\geq\varepsilon\right\}$$
$$\leq\sum_{j=1}^{J}\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in D_j}\left(\mathcal{E}^{\ell}(f)-\mathcal{E}^{\ell}_{\mathbf{z}}(f)\right)\geq\varepsilon\right\}.$$

Hence

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_{\oplus,\upsilon,R}}\left(\mathcal{E}^{\ell}(f)-\mathcal{E}^{\ell}_{\mathbf{z}}(f)\right)\geq\varepsilon\right\}$$
$$\leq\mathcal{N}\left(B_{\oplus,\upsilon,R},\frac{\varepsilon}{4L}\right)\exp\left\{-\frac{m\varepsilon^2}{8M_2^2}\right\}.$$

Combining the above estimate with $\|f^{\ell}_{\mathbf{z},\oplus}\|_{\oplus,\upsilon}\leq M/\sqrt{\gamma}$, we derive the desired result. ■

Appendix E
Proof of Theorem 2

The proof of Theorem 2 is dependent on the following lemma established in [15].

Lemma 3: Let $c_1, c_2 > 0$ and $p_1 > p_2 > 0$. Then, the equation

$$x^{p_1}-c_1x^{p_2}-c_2=0$$

has a unique positive zero solution x^*. In addition

$$x^*\leq\max\left\{(2c_1)^{\frac{1}{p_1-p_2}},(2c_2)^{\frac{1}{p_1}}\right\}.$$

Now, we present the proof of Theorem 2.

Proof: From [41, Lemma 3], we know $\mathcal{N}(B_{\oplus,\upsilon,R},\varepsilon/(4L))\leq\prod_{t=1}^{l}\mathcal{N}(B_{K_t,R},\varepsilon/(4lL)$. Moreover, $\mathcal{N}(B_{K_t,R},\varepsilon/(4lL))\leq\mathcal{N}(B_{K_t,1},\varepsilon/(4lLR))$. Observe that $\ln(\mathcal{N}(B_1,\eta))\leq C_d(\eta^{-s}+\mu^{-2(d+1)})$ has been presented in [16]. Here, C_d is a parameter depending on d and $s>0$ is a parameter, which can be arbitrarily small.

Setting $\prod_{t=1}^{l}\mathcal{N}(B_{K_t,1},\varepsilon/(4lLR)\exp\{-(m\varepsilon^2/8M_2^2)\}=\delta/2$, we get

$$C_dl\left(\frac{\varepsilon}{4lLR}\right)^{-s}+C_d\sum_{t=1}^{l}\mu_t^{-2(d+1)}-\frac{m\varepsilon^2}{2M_2^2}-\ln(2/\delta)\geq0.$$

Then

$$\varepsilon^{2+s}-\frac{2M_2^2}{m}\left(C_d\sum_{t=1}^{l}\mu_t^{-2(d+1)}+\ln(2/\delta)\right)\varepsilon^s$$
$$-\frac{2C_dlM_2^2(4lLR)^s}{m}\leq0.\quad(11)$$

Denote $p_1=2+s$, $p_2=s$, $c_1=(2M_2^2/m)(C_d\sum_{t=1}^{l}\mu_t^{-2(d+1)}+\ln(2/\delta))$, and $c_2=(2C_dlM_2^2(4lLR)^s/m)$. Now, take the equality in (11) to obtain the equation

$$\varepsilon^{p_1}-c_1\varepsilon^{p_2}-c_2=0.$$

Based on Lemma 3, this equation has only one positive zero ε^* satisfying that

$$\varepsilon^*\leq\max\left\{(2c_1)^{\frac{1}{p_1-p_2}},(2c_2)^{\frac{1}{p_1}}\right\}.$$

Therefore, from Proposition 3 and $\|f^{\ell}_{\mathbf{z},\oplus}\|_{\oplus,\upsilon}\leq(M/\sqrt{\gamma})$, we have

$$\mathcal{E}(f^{\ell}_{\mathbf{z},\oplus})-\mathcal{E}^{\ell}_{\mathbf{z}}(f^{\ell}_{\mathbf{z},\oplus})$$
$$\leq\max\left\{\frac{2M_2}{\sqrt{m}}\left(C_d\sum_{t=1}^{l}\mu_t^{-2(d+1)}+\ln(2/\delta)\right)^{\frac{1}{2}}\right.$$
$$\left.\times\left(\frac{4C_dlM_2^2(4lLR)^s}{m}\right)^{\frac{1}{2+s}}\right\}\quad(12)$$

with confidence at least $1-\delta/2$, where $R=(M/\sqrt{\gamma})$. Combining estimate (12) with Propositions 1 and 2, we get the statement in Theorem 2. ■

References

[1] S. Agarwal and P. Niyogi, "Generalization bounds for ranking algorithms via algorithmic stability," *J. Mach. Learn. Res.*, vol. 10, pp. 441–474, Feb. 2009.

[2] S. Agarwal, D. Dugar, and S. Sengupt, "Ranking chemical structures for drug discovery: A new machine learning approach," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 716–731, 2010.

[3] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.

[4] C. Burges *et al.*, "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn.*, Bonn, Germany, Aug. 2005, pp. 89–96.

[5] C. Burges, R. Ragno, and Q. Le, "Learning to rank with nonsmooth cost functions," in *Advances in Neural Information Processing Systems 19*. Vancouver, BC, Canada: MIT Press, Dec. 2006, pp. 193–200.

[6] H. Chen, "The convergence rate of a regularized ranking algorithm," *J. Approx. Theory*, vol. 164, no. 12, pp. 1513–1519, 2012.

[7] H. Chen and L. Q. Li, "Learning rates of multi-kernel regularized regression," *J. Statist. Planning Inference*, vol. 140, no. 9, pp. 2562–2568, 2010.

[8] H. Chen, Y. Tang, L. Li, Y. Yuan, X. Li, and Y. Tang, "Error analysis of stochastic gradient descent ranking," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 898–909, Jun. 2013.

[9] H. Chen, J. Peng, Y. Zhou, L. Li, and Z. Pan, "Extreme learning machine for ranking: Generalization analysis and applications," *Neural Netw.*, vol. 53, pp. 119–126, May 2014.

[10] S. Clemencon, G. Lugosi, and N. Vayatis, "Ranking and empirical minimization of U-statistics," *Ann. Statist.*, vol. 36, no. 2, pp. 844–874, 2008.

[11] D. Cossock and T. Zhang, "Statistical analysis of Bayes optimal subset ranking," *IEEE Trans. Inform. Theory*, vol. 54, no. 11, pp. 5140–5154, Nov. 2008.

[12] C. Cortes, M. Mohri, and A. Rastogi, "Magnitude-preserving ranking algorithms," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 169–176.

[13] C. Cortes, M. Mohri, and A. Rastogi, "An alternative ranking problem for search engines," in *Proc. 6th Workshop Experim. Algorithms*, Heidelberg, Germany, Jun. 2007, pp. 1–21.

[14] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, 2002.

[15] F. Cuker and S. Smale, "Best choices for regularization parameters in learning theory: On the bias–variance problem," *Found. Comput. Math.*, vol. 2, no. 4, pp. 413–428, 2002.

[16] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[17] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 465–479, Mar. 2012.

[18] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *J. Mach. Learn. Res.*, vol. 4, pp. 933–969, Dec. 2003.

[19] F. He and H. Chen, "Generalization performance of bipartite ranking algorithms with convex losses," *J. Math. Anal. Appl.*, vol. 404, no. 2, pp. 528–536, 2013.

[20] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in *Advances in Large Classifiers*. Cambridge, MA, USA: MIT Press, 2000.

[21] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou, "Learning theory approach to minimum error entropy criterion," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 377–397, 2013.

[22] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Edmonton, AB, Canada, Jul. 2002, pp. 133–142.

[23] K. Uematsu and Y. Lee, "Statistical optimality in multipartite ranking and ordinal regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1080–1094, May 2015.

[24] Y. Lei, L. Ding, and W. Zhang, "Generalization performance of radial basis function networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 551–564, Mar. 2015.

[25] X. Liu, S. Lin, J. Fang, and Z. Xu, "Is extreme learning machine feasible? A theoretical assessment (Part I)," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 7–20, Jan. 2015.

[26] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Boberg, and T. Salakoski, "Learning to rank with pairwise regularized least-squares," in *Proc. SIGIR Workshop Learn. Rank Inf. Retr.*, 2007, pp. 27–33.

[27] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Jarvinen, and J. Boberg, "An efficient algorithm for learning to rank from preference graphs," *Mach. Learn.*, vol. 75, no. 1, pp. 129–165, 2009.

[28] V. C. Raykar, R. Duraiswami, and B. Krishnapuram, "A fast algorithm for learning a ranking function from large-scale data sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1158–1170, Jul. 2008.

[29] W. Rejchel, "On ranking and generalization bounds," *J. Mach. Learn. Res.*, vol. 13, pp. 1373–1392, May 2012.

[30] C. Rudin and R. E. Schapire, "Margin-based ranking and an equivalence between AdaBoost and RankBoost," *J. Mach. Learn. Res.*, vol. 10, pp. 2193–2232, Dec. 2009.

[31] C. Rudin, "The P-norm push: A simple convex ranking algorithm that concentrates at the top of the list," *J. Mach. Learn. Res.*, vol. 10, pp. 2233–2271, Dec. 2009.

[32] I. Steinwart and C. Scovel, "Fast rates for support vector machines using Gaussian kernels," *Ann. Statist.*, vol. 35, no. 2, pp. 575–607, 2007.

[33] J. J. Sutherland, L. A. O'Brien, and D. F. Weaver, "A comparison of methods for modeling quantitative structure–activity relationships," *J. Med. Chem.*, vol. 47, no. 22, pp. 5541–5554, 2004.

[34] Y. Tang, L. Li, and X. Li, "Learning similarity with multikernel method," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 131–138, Feb. 2011.

[35] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.

[36] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[37] Q. Wu, Y. Ying, and D.-X. Zhou, "Learning rates of least-square regularized regression," *Found. Comput. Math.*, vol. 6, no. 2, pp. 171–192, 2006.

[38] Q. Wu, Y. Ying, and D.-X. Zhou, "Multi-kernel regularized classifiers," *J. Complex.*, vol. 23, no. 1, pp. 108–134, 2007.

[39] J. Xu, Y. Y. Tang, B. Zou, Z. Xu, L. Li, and Y. Lu, "The generalization ability of online SVM classification based on Markov sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 628–639, Mar. 2015.

[40] Y.-L. Xu and D.-R. Chen, "Partially-linear least-squares regularized regression for system identification," *IEEE Trans. Autom. Control*, vol. 54, no. 11, pp. 2637–2641, Nov. 2009.

[41] Y.-L. Xu, D.-R. Chen, H.-X. Li, and L. Liu, "Least square regularized regression in sum space," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 635–646, Apr. 2013.

[42] Y. Ying and D.-X. Zhou, "Learnability of Gaussians with flexible variances," *J. Mach. Learn. Res.*, vol. 8, pp. 249–276, Feb. 2007.

[43] C. Zhang and D. Tao, "Generalization bounds of ERM-based learning processes for continuous-time Markov chains," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1872–1883, Dec. 2012.

[44] Y. Zhang and F. Cao, "Analysis of convergence performance of neural networks ranking algorithm," *Neural Netw.*, vol. 34, pp. 65–71, Oct. 2012.

[45] D. Zheng, J. Wang, and Y. Zhao, "Non-flat function estimation with a multi-scale support vector regression," *Neurocomputing*, vol. 70, nos. 1–3, pp. 420–429, 2006.

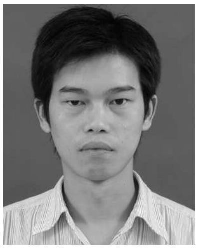
[46] D. X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1743–1752, Jul. 2003.

[47] B. Zou, L. Li, Z. Xu, T. Luo, and Y. Y. Tang, "Generalization performance of Fisher linear discriminant based on Markov sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 2, pp. 288–300, Feb. 2013.

Yicong Zhou (M'07–SM'14) received the B.S. degree from Hunan University, Changsha, China, in 1992, and the M.S. and Ph.D. degrees from Tufts University, Medford, MA, USA, in 2008 and 2010, all in electrical engineering.

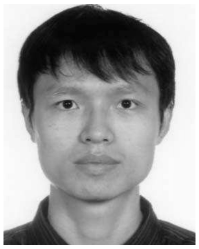
He is currently an Assistant Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His current research interests include multimedia security, image/signal processing, pattern recognition, and medical imaging.

Dr. Zhou is a member of the International Society for Photo-Optical Instrumentations Engineers and the Association for Computing Machinery. He received the third price of the Macau Natural Science Award in 2014.



Hong Chen received the B.Sc. and Ph.D. degrees from Hubei University, Wuhan, China, in 2003 and 2009, respectively.

He is currently an Associate Professor with the Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan. His current research interests include statistical learning theory, approximation theory, and machine learning.



Rushi Lan received the B.S. and M.S. degrees from the Nanjing University of Information Science and Technology, Nanjing, China, in 2008 and 2011, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer and Information Science, University of Macau, Macau, China.

His current research interests include image processing and pattern recognition.

Zhibin Pan received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2014.

He is currently an Associate Professor with the Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan. His current research interests include machine learning and pattern recognition.