Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2013, Article ID 927827, 8 pages
http://dx.doi.org/10.1155/2013/927827

Research Article

Regularized Ranking with Convex Losses and ℓ^1-Penalty

Heng Chen and Jitao Wu

Department of Mathematics, Beijing University of Aeronautics and Astronautics, Beijing 100191, China

Correspondence should be addressed to Heng Chen; chenheng_0913@163.com

Received 26 September 2013; Accepted 13 November 2013

Academic Editor: Yiming Ying

In the ranking problem, one has to compare two different observations and decide the ordering between them. It has received increasing attention both in the statistical and machine learning literature. This paper considers ℓ^1-regularized ranking rules with convex loss. Under some mild conditions, a learning rate is established.

1. Introduction

In the ranking problem, one has to compare two different observations and decide the ordering between them. The problem of ranking has become an interesting field for researchers in machine learning community. It has received increasing attention both in the statistical and machine learning literature.

The problem of ranking may be modeled in the framework of statistical learning (see [1, 2]). Let (X, Y) be a pair of random variables taking values in $\mathcal{X} \times \mathbb{R}$. The random observation X models some object and Y denotes its real-valued label. Let (X', Y') denote a pair of random variables identically distributed with (X, Y) (with respect to the probability $\mathbb{P}$) and independent of it. In the ranking problem one observes X and X' but not their labels Y and Y'. X is "better" than X' if $Y > Y'$. We are to construct a measurable function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, called a ranking rule, which predicts the ordering between objects in the following way: if $f(X, X') \geq 0$, we predict that X is better than X'. A ranking rule f has the property $f(x, x')f(x', x) \leq 0$. The performance of a ranking rule f is measured by the ranking error:

$$L(f) = \mathbb{P}\left(\operatorname{sign}\left(Y - Y'\right) f\left(X, X'\right) < 0\right), \tag{1}$$

that is, the probability that f ranks two randomly drawn instances incorrectly. It is easily seen that $L(f)$ attains its minimum L^*, over the class of all measurable functions, at the ranking rule

$$f^*(x, x') := \operatorname{sgn}\left(2\eta\left(x, x'\right) - 1\right) \quad \text{with } \eta\left(x, x'\right) = \mathbb{P}\left(Y > Y' \mid X = x, X' = x'\right). \tag{2}$$

In practice, the best rule f^* is unknown since the probability $\mathbb{P}$ is unknown. A widely used approach for estimating L^* is the empirical risk minimization with convex loss.

Definition 1. one says that $\phi : \mathbb{R} \to [0, \infty)$ is a ranking loss (function) if it is convex, differentiable at 0 with $\phi'(0) < 0$, and the smallest zero of ϕ is 1.

Examples of ranking loss include the least square loss $\phi(t) = (1 - t)^2$ and q-norm SVM loss $\phi_q(t) = (1 - t)_+^q$, where $q \geq 1$ and $t_+ = \max\{0, t\}$ for $t \in \mathbb{R}$.

The risk of a measurable function f is defined as $Q(f) = \mathbb{E}[\phi(\operatorname{sign}(Y - Y')f(X, X'))]$. Denote by f_ϕ a minimizer of $Q(f)$ over the set of all measurable and antisymmetric functions. For example, as in the classification case (see [3, 4]), $f_{\phi_1} = f^*$, and for $q > 1$,

$$f_{\phi_q}\left(x, x'\right) = \frac{\left(1 + f^*\left(x, x'\right)\right)^{1/(q-1)} - \left(1 - f^*\left(x, x'\right)\right)^{1/(q-1)}}{\left(1 + f^*\left(x, x'\right)\right)^{1/(q-1)} + \left(1 - f^*\left(x, x'\right)\right)^{1/(q-1)}}, \quad x, x' \in \mathcal{X}. \tag{3}$$

The following inequality holds for any f:

$$L(f) - L^* \le \begin{cases} Q(f) - Q^*, & \text{if } \phi(t) = (1-t)_+, \\ c\sqrt{Q(f) - Q^*}, & \text{if } \phi''(0) \ge 0, \end{cases} \tag{4}$$

where $Q^* = Q(f_\phi)$ and c is some constant.

Before proceeding further, we introduce the notion of Reproducing Kernel Hilbert Space (RKHS). Recall that a continuous function $K(\sigma, \sigma')$ is a Mercer kernel on a set Σ, if $K(\sigma, \sigma') = K(\sigma', \sigma), \forall \sigma, \sigma' \in \Sigma$, and given an arbitrary finite set $\{\sigma_1, \ldots, \sigma_n\} \subset \Sigma$, the matrix $\mathbf{K} = (K(\sigma_i, \sigma_j))_{i,j=1}^n$ is positive semidefinite. The RKHS $\mathscr{H}_K$ associated with the Mercer kernel K is the completion of $\operatorname{span}\{K_\sigma = K(\sigma, \cdot) \mid \sigma \in \Sigma\}$, with respect to the inner product given by $\langle K_\sigma, K_{\sigma'} \rangle_K = K(\sigma, \sigma')$. See [5] and ([6, Ch. 4]) for details.

For convenience, we assume hereafter that the Mercer kernels K on $\mathscr{X}^2 \times \mathscr{X}^2$ are symmetric in the sense that

$$K\left(\left(u, u'\right), \left(x, x'\right)\right) = K\left(\left(u', u\right), \left(x', x\right)\right), \quad \forall u, u', x, x' \in \mathscr{X}. \tag{5}$$

Such examples are Mercer kernels K of either form $K(s,t) = k(|s-t|_2)$ or $K(s,t) = k(\langle s, t \rangle)$, $s, t \in \mathscr{X}^2$, where $|\cdot|_2$ and $\langle \cdot, \cdot \rangle$ are the Euclidean norm and inner product, respectively.

Since the best ranking rule f^* is antisymmetric, it is reasonable that we restricted ourselves to the subspace $\mathscr{H}^{\text{as}}$ of anti-symmetric functions in $\mathscr{H}_K$; that is,

$$\mathscr{H}_K^{\text{as}} = \left\{ f \mid f \in \mathscr{H}_K, f\left(x, x'\right) = -f\left(x', x\right), \forall x, x' \in \mathscr{X} \right\}. \tag{6}$$

For any $x_i \in \mathscr{X}, i = 1, \ldots, n$, and $\alpha_{ij} \in \mathbb{R}$ with $\alpha_{ij} = -\alpha_{ji}, i, j = 1, \ldots, n$, it is easily seen that

$$f = \sum_{i,j=1}^n \alpha_{ij} K_{(x_i, x_j)} \in \mathscr{H}_K^{as}. \tag{7}$$

Conversely, any anti-symmetric function $f \in \operatorname{span}\{K_{(x_i,x_j)}\}_{i,j=1}^n$ with above expression should satisfy $\alpha_{ij} = -\alpha_{ji}, i, j = 1, \ldots, n$, provided $\det \mathbf{K} > 0$.

For a set of samples $\mathbf{z} = \{Z_1, \ldots, Z_n\} \subset \mathscr{X}^2$, let

$$\Omega_z(f) = \inf\left\{ \sum_{i,j=1}^n \left|\alpha_{ij}\right| \mid f = \sum_{i,j=1}^n \alpha_{ij} K_{(x_i,x_j)} \right\};$$

$$\mathscr{H}_{K,z} = \left\{ \sum_{i,j=1}^n \alpha_{ij} K_{(x_i,x_j)} \right\}, \tag{8}$$

$$\mathscr{H}_{K,z}^{\text{as}} = \left\{ \sum_{i,j=1}^n \alpha_{ij} K_{(x_i,x_j)} \mid \alpha_{ij} = -\alpha_{ji} \right\}.$$

For $\lambda > 0$, the ℓ^1-penalty regularized ranking rule $f_{z,\lambda}$ is the minimizer of the minimization problem

$$f_{z,\lambda} = \arg\min_{f \in \mathscr{H}_{K,z}^{\text{as}}} \left\{ Q_n(f) + \lambda \Omega_z(f) \right\}, \tag{9}$$

where $Q_n(f)$, known as empirical risk, is given by

$$Q_n(f) = \frac{1}{n(n-1)} \sum_{i \ne j} \phi\left(\operatorname{sign}\left(y_i - y_j\right) f\left(x_i, x_j\right)\right). \tag{10}$$

Associated with any ranking rule f, we construct another ranking rule as follows:

$$\pi(f)\left(x, x'\right) = \begin{cases} \operatorname{sign}\left(f\left(x, x'\right)\right), & \text{if} \left|f\left(x, x'\right)\right| > 1, \\ f\left(x, x'\right), & \text{if} \left|f\left(x, x'\right)\right| \le 1. \end{cases} \tag{11}$$

Clearly, $\pi(f)$ gives the same ranking rule as f and it satisfies

$$Q(\pi(f)) \le Q(f), \quad Q_n(\pi(f)) \le Q_n(f). \tag{12}$$

Hereafter, we denote $g_n = \pi(f_{z,\lambda})$. The goal of this paper is to bound the excess error $Q(g_n) - Q^*$, which in turn together with (4) up bounds the excess ranking error $L(g_n) - L^*$. The main result of this paper is, under mild conditions, to establish a learning rate for ℓ^1-penalty regularized ranking rules with convex loss.

Classification with convex loss, in particular for q-norm SVMs, has been the subject of many theoretical considerations in recent years. The 1-norm SVMs with regularizer being RKHS norm for ranking was investigated in [2, 7]. The ℓ^1-penalty has been used in [8, 9] for classification problems under the framework of SVMs. It is well known that ℓ^1-regularization usually leads to solution with sparse representation (see, e.g., [10–13]). In this paper, we consider ranking with convex loss and ℓ^1-penalty.

In [2], the RKHS-norm SVMs for ranking was proposed. But it was implemented over a ball $B_R = \{f \in \mathscr{H} \mid \|f\|_K \le R\}$ of $\mathscr{H}_K$, not the whole RKHS $\mathscr{H}_K$; that is, it solves the minimization problem

$$f_n = \arg\min_{f \in B_R} Q_n(f) + \lambda \|f\|_K^2. \tag{13}$$

A convergence rate for $Q(f_n) - \inf_{f \in B_R} Q(f)$ has been established for Gaussian kernel. The approximation error $\inf_{f \in B_R} Q(f) - Q^*$ was not considered there. The asymptotic behavior of the same algorithm implemented over the whole RKHS $\mathscr{H}_K$ was investigated in [7]. Moreover, a fast learning rate $O(1/n)$ is obtained under some conditions.

We would like to mention a recent paper [14], where the error $\mathscr{E}(f)$ of function f is defined by $\mathscr{E}(f) = \mathbb{E}((Y - Y' - f(X) + f(X'))^2)$. A convergence rate for the minimizer of the regularized empirical error was established. The author made use of the technique of estimation via integral operator developed in [15].

The rest of the paper is organized as follows. In Section 2, after making some assumptions, we state the main result, an upper bound for $Q(g_n) - Q^* + \lambda \Omega_z(f_{z,\lambda})$. As usual, it is decomposed as a sum of three terms, sample error, hypothesis error, and approximation error. Sections 3 and 4 are devoted to the estimations of hypothesis error and sample error, respectively. A proof of the main result is given in Section 5.

2. Assumptions and Main Results

For the statement of the main results, we need to introduce some notions and make some assumptions.

Denote $g_f = \phi(\text{sign}(y - y')f(x,x')) - \phi(\text{sign}(y - y')f_\phi(x,x'))$. The following assumption is a bound for variance of g_f, which is adopted by many authors.

Assumption 2. There is a constant $\alpha \in [0,1]$ such that for any M,

$$\mathbb{E}\left(g_f^2\right) \le c_1\left(\mathbb{E}\left(g_f\right)\right)^\alpha, \quad \forall f : \mathcal{X}^2 \longrightarrow [-M, M], \tag{14}$$

where $c_1 = c_1(M)$ is a constant.

For $\phi_q(t) = (1-t)_+^q$, the assumption is satisfied [16] for

$$\alpha = \begin{cases} 1, & \text{if } 1 < q \le 2, \\ \dfrac{2}{q}, & \text{if } 2 < q. \end{cases} \tag{15}$$

It is known in [17] that if there is some positive constant C such that

$$\mathbb{P}\left(\left|2\eta\left(X, X'\right) - 1\right| \le \xi\right) \le C\xi^{\alpha/(1-\alpha)}, \quad \forall \xi > 0, \tag{16}$$

then the assumption is satisfied for $\phi_1(t) = (1-t)_+$.

Suppose hereafter $\kappa := \sup_{s\in\mathcal{X}^2} K(s,s) < \infty$. We note that, for any Mercer kernel K,

$$\sup_{s\in\mathcal{X}^2} K(s,s) = \sup_{s,s'\in\mathcal{X}^2} K\left(s,s'\right). \tag{17}$$

We now construct a set of functions which contains $\mathcal{H}_{K,z}$ and is independent of the samples.

Definition 3. The Banach space $\mathcal{H}$ is defined as the function set on $\mathcal{X}^2$ containing all functions of the form

$$f = \sum_{i=1}^{\infty} a_i K_{s_i}, \quad \{a_i\}_{i=1}^{\infty} \in \ell^1, \ s_i \in \mathcal{X}^2, \tag{18}$$

with the norm

$$\|f\| := \inf\left\{\sum_{i=1}^{\infty} |a_i| \ f = \sum_{i=1}^{\infty} a_i K_{s_i}\right\}. \tag{19}$$

Obviously, $\mathcal{H}_{K,z} \subset \mathcal{H}, \forall \mathbf{z} \in \mathbf{Z}^n$. By the definition of κ and (17), one has

$$\begin{aligned} \left\|\sum_{i=N_1}^{N_2} a_i K_{s_i}\right\|_K^2 &= \sum_{i,j=N_1}^{N_2} a_i a_j K\left(s_i, s_j\right) \\ &\le \kappa\left(\sum_{i=N_1}^{N_2} |a_i|\right)^2, \end{aligned} \tag{20}$$

which implies that the series $\sum_{i=1}^{\infty} a_i K_{s_i}$ converges in $\mathcal{H}_K$. Consequently, $f \in \mathcal{H}_K$ and $\|f\|_K \le \sqrt{\kappa}\|f\|$. The following also holds: $\|f\|_C \le \kappa\|f\|, \forall f \in \mathcal{H}$, where $\|f\|_C = \sup_{s\in\mathcal{X}^2}|f(s)|$ for $f \in C(\mathcal{X}^2)$.

Denote $\mathcal{H}^{\text{as}} = \{f \in \mathcal{H} \mid f(x,x') = -f(x',x)\}$. The approximation error of Q^* by $Q(f)$ with $f \in \mathcal{H}^{\text{as}}$ is defined as

$$D(\lambda) := \inf_{f\in\mathcal{H}^{\text{as}}} \left\{Q(f) - Q^* + \lambda\|f\|\right\}. \tag{21}$$

Denote the minimizer

$$f_\lambda := \arg\min_{f\in\mathcal{H}^{\text{as}}} \left\{Q(f) + \lambda\|f\|\right\}, \quad \lambda > 0. \tag{22}$$

The next assumption is concerned with the approximation power of $\mathcal{H}^{\text{as}}$ to f_ϕ.

Assumption 4. There are positive constants c_2 and β such that

$$\mathcal{D}(\lambda) = Q\left(f_\lambda\right) - Q^* + \lambda\left\|f_\lambda\right\| \le c_2\lambda^\beta, \quad \forall\lambda > 0. \tag{23}$$

Recall $f_\phi(x,x') = -f_\phi(x',x)$. The above assumption is not too restrict.

Assumption 5. (i) The kernel K satisfies a Lipschitz condition of order γ with $0 < \gamma < 1$; that is, there exists some $c_3 > 0$ such that

$$\left|K(s,t) - K\left(s,t'\right)\right| \le c_3\left|t - t'\right|_2^\gamma. \tag{24}$$

(ii) The ranking loss has an increment exponent $\theta \ge 1$; that is, there exist some constants $\theta \ge 1, c_4$ such that

$$\phi(t) \le c_4(1 + |t|)^\theta, \quad \phi'_\pm(t) \le c_4(1 + |t|)^{\theta-1}, \quad t \in \mathbb{R}, \tag{25}$$

where $\phi'_\pm$ denotes the right- and left-sided derivatives of ϕ, respectively.

Assumption 6. The margin distribution ρ_X satisfies condition L_τ with $0 < \tau < \infty$; that is, for some $c_5 > 0$ and any ball $B(x,\delta) := \{u \in \mathcal{X} \mid |u - x|_2 < \delta\}$, one has

$$\rho_X(B(x,\delta)) \ge c_5\delta^\tau, \quad \forall x \in \mathcal{X}, 0 < \delta \le 1. \tag{26}$$

The last assumption concerns covering numbers. For a subset $\mathcal{S}$ of a space with pseudometric ρ and $\delta > 0$. The covering number $\mathcal{N}(\delta, \mathcal{S}, \rho)$ is defined to be the minimal number l such that there exist l disks with radius δ covering $\mathcal{S}$. When Σ is compact this number is finite.

Assumption 7. (i) There are some $\alpha > 0$ and $c_\alpha > 0$ such that

$$\mathcal{N}\left(\delta, \mathcal{X}, |\cdot|_2\right) \le c_6\left(\frac{1}{\delta}\right)^\alpha, \quad \forall\delta > 0. \tag{27}$$

(ii) For $R > 0$, let $B_R = \{f \in \mathcal{H}^{\text{as}} \mid \|f\| \le R\}$. There are some constant $s \in (0,1), c_7 > 0$ such that

$$\log\mathcal{N}\left(\delta, B_1, \|\cdot\|_\infty\right) \le c_7\delta^{-s}, \quad \forall\delta > 0. \tag{28}$$

It was shown in [18], under Assumptions 5(i), 6, and 7(i), that the following holds:

$$\log\mathcal{N}\left(\delta, \mathcal{X}, |\cdot|_2\right) \le c_6\left(\frac{4c_3}{\delta}\right)^{\alpha/\gamma}\log\left(2 + \frac{4\kappa}{\delta}\right), \quad \forall 0 < \delta < 1. \tag{29}$$

Therefor (ii) in Assumption 7 holds provided that $\alpha/\gamma < 1$.

We are in a position to state the main result of this paper. The proof is given in Section 5.

Theorem 8. *For any $\varepsilon \in (0, \beta)$, under the Assumptions 2–7, one has confidence at least $1 - C_\varepsilon e^{-t}$*

$$\begin{aligned} &Q(g_n) - Q^* + \lambda\Omega_z(f_{z,\lambda}) \\ &\quad \le C'\left(\max\left\{(\log n + t)^{\gamma/\tau}, t\right\}\right)^{(2-\alpha+s)/(2-\alpha)} n^{-(\beta-\varepsilon)\mu}, \end{aligned} \tag{30}$$

where $\mu = \min\{\gamma/\tau(\beta + (1-\beta)\theta), 1/(\beta + (1-\beta)\theta), 1/((2-\alpha)\beta + s)\}$ and C_ε, C' are constants independent of t or n.

The first step of the proof is to decompose $Q(g_n) - Q^* + \lambda\Omega_z(f_{z,\lambda})$ into errors of different types as the following:

$$Q(g_n) - Q^* + \lambda\Omega_z(f_{z,\lambda}) = S(\mathbf{z}, \lambda) + P(\mathbf{z}, \lambda) + D(\lambda), \tag{31}$$

where

$$S(\mathbf{z}, \lambda) = \{Q(g_n) - Q_n(g_n)\} + \{Q_n(f_\lambda) - Q(f_\lambda)\}, \tag{32}$$

referred to as sample error and

$$P(\mathbf{z}, \lambda) = \{Q_n(g_n) + \lambda\Omega_z(f_{z,\lambda})\} - \{Q_n(f_\lambda) + \lambda\|f_\lambda\|\}, \tag{33}$$

referred to as hypothesis error. We bound hypothesis error and sample error in the next two sections, respectively.

In the estimation of sample error, Hoeffding's decomposition of U-statistic, which breaks U-statistic into a sum of iid random variables and a degenerate U-statistic (see Section 4 for details), is a useful tool.

3. Hypothesis Error

In this section, we bound hypothesis error $P(\mathbf{z}, \lambda)$. This error is caused as we switch from the minimizer $f_{z,\lambda}$ of $Q_n(f) + \lambda\Omega_z(f)$ in $\mathcal{H}^{\text{as}}_{K,z}$ to the minimizer f_λ of $Q(f) + \lambda||f||$ in $\mathcal{H}^{\text{as}}$. Such errors are estimated in some papers, for example, [7, 18], and so forth. We note that, different from [18, 19], the underlying spaces $\mathcal{H}^{\text{as}}_{K,z}$ and $\mathcal{H}^{\text{as}}$ are sets of antisymmetric functions. We begin with the representations of the functions.

Lemma 9. *Let $f \in \mathcal{H}^{as}$. For any $\eta > 0$, one has a representation:*

$$\begin{aligned} f &= \frac{1}{2}\sum_{i=1}^{\infty} a_i\left(K_{(x_i,u_i)} - K_{(u_i,x_i)}\right), \quad x_i, u_i \in \mathcal{X}, \\ &\sum_{i=1}^{\infty}|a_i| \le \|f\| + \eta. \end{aligned} \tag{34}$$

Proof. For any $\eta > 0$, there are sequences $\{s_i\}_{i=1}^{\infty} \in \mathcal{X}^2$ and $\{a_i\}_{i=1}^{\infty} \in \ell^1$ such that $f = \sum_{i=1}^{\infty} a_i K_{s_i}$, $s_i \in \mathcal{X}^2$ and $\sum_{i=1}^{\infty}|a_i| \le \|f\| + \eta$.

Denote $s_i = (x_i, u_i)$, $x_i, u_i \in \mathcal{X}$, $i = 1, 2, \ldots$. It follows from (5) that

$$f(x', x) = \sum_{i=1}^{\infty} a_i K\left((u_i, x_i), (x, x')\right). \tag{35}$$

The proof is complete by $f(x, x') = 1/2(f(x, x') - f(x', x))$. □

A set $\{x_i\}_{i=1}^n \subset \mathcal{X}$ is said to be Δ-dense in $\mathcal{X}$ if for any $x \in \mathcal{X}$ there exists some $1 \le j \le n$ such that $|x - x_j|_2 < \Delta$.

Proposition 10 ([19], Proposition 3.1). *Let $\{X_j\}_{j=1}^n$ be drawn independently according to ρ_X. Then for any $t > 1$, under Assumption 6 and (i) in Assumption 7, with confidence at least $1 - e^{-t}$, $\{X_j\}_{j=1}^n$ is $c_{\alpha,\tau}((\log n + t)/n)^{\gamma/\tau}$-dense in $\mathcal{X}$, where $c_{\alpha,\tau}$ is a constant that depends only on α and τ.*

The hypothesis error $P(\mathbf{z}, \lambda)$ is bounded by the following proposition.

Proposition 11. *Assume Assumptions 5 and 6. Then for any $t > 1$, with confidence at least $1 - e^{-t}$, there holds*

$$\begin{aligned} P(\mathbf{z}, \lambda) &\le C\|f_\lambda\|_\infty\left(\frac{\log n + t}{n}\right)^{\gamma/\tau} \\ &\times\left(1 + \|f_\lambda\|_\infty + \|f_\lambda\|_\infty\left(\frac{\log n + t}{n}\right)^{\gamma/\tau}\right)^{\theta-1}, \end{aligned} \tag{36}$$

where C' is a constant independent of $\mathbf{z}, \lambda, m$, and t. (hereafter, C and C' are constants which are independent of R, t, n, or λ, and may changes from line to line.)

Proof. The proof follows the line of [19, 20]. For any $\eta > 0$, let the representation $f_\lambda = \sum_{i=1}^{\infty} b_i(K_{(x_i,u_i)} - K_{(u_i,x_i)})$ with $\sum_{i=1}^{\infty}|b_i| \le (\|f_\lambda\|_\infty + \eta)/2$ be given in Lemma 9.

By Proposition 11, with confidence at least $1 - e^{-t}$, for any $i = 1, 2, \ldots, n$, there are some $X_{j_i}, X_{k_i} \in \{X_i\}_{i=1}^n$ such that $\max\{|x_i - X_{j_i}|_2, |u_i - X_{k_i}|\} \le c_{\alpha,\tau}((\log n + t)/n)^{1/\tau}$. For an integer N, which will be determined later, denote $f = \sum_{i=1}^{N} b_i(K_{(X_{j_i},X_{k_i})} - K_{(X_{k_i},X_{j_i})}) \in \mathcal{H}^{\text{as}}_{K,z}$. So by Assumption 7,

$$\begin{aligned} &\left\|f - \sum_{i=1}^{N} b_i\left(K_{(x_i,u_i)} - K_{(u_i,x_i)}\right)\right\|_\infty \\ &\quad \le C\left(\|f_\lambda\| + \eta\right)\left(\frac{\log n + t}{n}\right)^{\gamma/\tau}, \end{aligned} \tag{37}$$

where $C = c_3 c^\gamma_{\alpha,\tau}$.

Choose N such that $\sum_{j=N+1}^{\infty}|b_j| < \eta/2$. Therefore

$$\left\|\sum_{i=1}^{N} b_i\left(K_{(x_i,u_i)} - K_{(u_i,x_i)}\right) - f_\lambda\right\|_\infty \le 2\kappa\sum_{j=N+1}^{\infty}\left|b_j\right| \le \kappa\eta. \tag{38}$$

Consequently, $\|f - f_\lambda\|_\infty \le D_\eta := C(\|f_\lambda\|_\infty + \eta)((\log n + t)/n)^{\gamma/\tau} + \kappa\eta$, which together with

$$\begin{aligned} &|Q_n(f_1) - Q_n(f_2)| \\ &\quad \le c_4(1 + \max\{\|f_1\|_\infty, \|f_2\|_\infty\})^{\theta-1}\|f_1 - f_2\|_\infty \end{aligned} \tag{39}$$

yields, with confidence at least $1 - e^{-t}$,

$$Q_n(f) \le Q_n(f_\lambda) + c_4\left(1 + \|f_\lambda\|_\infty + D_\eta\right)^{\theta-1} D_\eta. \tag{40}$$

On the other hand, since $f \in \mathscr{H}^{\mathrm{as}}_{K,z}$, the following holds by (12) and (9), with confidence at least $1 - e^{-t}$:

$$\begin{aligned} Q_n(g_n) + \lambda\Omega_{\mathbf{z}}(f_{z,\lambda}) &\le Q_n(f_{z,\lambda}) + \lambda\Omega_z(f_{z,\lambda}) \\ &\le Q_n(f) + 2\lambda\sum_{j=1}^{N}\left|b_j\right| \\ &\le Q_n(f) + \lambda\left(\|f_\lambda\| + \eta\right), \end{aligned} \tag{41}$$

which together with (40) completes the proof by letting $\eta \to 0$. □

The above bound for $P(\mathbf{z}, \lambda)$ is the same as in [20], both using of the same density of $\{X_i\}_{i=1}^n$ in $\mathscr{X}$. However, the functions considered there are defined on $\mathscr{X}$, instead of $\mathscr{X}^2$ as present.

4. Sample Error

As in [2, 7], Hoeffding's decomposition plays an important role in the estimation for sample error. For any $f(x, x'), x, x' \in \mathscr{X}$, denote

$$\begin{aligned} \varphi_f(z, z') &= \phi\left(\operatorname{sign}\left(y - y'\right) f\left(x, x'\right)\right), \\ z &= (x, y), \qquad z' = \left(x', y'\right). \end{aligned} \tag{42}$$

By Hoeffding's decomposition of U-statistic, we have

$$\begin{aligned} &Q\left(f - f_\phi\right) - Q_n\left(f - f_\phi\right) \\ &\quad = 2\left(Q\left(f - f_\phi\right) - P_n\left(P\varphi_f - P\varphi_{f_\phi}\right)\right) - U_n\left(h_f - h_{f_\phi}\right), \end{aligned} \tag{43}$$

where $h_f(z, z') = \varphi_f(z, z') - P\varphi_f(z) - P\varphi_f(z') + Q(f)$ and, for any $g(z, z')$,

$$\begin{aligned} Pg(z) &= \mathbb{E}\left(g\left(Z, Z'\right) \mid Z = z\right), \\ P_n(g) &= \frac{1}{n}\sum_{i=1}^{n} g\left(Z_i\right), \\ U_n(g) &= \frac{1}{n(n-1)}\sum_{i \ne j} g\left(Z_i, Z_j\right). \end{aligned} \tag{44}$$

Moreover, for any ranking rule f, we denote $g_f(z) = P\varphi_f(z) - P\varphi_{f_\phi}(z)$. Then

$$Q\left(f - f_\phi\right) - P_n\left(P\varphi_f - P\varphi_{f_\phi}\right) = \mathbb{E}\left(g_f\right) - P_n\left(g_f\right). \tag{45}$$

It is the deviation of sum of independent random variables from their mean.

As seen, in Hoeffding's decomposition (43), the first term is a sum of iid variables and the second term $U_n(h_f - h_{f_\phi})$ is a degenerate U-statistic. The degeneration means $\mathbb{E}(U_n(h_f(Z, Z') - h_{f_\phi}(Z, Z')) \mid Z) = 0, \forall Z$.

Denote

$$\begin{aligned} S_1 &= Q(g_n) - Q\left(f_\phi\right) - \left(Q_n(g_n) - Q_n\left(f_\phi\right)\right) \\ &= Q\left(g_n - f_\phi\right) - Q_n\left(g_n - f_\phi\right), \\ S_2 &= Q_n(f_\lambda) - Q_n\left(f_\phi\right) - \left(Q(f_\lambda) - Q\left(f_\phi\right)\right) \\ &= Q_n\left(f_\lambda - f_\phi\right) - Q\left(f_\lambda - f_\phi\right), \end{aligned} \tag{46}$$

so that the sample error $Q(g_n) - Q_n(g_n) + Q_n(f_\lambda) - Q(f_\lambda) = S_1 + S_2$.

We first estimate S_1. Since g_n depends on $\mathbf{z}$, by (43) and (45), we need to consider the suprema of the sets

$$\begin{aligned} &\left\{\mathbb{E}\left(g_{\pi(f)}\right) - P_n\left(g_{\pi(f)}\right) \mid F \in \mathscr{F}\right\}, \\ &\left\{\left|U_n\left(h_{\pi(f)} - h_{f_\phi}\right)\right| \mid F \in \mathscr{F}\right\}, \end{aligned} \tag{47}$$

where $\mathscr{F}$, containing g_n, is (a subset of) a ball in $\mathscr{H}^{\mathrm{as}}$.

With the above decomposition and the methods in [21], the following proposition is established for the ball $\{f \in \mathscr{H}^{\mathrm{as}}_K \mid \|f\|_K \le R\}$ and $\phi(t) = (1 - t)_+$ in [7]. The Assumption 2 and the condition on the covering number of $\{f \in \mathscr{H}^{\mathrm{as}}_K \mid \|f\|_K \le R\}$ play a crucial role. The arguments also work in the present setting; that is, for the ball $B_R = \{f \in \mathscr{H}^{\mathrm{as}} \mid \|f\| \le R\}$. For this we note that ϕ satisfies $|\phi(t) - \phi(t')| \le C|t - t'|, t, t' \in [-1, 1]$. Therefore, under (ii) of Assumption 5 and (ii) of Assumption 7, the covering number $\mathscr{N}(\delta, \mathscr{G}, \|\cdot\|_\infty)$ of $\mathscr{G} = \{g_{\pi(f)} \mid f \in B_R\}$ satisfies $\log \mathscr{N}(\delta, \mathscr{G}, \|\cdot\|_\infty) \le C(R/\delta)^s$. The interested reader may refer to [7, 21] for the details.

Proposition 12. *Let $R > 0$ and $t > 0$. Under Assumption 2, (ii) of Assumption 5, and (ii) of Assumption 7, one has confidence at least $1 - e^{-t}$,*

$$\mathbb{E}\left(g_{\pi(f)}\right) - P_n\left(g_{\pi(f)}\right) \le \delta_0 + \delta_0^{1-(\alpha/2)}\left(\mathbb{E}\left(g_{\pi(f)}\right)\right)^{(\alpha/2)}, \quad \forall f \in B_R, \tag{48}$$

where δ_0 is bounded by

$$\delta_0 \le C\left(\left(\frac{t}{n}\right)^{1/(2-\alpha)} + \left(\frac{R^s}{n}\right)^{1/(2-\alpha+s)}\right), \tag{49}$$

with C being a constant independent of R, t, and n.

The estimation for the supremum $\sup_{f \in \mathscr{F}_R}|U_n(h_{\pi(f)} - h_{f_\phi})|$ of U-process is much involved. The supremum of U-processes has been studied in a few papers. The following lemma follows from the proof of ([22, Theorem 3.2, $m = 2$]).

Lemma 13. *Suppose that a function class $\mathscr{F}$ satisfies the following conditions.*

(i) *For any* $f \in \mathcal{F}$, $f(z,z') = f(z',z)$ *and* $\mathbb{E}(f(Z,Z') \mid Z) = 0$.

(ii) $\mathcal{F}$ *is uniformly bounded by a universal constant* C_0.

(iii) $C_{\mathcal{F}} := \int_0^\infty \log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty) d\delta < \infty$.

Then

$$\mathbb{E} \exp\left(\lambda \sqrt{\Gamma_n}\right) \le C \exp\left(C' \lambda^2 C_{\mathcal{F}}\right), \quad \forall \lambda > 0, \tag{50}$$

where C, C' *are some constants, independent of* $C_{\mathcal{F}}$, *and*

$$\Gamma_n = \sup_{f \in \mathcal{F}} \left|(n-1) U_n(f)\right|. \tag{51}$$

Proposition 14. *Suppose that (ii) in Assumption 7 holds. Then one has with confidence at least* $1 - Ce^{-t}$

$$\sup_{f \in \mathcal{F}_R} \left|U_n\left(h_{\pi(f)} - h_{f_\phi}\right)\right| \le \frac{C'\left(R^s + 1\right) t}{n}, \tag{52}$$

where $\mathcal{F}_R = \{h_{\pi(f)} \mid f \in B_R\}$ *and* C, C' *are some positive constants.*

Proof. We first claim that $\mathcal{F}_R$ satisfies the conditions in Lemma 13. Indeed, condition (i) holds by definition of $h_{\pi(f)}$ and $f(x,x') = -f(x',x)$ for any $f \in B_R$. Also, by definition of h_f, for any $f \in \mathcal{F}_R$, $\|h_{\pi(f)}\|_\infty \le 4\phi(-1)$, implying (ii). Moreover, by (ii) of Assumption 5

$$\left\|h_{\pi(f_1)} - h_{\pi(f_2)}\right\|_\infty \le C\left\|\pi(f_1) - \pi(f_2)\right\|_\infty \le C\left\|f_1 - f_2\right\|_\infty \tag{53}$$

we have

$$\log \mathcal{N}\left(\mathcal{F}_R, \delta, \|\cdot\|_\infty\right) \le \log \mathcal{N}\left(B_R, \frac{\delta}{C}, \|\cdot\|_\infty\right) \le c_7 (CR)^s \delta^{-s} \tag{54}$$

provided (ii) in Assumption 7. This establishes (iii), as claimed.

Applying Markov's inequality to $\exp(\lambda\sqrt{\Gamma_n})$, with $\lambda = \sqrt{t}/2C'R^s$ and appealing to Lemma 13, we have $\mathbb{P}(\Gamma_n > t) \le Ce^{-t/(4C'R^s)}$, $\forall t > 0$; that is,

$$\mathbb{P}\left(\Gamma_n > 4C'R^s t\right) \le Ce^{-t}, \quad \forall t > 0. \tag{55}$$

For single function h_{f_ϕ}, it is known in ([23, Proposition 2.3]) that

$$\mathbb{P}\left(\left|U_n\left(h_{f_\phi}\right)\right| \ge \frac{C't}{n-1}\right) \le Ce^{-t}, \tag{56}$$

which together with (55) completes the proof. □

The estimation for S_2 is easy since f_λ does not change with the set $\mathbf{z}$ of samples.

Proposition 15. *Assume Assumption 2. For any* $t > 0$, *one has with confidence at least* $1 - Ce^{-t}$

$$S_2 \le C' t \left(\frac{\left(1 + \kappa\left\|f_\lambda\right\|_\infty\right)^\theta}{3n} + \left(\frac{1}{n}\right)^{1/(2-\alpha)}\right) + D(\lambda), \tag{57}$$

where C, C' *are some constants, independent of* λ, t *and* n.

Proof. Clearly, the function $g_{f_\lambda}(Z)$ satisfies $\|g_{f_\lambda}\|_\infty \le C(1 + \|f_\lambda\|_\infty)^\theta$. Then, by Assumption 2, we conclude, as in [20, 21], with confidence at least $1 - e^{-t}$, that

$$\begin{aligned} &P_n\left(g_{f_\lambda}\right) - \left(Q\left(f_\lambda\right) - Q^*\right) \\ &\le C' t \left(\frac{\left(1 + \left\|f_\lambda\right\|_\infty\right)^\theta}{3n} + \left(\frac{1}{n}\right)^{1/(2-\alpha)}\right) + D(\lambda). \end{aligned} \tag{58}$$

It remains to estimate $U_n(h_{f_\lambda} - h_{f_\phi})$. For any single function g with $g(z,z') = g(z',z)$ and $\mathbb{E}[g(Z,Z') \mid Z] = 0, \forall Z$, we have by [23, Proposition 2.3]

$$\mathbb{P}\left(\left|U_n(g)\right| \ge \frac{C't\|g\|_\infty}{n-1}\right) \le Ce^{-t}, \tag{59}$$

which together with $\|h_{f_\lambda} - h_{f_\phi}\|_\infty \le C'(1 + \|f_\lambda\|_\infty)^\theta$ implies, with confidence at least $1 - Ce^{-t}$, that

$$\left|U_n\left(h_{f_\lambda} - h_{f_\phi}\right)\right| \le \frac{C'\left(1 + \left\|f_\lambda\right\|_\infty\right)^\theta t}{n}, \tag{60}$$

where C and C' are constants. The proof is complete. □

5. Proof of Theorem 8

Theorem 16. *For* R *such that* $\Omega_z(f_{z,\lambda}) \le R$ *and any* $t > 1$, *under Assumptions 2–7, one has with confidence at least* $1 - Ce^{-t}$

$$\begin{aligned} &Q(g_n) - Q^* + \lambda \Omega_z\left(f_{z,\lambda}\right) \\ &\le C'\left(\left(\frac{\log n + t}{n}\right)^{\gamma/\tau} \lambda^{(\beta-1)\theta} + \left(\frac{t}{n}\right)^{1/(2-\alpha)}\right. \\ &\qquad \left. + \left(\frac{R^s}{n}\right)^{1/(2-\alpha+s)} + \frac{\left(R^s + 1\right) t}{n} + \frac{t\lambda^{(\beta-1)\theta}}{n} + \lambda^\beta\right), \end{aligned} \tag{61}$$

where C, C' *are constant independent of* $R, t,$ *or* n.

Proof. We note that

$$\left\|f_\lambda\right\|_\infty \le \kappa \left\|f_\lambda\right\| \le \frac{\kappa D(\lambda)}{\lambda} \le C\lambda^{\beta-1}. \tag{62}$$

By $\Omega_z(f_{z,\lambda}) \le R$, a combination of Propositions 11, 12, 14, and 15 yields that, with confidence at least $1 - Ce^{-t}$,

$$\begin{aligned} &Q(g_n) - Q^* + \lambda \Omega_z\left(f_{z,\lambda}\right) \\ &\le C_1\left(\frac{\log n + t}{n}\right)^{\gamma/\tau} \lambda^{(\beta-1)\theta} + \delta_0 + \delta_0^{1-\alpha/2}\left\{Q(f) - Q^*\right\}^{\alpha/2} \\ &\quad + \frac{C_2\left(R^s + 1\right) t}{n} + C_3 t\left(\frac{\lambda^{(\beta-1)\theta}}{3n} + \left(\frac{1}{n}\right)^{1/(2-\alpha)}\right) + 2\lambda^\beta, \end{aligned} \tag{63}$$

where $C_i, i = 1, 2, 3$, are constants, and δ_0 is bounded by (49).

Putting $x = Q(g_n) - Q^* + \lambda\Omega_z(f_{z,\lambda})$ and $\nu = \alpha/2$ into the implication relation

$$x \le ax^{\nu} + b, a, b, x > 0 \Longrightarrow x \le \max\left\{(2a)^{1/(1-\nu)}, 2b\right\}, \quad (64)$$

we obtain (61) from (63). Therefore, the conditional probability of the event that inequality (61) holds, given the event $\Omega_z(f_{z,\lambda}) \le R$, is at least $1 - Ce^{-t}$. The proof is complete. □

For any R, denote the random event $\Omega_z(f_{z,\lambda}) \le R$ by ξ_R. Obviously, $\mathbb{P}(\xi_R) = 1$ for $R = \phi(0)/\lambda$. However, to prove Theorem 8, a smaller R with $\mathbb{P}(\xi_R) = 1$ is desired. To this end, we apply the iteration technique for estimation of $\Omega_z(f_{z,\lambda})$ introduced in [17].

Recall that μ is given in Theorem 8. It is easily seen that, for $\lambda = n^{-\mu}, R \le n^{1/s}$,

$$\left(\frac{1}{n}\right)^{\gamma/\tau}\lambda^{(\beta-1)\theta} \le \lambda^{\beta}, \qquad \frac{1}{n}\lambda^{(\beta-1)\theta} \le \lambda^{\beta},$$

$$\left(\frac{1}{n}\right)^{1/(2-\alpha)} \le \frac{(R^s+1)}{n} \le \left(\frac{R^s}{n}\right)^{1/(2-\alpha+s)} \le \lambda^{\beta}\left(\lambda^{1-\beta}R\right)^{s/(2-\alpha+s)}. \quad (65)$$

Therefore, we have, by Theorem 16, with the conditional probability at least $1 - Ce^{-t}$, given ξ_R,

$$Q(g_n) - Q^* + \lambda\Omega_z(f_{z,\lambda}) \le C' \max\left\{(\log n + t)^{\gamma/\tau}, t\right\}\lambda^{\beta}\left(\left(\lambda^{1-\beta}R\right)^{s/(2-\alpha+s)} + 2\right). \quad (66)$$

If $\lambda^{\beta-1}R^{-1} = O(1)$, the above inequality becomes

$$Q(g_n) - Q^* + \lambda\Omega_z(f_{z,\lambda}) \le C' \max\left\{(\log n + t)^{\gamma/\tau}, t\right\}\lambda^{\beta}\left(\lambda^{1-\beta}R\right)^{s/(2-\alpha+s)}. \quad (67)$$

Consequently, given event ξ_R with $\lambda^{\beta-1}R^{-1} = O(1), R \le n^{1/s}$, we have with confidence at least $1 - Ce^{-t}$,

$$\|f_{z,\lambda}\| \le \Omega_z(f_{z,\lambda}) \le r(R) := C' \max\left\{(\log n + t)^{\gamma/\tau}, t\right\} \times \lambda^{(\beta-1)((1-s)/(2-\alpha+s))}R^{s/(2-\alpha+s)}. \quad (68)$$

Let $R^{(0)} = \phi(0)/\lambda, R^{(k)} = r(R^{(k-1)}), k = 1, 2, \ldots$. By induction, it is easy to prove $\lambda^{\beta-1}(R^{(k)})^{-1} = O(1), R^{(k)} \le n^{1/s}$. Since $\mathbb{P}\{\xi_{R_0}\} = 1$, we have with confidence at least $1 - kCe^{-t}$ $\Omega_z(f_{z,\lambda}) \le R^{(k)}, k = 1, 2, \ldots,$. Clearly,

$$R^{(k)} \le \left(C' \max\left\{(\log n + t)^{\gamma/\tau}, t\right\}\right)^{(1-\nu^{k+1})/(1-\nu)} \times \lambda^{\beta-1}\lambda^{-(1/(2-\alpha+s))^{k+1}}, \quad k = 1, 2, \ldots, \quad (69)$$

where $\nu = s/(2 - \alpha + s) < 1$.

For any $\varepsilon > 0$, let k be the smallest integer such that $(1/(2 - \alpha + s))^{k+1} < \varepsilon$. Substituting $R = R^{(k)}$ into (67), we bound the right hand side of (67) by $(C' \max\{(\log n + t)^{\gamma/\tau}, t\})^{1/(1-\nu)}\lambda^{\beta-\varepsilon}$, with confidence at least $1 - C_\varepsilon e^{-t}$, where $C_\varepsilon = kC$. This completes the proof.

Acknowledgment

The authors thank Professor Di-Rong Chen for his help.

References

[1] S. Clémençon, G. Lugosi, and N. Vayatis, "Ranking and empirical minimization of U-statistics," *The Annals of Statistics*, vol. 36, no. 2, pp. 844–874, 2008.

[2] W. Rejchel, "On ranking and generalization bounds," *Journal of Machine Learning Research*, vol. 13, pp. 1373–1392, 2012.

[3] Y. Lin, "Support vector machines and the Bayes rule in classification," *Data Mining and Knowledge Discovery*, vol. 6, no. 3, pp. 259–275, 2002.

[4] D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou, "Support vector machine soft margin classifiers: error analysis," *Journal of Machine Learning Research*, vol. 5, pp. 1143–1175, 2003/04.

[5] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[6] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, vol. 24, Cambridge University Press, Cambridge, UK, 2007.

[7] H. Chen and J. T. Wu, "Supportvecor machine for ranking," *submitted*.

[8] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proceedings of the 15th International Conference on Machine Learning (ICML '98)*, J. Shavlik, Ed., Morgan Kaufmann, 1998.

[9] M. Song, C. M. Breneman, J. Bi et al., "Prediction of protein retention times in anion-exchange chromatography systems using support vector regression," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 6, pp. 1347–1357, 2002.

[10] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[11] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," *Advances in Neural Information Processing Systems*, vol. 16, pp. 49–56, 2004.

[12] B. Tarigan and S. A. van de Geer, "Classifiers of support vector machine type with ℓ^1 complexity regularization," *Bernoulli*, vol. 12, no. 6, pp. 1045–1076, 2006.

[13] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

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

[15] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive Approximation*, vol. 26, no. 2, pp. 153–172, 2007.

[16] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[17] I. Steinwart and C. Scovel, "Fast rates for support vector machines using Gaussian kernels," *The Annals of Statistics*, vol. 35, no. 2, pp. 575–607, 2007.

[18] Q.-W. Xiao and D.-X. Zhou, "Learning by nonsymmetric kernels with data dependent spaces and ℓ^1-regularizer," *Taiwanese Journal of Mathematics*, vol. 14, no. 5, pp. 1821–1836, 2010.

[19] H. Tong, D.-R. Chen, and F. Yang, "Support vector machines regression with ℓ^1-regularizer," *Journal of Approximation Theory*, vol. 164, no. 10, pp. 1331–1344, 2012.

[20] H. Tong, D.-R. Chen, and F. Yang, "Learning rates for ℓ^1-regularized kernel classifiers," *Journal of Applied Mathematics*, vol. 2013, Article ID 496282, 11 pages, 2013.

[21] H. Tong, D.-R. Chen, and L. Peng, "Analysis of support vector machines regression," *Foundations of Computational Mathematics*, vol. 9, no. 2, pp. 243–257, 2009.

[22] M. A. Arcones and E. Giné, "U-processes indexed by Vapnik-Červonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters," *Stochastic Processes and their Applications*, vol. 52, no. 1, pp. 17–38, 1994.

[23] M. A. Arcones and E. Giné, "Limit theorems for U-processes," *The Annals of Probability*, vol. 21, no. 3, pp. 1494–1542, 1993.