Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Approximation Theory 164 (2012) 1331–1344

JOURNAL OF Approximation Theory

www.elsevier.com/locate/jat

Full length article

Support vector machines regression with l^1-regularizer☆

Hongzhi Tong[a,*], Di-Rong Chen[b], Fenghong Yang[c]

[a] *School of Information Technology & Management, University of International Business and Economics, Beijing 100029, PR China*

[b] *Department of Mathematics and LMIB, Beijing University of Aeronautics and Astronautics, Beijing 100083, PR China*

[c] *School of Applied Mathematics, Central University of Finance and Economics, Beijing 100081, PR China*

Received 1 November 2010; received in revised form 6 June 2012; accepted 19 June 2012
Available online 28 June 2012

Communicated by Ding-Xuan Zhou

Abstract

The classical support vector machines regression (SVMR) is known as a regularized learning algorithm in reproducing kernel Hilbert spaces (RKHS) with a ε-insensitive loss function and an RKHS norm regularizer. In this paper, we study a new SVMR algorithm where the regularization term is proportional to l^1-norm of the coefficients in the kernel ensembles. We provide an error analysis of this algorithm, an explicit learning rate is then derived under some assumptions.

Keywords: Support vector machines regression; Coefficient regularization; Learning rate; Reproducing kernel Hilbert spaces; Error decomposition

1. Introduction

Let X be a compact subset of $\mathbb{R}^n$, $Y \subset [-M, M]$ for some $M > 0$. The relation between the input $x \in X$ and the output $y \in Y$ is described by a probability distribution ρ on $Z := X \times Y$,

☆ Research supported by NSF of China under grants 10871015, 10872009 and National Basic Research Program of China under grant 973-2006CB303102.

* Corresponding author.
E-mail addresses: tonghz@uibe.edu.cn (H. Tong), drchen@buaa.edu.cn (D.-R. Chen), fhyang@cufe.edu.cn (F. Yang).

0021-9045/$ - see front matter
doi:10.1016/j.jat.2012.06.005

but ρ is known only through a set of samples $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ independently drawn according to ρ. Given samples $\mathbf{z}$ the regression problem in learning theory aims at finding a function $f_{\mathbf{z}} : X \to \mathbb{R}$, such that $f_{\mathbf{z}}(x)$ is a satisfactory estimate of output y when a new input x is given.

Support vector machines regression (SVMR) used the ε-insensitive loss function (ILF)

$$V(y, f(x)) = |y - f(x)|_\varepsilon = \begin{cases} 0, & \text{if } |y - f(x)| < \varepsilon, \\ |y - f(x)| - \varepsilon, & \text{otherwise,} \end{cases} \tag{1.1}$$

to measure the cost paid by replacing the true y with the estimate $f(x)$. In [11], an interpretation of ILF for SVMR is presented. It demonstrates that it is appropriate to use the ILF rather than the quadratic loss function (QLF) used in least square regression (LSR) (see e.g. [5,10,8,17]) under the assumption that the noise affecting the data is additive and Gaussian, but not necessarily zero mean, and that its variance and mean are random variables, the mean has a distribution which is uniform in the interval $[-\varepsilon, \varepsilon]$. This also helps us to understand the role of the parameter ε in (1.1).

The error for a measurable function f is measured by the expected risk

$$\mathcal{E}(f) := \int_Z V(y, f(x)) d\rho = \int_X \int_Y V(y, f(x)) d\rho(y|x) d\rho_X(x),$$

where ρ_X is the marginal distribution on X and $\rho(\cdot|x)$ is the conditional probability measure at x induced by ρ. We will denote

$$f^* := \arg\min \mathcal{E}(f),$$

where the minimum is taken over all measurable functions. Obviously f^* is our ideal estimator and it is often called the target function. By Theorem 4.1 in [15], we know f^* exists and

$$|f^*(x)| \le M + \varepsilon, \quad \forall x \in X.$$

We are interested in the kernel based learning algorithms. Recall a Mercer kernel K on $X \times X$ which is continuous, symmetric and positive semidefinite, i.e. for any finite set of distinct points $\{x_1, x_2, \ldots, x_l\} \subset X$, the matrix $\left(K(x_i, x_j)\right)_{i,j=1}^l$ is positive semidefinite. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with a Mercer kernel K is defined (see [1]) as the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying

$$\langle K_x, K_u \rangle_K = K(x, u),$$

and the reproducing property is given by

$$\langle K_x, f \rangle_K = f(x), \quad \forall x \in X, f \in \mathcal{H}_K.$$

The classical SVMR (see e.g. [4,10,16]) is then given by the following scheme:

$$\tilde{f}_{\mathbf{z},\lambda} := \arg \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_K^2 \right\}, \tag{1.2}$$

where $\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^m V(y_i, f(x_i))$ is the empirical error with respect to $\mathbf{z}$. The term $\lambda \|f\|_K^2$ is called the regularization term, λ is the regularization parameter, which is usually chosen to be some function of m and $\lim_{m \to \infty} \lambda(m) = 0$, $\|f\|_K^2$ is the regularizer.

The theoretical analysis of the regularized learning algorithms with the square RKHS norm regularizer is well understood (see e.g. [5,6]). Especially [15] gave a quantitative convergence result for scheme (1.2). In this paper, we consider a different regularized SVMR algorithm. In our setting, the regularizer is rather than an RKHS norm but a l^1-norm of the coefficients in the kernel ensembles.

Definition 1.1. Let

$$\mathcal{H}_{K,\mathbf{z}} := \left\{ \sum_{i=1}^{m} a_i K_{x_i} : a_i \in \mathbb{R}, i = 1, 2, \ldots, m \right\},$$

and

$$\Omega_{\mathbf{z}}(f) = \inf \left\{ \sum_{i=1}^{m} |a_i| : f = \sum_{i=1}^{m} a_i K_{x_i} \right\}.$$

Then SVMR with l^1-regularizer is given as

$$f_{\mathbf{z},\lambda} := \arg \min_{f \in \mathcal{H}_{K,\mathbf{z}}} \{\mathcal{E}_{\mathbf{z}}(f) + \lambda \Omega_{\mathbf{z}}(f)\}. \tag{1.3}$$

l^1-coefficient regularization generally leads to sparse representation, that is, it tends to result in a solution with a few non-zero coefficients (see [9,21]), thus these methods have attracted much attention recently (see e.g. [2,7,14]). However, it should be noticed that there exist essential differences between algorithm (1.2) and (1.3). For example, in (1.3) both hypothesis space $\mathcal{H}_{K,\mathbf{z}}$ and regularizer $\Omega_{\mathbf{z}}(f)$ are dependent of samples $\mathbf{z}$, this causes a consequence that the classical error decomposition approach (see [15]) cannot be applied to (1.3) any longer. There are some studies in the theoretical analysis of the l^1-regularized learning algorithms, but the research is not very rich yet. Wu and Zhou [18] provide an error analysis of l^1-regularized support vector machines classification (SVMC) by setting a stepping-stone between the linear programming SVMC and the classical quadratic programming SVMC, but it seems that this approach cannot be applied to the loss functions other than hinge loss used in SVMC. A general analysis framework is established in [19] for learning algorithms with sample dependent hypothesis space and coefficient based regularization, they introduce a modified error decomposition technique by means of an extra hypothesis error, while the sample errors and learning rates have not been considered there. Xiao and Zhou [20] studied LSR with l^1-regularizer, its key ideas for bounding hypothesis error are from [19]. Unlike the QLF used in LSR, the ILF is not differentiable. It causes some new technical difficulties, for example, an explicit expression of the solution $f_{\mathbf{z},\lambda}$ or its coefficient $a_{\mathbf{z}}$ like Theorem 3.1 in [13] or Theorem 4 in [19] is not derived. As the same time, the inequality (5.2) used to bound the sample error in [20] is not valid for ILF. We overcome this difficulty by introducing a more general inequality under assumption (4.1) (see Section 4 below).

In this paper, we adopt the ideas from [19] and provide an error analysis of scheme (1.3), we will mainly focus on estimating the excess risk

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*), \tag{1.4}$$

where $\pi(\cdot)$ is a projection operator defined in Section 2. The rest of paper is organized as follows. In Section 2, we introduce the notations and some preliminary results. We bound the hypothesis error and sample error in Sections 3 and 4, respectively. An explicit learning rate of scheme (1.3) is derived eventually in Section 5.

2. Preliminaries

We will write $|\cdot|_2$ for the Euclidean norm in $\mathbb{R}^n$. We denote $L^1_{\rho_X}$ the measurable functions on X with norm $\|f\|_{L^1_{\rho_X}} := \int_X |f(x)| d\rho_X(x)$, we also denote $C(X)$ as the space of continuous functions on X with the uniform norm $\|\cdot\|_\infty$.

Since the target function $|f^*(x)| \le M + \varepsilon$, it is reasonable to restrict its approximation functions to those also contained in $[-M-\varepsilon, M+\varepsilon]$.

Definition 2.1. The projection operator $\pi = \pi_{M+\varepsilon}$ is defined on the space of measurable functions $f : X \to \mathbb{R}$ as

$$\pi(f)(x) = \begin{cases} M+\varepsilon, & \text{if } f(x) > M+\varepsilon, \\ -M-\varepsilon, & \text{if } f(x) < -M-\varepsilon, \\ f(x), & \text{if } -M-\varepsilon \le f(x) \le M+\varepsilon. \end{cases}$$

Since $V(y, \pi(f)(x)) \le V(y, f(x))$, we have that

$$\mathcal{E}(\pi(f)) \le \mathcal{E}(f), \qquad \mathcal{E}_{\mathbf{z}}(\pi(f)) \le \mathcal{E}_{\mathbf{z}}(f). \tag{2.1}$$

So we take $\pi(f_{\mathbf{z},\lambda})$ instead of $f_{\mathbf{z},\lambda}$ as our empirical target function and analyze the related learning rates.

We have seen that the hypothesis space $\mathcal{H}_{K,\mathbf{z}}$ depends on samples, to characterize the approximation ability (independent of samples) of algorithm (1.3), we adopt the idea in [19] of using a larger function space containing all of the possible hypothesis spaces.

Definition 2.2. Banach space $\mathcal{H}$ is defined as the function set on X containing all functions of the form

$$f = \sum_{j=1}^{\infty} a_j K_{x_j}, \qquad \{a_j\}_{j=1}^{\infty} \in l^1, \qquad \{x_j\}_{j=1}^{\infty} \subset X,$$

with the norm

$$\|f\| := \inf\left\{ \sum_{j=1}^{\infty} |a_j| : f = \sum_{j=1}^{\infty} a_j K_{x_j} \right\}.$$

Obviously,

$$\mathcal{H}_{K,\mathbf{z}} \subset \mathcal{H}, \quad \forall \mathbf{z} \in Z^m.$$

By the continuity of K and compactness of X, we have

$$\kappa := \sup_{x \in X} K(x,x) < \infty.$$

It implies that $\mathcal{H}$ is a subset of $C(X)$, and

$$\|f\|_\infty \le \kappa \|f\|, \quad \forall f \in \mathcal{H}. \tag{2.2}$$

$\mathcal{H}$ is called the universal hypothesis space associated with scheme (1.3). The approximation error of f^* in $\mathcal{H}$ is defined as

$$D(\lambda) := \inf_{f \in \mathcal{H}} \{\mathcal{E}(f) - \mathcal{E}(f^*) + \lambda \|f\|\}.$$

$D(\lambda)$ is independent of sample, it is usually estimated by rich knowledge from approximation theory (see [12]). It is easy to see that $\mathcal{E}(f) - \mathcal{E}(f^*) \le \|f - f^*\|_{L^1_{\rho_X}}$. Hence $D(\lambda)$ concerns the approximation of f^* in $L^1_{\rho_X}$ by functions from $\mathcal{H}$, it can be characterized by requiring f^* to lie in some interpolation spaces of the pair $(L^1_{\rho_X}, \mathcal{H})$ (see e.g. [3,15]).

Definition 2.3. We say the target function f^* can be approximated in $\mathcal{H}$ with exponent $0 < \beta \le 1$ if there exists a constant c_β such that

$$D(\lambda) \le c_\beta \lambda^\beta, \quad \forall \lambda > 0. \tag{2.3}$$

To formulate a error decomposition of (1.4), we introduce

$$f_\lambda := \arg\min_{f \in \mathcal{H}} \{\mathcal{E}(f) + \lambda \|f\|\},$$

and thus

$$D(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f^*) + \lambda \|f_\lambda\|. \tag{2.4}$$

Now we can give the following error decomposition for the excess risk (1.4).

Proposition 2.1. *Let*

$$S(\mathbf{z}, \lambda) := \left\{\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\right\} + \left\{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\right\},$$

and

$$P(\mathbf{z}, \lambda) := \left\{\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\right\} - \left\{\mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda \|f_\lambda\|\right\}.$$

Then we have

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) \le S(\mathbf{z}, \lambda) + P(\mathbf{z}, \lambda) + D(\lambda).$$

Proof. Since

$$\begin{aligned}
&\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) \\
&\quad \le \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \\
&\quad = \left\{\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\right\} + \left\{\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\right\} \\
&\qquad + \left\{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\right\} - \left\{\mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda \|f_\lambda\|\right\} + \left\{\mathcal{E}(f_\lambda) - \mathcal{E}(f^*) + \lambda \|f_\lambda\|\right\} \\
&\quad = S(\mathbf{z}, \lambda) + P(\mathbf{z}, \lambda) + D(\lambda),
\end{aligned}$$

the conclusion is proved. □

$S(\mathbf{z}, \lambda)$ is usually called the sample error, and $P(\mathbf{z}, \lambda)$ is called the hypothesis error. We will estimate them respectively in the next two sections.

3. Hypothesis error

The hypothesis error $P(\mathbf{z}, \lambda)$ is caused by replacing the hypothesis space $\mathcal{H}_{K,\mathbf{z}}$ by the universal hypothesis space $\mathcal{H}$. In order to estimate it, we need to give some assumptions on input X, kernel function K and marginal distribution ρ_X.

Definition 3.1. Let $\mathcal{F}$ be a subset of a metric space. For any $r > 0$ the covering number $\mathcal{N}(\mathcal{F}, r)$ is defined to be the minimal integer $l \in \mathbb{N}$ such that there exist l balls with radius r covering $\mathcal{F}$.

Covering numbers are used to describe the complexity of X, we shall assume that for some $\alpha > 0$ and $c_\alpha > 0$,

$$\mathcal{N}(X, r) \leq c_\alpha (1/r)^\alpha, \quad \forall r > 0. \tag{3.1}$$

Definition 3.2. We say the kernel K satisfies a Lipschitz condition of order γ with $0 < \gamma \leq 1$ if there exists some $c_\gamma > 0$ such that

$$|K(x, u) - K(x, u')| \leq c_\gamma |u - u'|_2^\gamma, \quad \forall x, u, u' \in X. \tag{3.2}$$

Definition 3.3. The marginal distribution ρ_X is said to satisfy condition L_τ with $0 < \tau < \infty$ if for some $c_\tau > 0$ and any ball $B(x, r) := \{u \in X : |u - x|_2 < r\}$, one has

$$\rho_X(B(x, r)) \geq c_\tau r^\tau, \quad \forall x \in X, 0 < r \leq 1. \tag{3.3}$$

Definition 3.4. A set $\{x_1, x_2, \ldots, x_m\} \subset X$ is said to be Δ-dense if for any $x \in X$ there exists some $1 \leq i \leq m$ such that $|x - x_i|_2 < \Delta$.

By Lemma 3 in [20], we can get the following.

Proposition 3.1. *If X satisfies (3.1), ρ_X satisfies (3.3), and $\{x_i\}_{i=1}^m$ is drawn independently according to ρ_X. Then for any $t > 1$, with confidence at least $1 - e^{-t}$, $\{x_i\}_{i=1}^m$ is $c_{\alpha,\tau} \left(\frac{\log m + t}{m}\right)^{\frac{1}{\tau}}$-dense, where $c_{\alpha,\tau}$ is a constant depends only on α and τ.*

Proof. We know from [20, Lemma 3] that if ρ_X satisfies condition L_τ with $\tau > 0$ and Δ satisfies

$$-\log \mathcal{N}\left(X, \frac{\Delta}{2}\right) + \frac{m c_\tau \Delta^\tau}{2^\tau} \geq t, \tag{3.4}$$

then $\{x_i\}_{i=1}^m$ is Δ-dense with confidence at least $1 - e^{-t}$. So what we only to do is finding a solution Δ to (3.4). To this end, we consider a strictly increasing function

$$h_1(v) := -\log \mathcal{N}\left(X, \frac{v}{2}\right) + \frac{m c_\tau v^\tau}{2^\tau}, \quad v \in (0, +\infty).$$

Let v^* is the unique positive solution of the equation $h_1(v) = t$, by assumption (3.1),

$$t = h_1(v^*) \geq \frac{m c_\tau v^{*\tau}}{2^\tau} - \log\left(c_\alpha \left(\frac{2}{v^*}\right)^\alpha\right).$$

If $v^* > 2\left(\frac{1}{m}\right)^{\frac{1}{\tau}}$, we have

$$t \geq \frac{m c_\tau v^{*\tau}}{2^\tau} - \log c_\alpha - \frac{\alpha}{\tau} \log m.$$

Hence

$$\begin{aligned}\nu^* &\le 2\left(\frac{\log c_\alpha + \frac{\alpha}{\tau}\log m + t}{m c_\tau}\right)^{\frac{1}{\tau}} \\ &\le 2\left[\left(\frac{\log c_\alpha + \frac{\alpha}{\tau} + 1}{c_\tau}\right)^{\frac{1}{\tau}} + 1\right]\left(\frac{\log m + t}{m}\right)^{\frac{1}{\tau}}. \end{aligned} \tag{3.5}$$

If $\nu^* \le 2\left(\frac{1}{m}\right)^{\frac{1}{\tau}}$, (3.5) still holds.

Consequently by taking $\Delta = c_{\alpha,\tau}\left(\frac{\log m+t}{m}\right)^{\frac{1}{\tau}}$ with $c_{\alpha,\tau} := 2\left[\left(\frac{\log c_\alpha + \frac{\alpha}{\tau}+1}{c_\tau}\right)^{\frac{1}{\tau}} + 1\right]$, we get $\nu^* \le \Delta$, and thus Δ satisfies (3.4). □

We now bound $P(\mathbf{z}, \lambda)$ by the following theorem.

Theorem 3.1. *If X satisfies (3.1), K satisfies (3.2), and ρ_X satisfies (3.3), then for any $t > 1$, with confidence at least $1 - e^{-t}$, there holds*

$$P(\mathbf{z}, \lambda) \le C_1\left(\frac{\log m + t}{m}\right)^{\frac{\gamma}{\tau}} \frac{D(\lambda)}{\lambda},$$

where C_1 is a constant independent of λ, m or t.

Proof. We know from (2.4) that $\|f_\lambda\| \le \frac{D(\lambda)}{\lambda}$. So for any $\eta > 0$, f_λ can be written as $f_\lambda = \sum_{j=1}^{\infty} b_j K_{u_j}$ with $u_j \in X$ and

$$\|f_\lambda\| \le \sum_{j=1}^{\infty} |b_j| < \|f_\lambda\| + \eta \le \frac{D(\lambda)}{\lambda} + \eta. \tag{3.6}$$

At the same time, there exists some $N \in \mathbb{N}$ such that $\sum_{j=N}^{\infty} |b_j| < \eta$, and thus

$$\left\|\sum_{j=1}^{N} b_j K_{u_j} - f_\lambda\right\|_\infty \le \kappa \sum_{j=N}^{\infty} |b_j| \le \kappa\eta. \tag{3.7}$$

Proposition 3.1 ensures us $\{x_i\}_{i=1}^m$ is $c_{\alpha,\tau}\left(\frac{\log m+t}{m}\right)^{\frac{1}{\tau}}$-dense with confidence at least $1 - e^{-t}$, it implies that under the same confidence, for each u_j there is some $x(u_j) \in \{x_i\}_{i=1}^m$ such that $|x(u_j) - u_j|_2 \le c_{\alpha,\tau}\left(\frac{\log m+t}{m}\right)^{\frac{1}{\tau}}$. So by (3.2) and (3.6),

$$\begin{aligned}\left\|\sum_{j=1}^{N} b_j K_{u_j} - \sum_{j=1}^{N} b_j K_{x(u_j)}\right\|_\infty &\le c_\gamma c_{\alpha,\tau}^{\gamma}\left(\frac{\log m + t}{m}\right)^{\frac{\gamma}{\tau}} \sum_{j=1}^{N} |b_j| \\ &\le C_1\left(\frac{\log m + t}{m}\right)^{\frac{\gamma}{\tau}}\left(\frac{D(\lambda)}{\lambda} + \eta\right), \end{aligned} \tag{3.8}$$

where $C_1 := c_\gamma c_{\alpha,\tau}^\gamma$. Since $\sum_{j=1}^N b_j K_{x(u_j)} \in \mathcal{H}_{K,\mathbf{z}}$, we know from (1.3) and (2.1)

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) &\le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \\ &\le \mathcal{E}_{\mathbf{z}}\left(\sum_{j=1}^N b_j K_{x(u_j)}\right) + \lambda\sum_{j=1}^N |b_j| \\ &\le \mathcal{E}_{\mathbf{z}}\left(\sum_{j=1}^N b_j K_{x(u_j)}\right) + \lambda(\|f_\lambda\| + \eta).\end{aligned}$$

Note that for any $f_1, f_2 \in \mathcal{H}$,

$$|V(y, f_1(x)) - V(y, f_2(x))| \le \|f_1 - f_2\|_\infty.$$

This together with (3.7) and (3.8) implies

$$\begin{aligned}\left|\mathcal{E}_{\mathbf{z}}\left(\sum_{j=1}^N b_j K_{x(u_j)}\right) - \mathcal{E}_{\mathbf{z}}(f_\lambda)\right| &\le \left\|\sum_{j=1}^N b_j K_{x(u_j)} - f_\lambda\right\|_\infty \\ &\le \left\|\sum_{j=1}^N b_j K_{x(u_j)} - \sum_{j=1}^N b_j K_{u_j}\right\|_\infty \\ &\quad + \left\|\sum_{j=1}^N b_j K_{u_j} - f_\lambda\right\|_\infty \\ &\le C_1\left(\frac{\log m + t}{m}\right)^{\frac{\gamma}{\tau}}\left(\frac{D(\lambda)}{\lambda} + \eta\right) + \kappa\eta.\end{aligned}$$

Hence,

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) &\le \mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\| + \lambda\eta \\ &\quad + C_1\left(\frac{\log m + t}{m}\right)^{\frac{\gamma}{\tau}}\left(\frac{D(\lambda)}{\lambda} + \eta\right) + \kappa\eta.\end{aligned}$$

Let $\eta \to 0$, we get

$$\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \le \mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\| + C_1\left(\frac{\log m + t}{m}\right)^{\frac{\gamma}{\tau}}\frac{D(\lambda)}{\lambda}.$$

This proves the theorem. □

4. Sample error

For a measurable function $f : Z \to \mathbb{R}$, denote $\mathbb{E}f := \int_Z f(z)d\rho$. In order to estimate the sample error, one always assumes a variance–expectation bound for the pair (V, ρ) with the exponent $s \in [0, 1]$ and some $c_s > 0$

$$\mathbb{E}\{V(y, f(x)) - V(y, f^*(x))\}^2 \le c_s\{\mathcal{E}(f) - \mathcal{E}(f^*)\}^s, \quad \forall \|f\|_\infty \le M + \varepsilon. \tag{4.1}$$

It is easy to see that assumption (4.1) always holds for $s = 0$ and $c_s = 4(M + \varepsilon)^2$.

Write $S(\mathbf{z}, \lambda)$ as

$$\begin{aligned} S(\mathbf{z}, \lambda) = & \{[\mathcal{E}(\pi(f_{z,\boldsymbol{\lambda}})) - \mathcal{E}(f^*)] - [\mathcal{E}_{\mathbf{z}}(\pi(f_{z,\boldsymbol{\lambda}})) - \mathcal{E}_{\mathbf{z}}(f^*)]\} \\ & + \{[\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f^*)] - [\mathcal{E}(f_\lambda) - \mathcal{E}(f^*)]\} \\ =: & S_1(\mathbf{z}, \lambda) + S_2(\mathbf{z}, \lambda). \end{aligned}$$

To bound $S_2(\mathbf{z}, \lambda)$, we need the following one-sided Bernstein inequality (see [5]).

Let ξ be a random variable on a probability space Z with mean $\mathbb{E}\xi = \mu$ and variance $\sigma^2(\xi) = \sigma^2$. If $|\xi - \mu| \le B$ almost everywhere, then for all $\eta > 0$

$$Prob_{\mathbf{z}\in Z^m}\left\{\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu \ge \eta\right\} \le \exp\left\{-\frac{m\eta^2}{2\left(\sigma^2 + \frac{1}{3}B\eta\right)}\right\}.$$

Proposition 4.1. *If assumption (4.1) holds, then for any $t > 1$ with confidence at least $1 - 2e^{-t}$, one has*

$$S_2(\mathbf{z}, \lambda) \le \left(\frac{7\kappa D(\lambda)}{6m\lambda} + \frac{8(M+\varepsilon)}{3m} + \left(\frac{2c_s}{m}\right)^{\frac{1}{2-s}}\right) t + D(\lambda).$$

Proof. Denote $\xi_1 := V(y, f_\lambda(x)) - V(y, \pi(f_\lambda)(x))$, $\xi_2 := V(y, \pi(f_\lambda)(x)) - V(y, f^*(x))$, then $S_2(\mathbf{z}, \lambda) = \{\frac{1}{m}\sum_{i=1}^m \xi_1(z_i) - \mathbb{E}\xi_1\} + \{\frac{1}{m}\sum_{i=1}^m \xi_2(z_i) - \mathbb{E}\xi_2\}$. By (2.2) and (2.4), we can see that

$$\|f_\lambda\|_\infty \le \kappa\|f_\lambda\| \le \kappa\frac{D(\lambda)}{\lambda}.$$

So it is easy to check that $0 \le \xi_1 \le \kappa\frac{D(\lambda)}{\lambda}$ and $\sigma^2(\xi_1) \le \kappa\frac{D(\lambda)}{\lambda}\mathbb{E}\xi_1$.

Applying the one-sided Bernstein inequality to ξ_1 we have with confidence at least $1 - e^{-t}$,

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m}\xi_1(z_i) - \mathbb{E}\xi_1 &\le \frac{2\kappa t D(\lambda)}{3m\lambda} + \sqrt{\frac{2\kappa t D(\lambda)\mathbb{E}(\xi_1)}{m\lambda}} \\ &\le \frac{2\kappa t D(\lambda)}{3m\lambda} + \frac{\kappa t D(\lambda)}{2m\lambda} + \mathbb{E}\xi_1 \\ &= \frac{7\kappa t D(\lambda)}{6m\lambda} + \mathbb{E}\xi_1. \end{aligned}$$

For ξ_2, noting that both $\pi(f_\lambda)(x)$ and $f^*(x)$ are contained in $[-M-\varepsilon, M+\varepsilon]$, we know from assumption (4.1)

$$|\xi_2| \le |\pi(f_\lambda)(x) - f^*(x)| \le 2(M+\varepsilon), \qquad \sigma^2(\xi_2) \le c_s(\mathbb{E}(\xi_2))^s.$$

Applying the one-sided Bernstein inequality again, with confidence at least $1 - e^{-t}$, we have

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m}\xi_2(z_i) - \mathbb{E}\xi_2 &\le \frac{8(M+\varepsilon)t}{3m} + \sqrt{\frac{2tc_s(\mathbb{E}\xi_2)^s}{m}} \\ &\le \frac{8(M+\varepsilon)t}{3m} + \frac{s}{2}\mathbb{E}\xi_2 + \left(1 - \frac{s}{2}\right)\left(\frac{2c_s t}{m}\right)^{\frac{1}{2-s}} \\ &\le \frac{8(M+\varepsilon)t}{3m} + \left(\frac{2c_s t}{m}\right)^{\frac{1}{2-s}} + \mathbb{E}\xi_2, \end{aligned}$$

where in the second inequality we have used the elementary inequality

$$\frac{1}{p}+\frac{1}{q}=1, \quad \text{with } p,q>1 \Rightarrow ab \le \frac{1}{p}a^p+\frac{1}{q}b^q, \forall a,b>0. \tag{4.2}$$

Since $\mathbb{E}\xi_1+\mathbb{E}\xi_2=\mathcal{E}(f_\lambda)-\mathcal{E}(f^*)\le D(\lambda)$, the proposition is proved. □

It is more difficult to bound $S_1(\mathbf{z},\lambda)$, because $f_{\mathbf{z},\lambda}$ depends on samples $\mathbf{z}$ and thus runs over a set of functions in $\mathcal{H}$. We need a uniform probability inequality which involves the complexity of $\mathcal{H}$ described by covering number.

For $R>0$, denote $\mathcal{B}_R=\{f\in\mathcal{H}:\|f\|\le R\}$. By (2.2) $\mathcal{B}_R$ is a bound set in $C(X)$, we denote the uniform covering number of unit ball $\mathcal{B}_1$ as $\mathcal{N}(r):=\mathcal{N}(\mathcal{B}_1,r),\forall r>0$. The following lemma has been proved in [20].

Lemma 4.1. *If X satisfies (3.1) and K satisfies (3.2), then for any $0<r\le 1$,*

$$\log\mathcal{N}(r)\le c_\alpha\left(\frac{4c_\gamma}{r}\right)^{\frac{\alpha}{\gamma}}\log\left(2+\frac{4\kappa}{r}\right).$$

The next lemma is adopted from [18], it is a uniform law of large numbers for a class of function.

Lemma 4.2. *Let $0\le s\le 1$, $B>0$, $c\ge 0$, and $\mathcal{G}$ be a set of functions on Z such that for every $g\in\mathcal{G}$, $\mathbb{E}g\ge 0$, $|\mathbb{E}g-g|\le B$ and $\mathbb{E}g^2\le c(\mathbb{E}g)^s$. Then for any $\eta>0$,*

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}}\frac{\mathbb{E}g-\frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{(\mathbb{E}g)^s+\eta^s}}>4\eta^{1-\frac{s}{2}}\right\}\le\mathcal{N}(\mathcal{G},\eta)\exp\left\{\frac{-m\eta^{2-s}}{2\left(c+\frac{1}{3}B\eta^{1-s}\right)}\right\}.$$

Proposition 4.2. *Let $R>0$, if X satisfies (3.1), K satisfies (3.2) and (4.1) holds, then for any $t>1$, with confidence at least $1-e^{-t}$, there holds*

$$\begin{aligned}&\{\mathcal{E}(\pi(f))-\mathcal{E}(f^*)\}-\{\mathcal{E}_{\mathbf{z}}(\pi(f))-\mathcal{E}_{\mathbf{z}}(f^*)\}\\&\quad\le\frac{1}{2}\{\mathcal{E}(\pi(f))-\mathcal{E}(f^*)\}+C_2t\left\{\left(\frac{R^{\frac{\alpha}{\gamma}}}{m}\right)^{\frac{1}{2-s+\alpha/\gamma}}+\left(\frac{R^{\frac{\alpha}{\gamma}+1}}{m}\right)^{\frac{1}{3-s+\alpha/\gamma}}+\left(\frac{1}{m}\right)^{\frac{1}{2-s}}\right\},\end{aligned}$$

for all $f\in\mathcal{B}_R$, where C_2 is a constant independent of m, R or t.

Proof. Let $\mathcal{F}_R:=\{V(y,\pi(f)(x))-V(y,f^*(x)):f\in\mathcal{B}_R\}$, then for each function $g\in\mathcal{F}_R$

$$\|g\|_\infty\le 2(M+\varepsilon),\qquad |g-\mathbb{E}g|\le 4(M+\varepsilon),$$
$$\mathbb{E}g=\mathcal{E}(\pi(f))-\mathcal{E}(f^*)\ge 0,\qquad \frac{1}{m}\sum_{i=1}^m g(z_i)=\mathcal{E}_{\mathbf{z}}(\pi(f))-\mathcal{E}_{\mathbf{z}}(f^*).$$

Assumption (4.1) tells us $\mathbb{E}g^2\le c_s(\mathbb{E}g)^s$. So applying Lemma 4.2 to function set $\mathcal{F}_R$, we have

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{B}_R}\frac{\{\mathcal{E}(\pi(f))-\mathcal{E}(f^*)\}-\{\mathcal{E}_{\mathbf{z}}(\pi(f))-\mathcal{E}_{\mathbf{z}}(f^*)\}}{\sqrt{(\mathcal{E}(\pi(f))-\mathcal{E}(f^*))^s+\eta^s}}>4\eta^{1-\frac{s}{2}}\right\}$$

$$\leq \mathcal{N}(\mathcal{F}_R, \eta) \exp\left\{\frac{-m\eta^{2-s}}{2\left(c_s + \frac{4}{3}(M+\varepsilon)\eta^{1-s}\right)}\right\}.$$

Note that for any $f_1, f_2 \in \mathcal{B}_R$ and $(x, y) \in Z$,

$$|V(y, \pi(f_1)(x)) - V(y, \pi(f_2)(x))| \leq |\pi(f_1)(x) - \pi(f_2)(x)| \leq \|f_1 - f_2\|_\infty.$$

We get

$$\log \mathcal{N}(\mathcal{F}_R, \eta) \leq \log \mathcal{N}(\mathcal{B}_R, \eta) = \log \mathcal{N}\left(\frac{\eta}{R}\right) \leq c_\alpha \left(\frac{4c_\gamma R}{\eta}\right)^{\frac{\alpha}{\gamma}} \log\left(2 + \frac{4\kappa R}{\eta}\right).$$

Here the last inequality follows from Lemma 4.1. Choose $\tilde{\eta}$ to be the positive solution of the equation

$$h_2(\eta) := \frac{m\eta^{2-s}}{2\left(c_s + \frac{4}{3}(M+\varepsilon)\eta^{1-s}\right)} - c_\alpha \left(\frac{4c_\gamma R}{\eta}\right)^{\frac{\alpha}{\gamma}} \log\left(2 + \frac{4\kappa R}{\eta}\right) = t,$$

then for any $f \in \mathcal{B}_R$ with confidence at least $1 - e^{-t}$,

$$\begin{aligned}
&\{\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)\} - \{\mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f^*)\} \\
&\quad \leq 4\tilde{\eta}^{1-\frac{s}{2}} \sqrt{(\mathcal{E}(\pi(f)) - \mathcal{E}(f^*))^s + \tilde{\eta}^s} \\
&\quad \leq 4\tilde{\eta} + \frac{s}{2}\{\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)\} + \left(1 - \frac{s}{2}\right) 4^{\frac{2}{2-s}} \tilde{\eta} \\
&\quad \leq 20\tilde{\eta} + \frac{1}{2}\{\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)\},
\end{aligned}$$

where in the second inequality we have used the elementary inequality (4.2) again.

It remains to estimate $\tilde{\eta}$. Since

$$\{\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)\} - \{\mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f^*)\} \leq 4(M+\varepsilon),$$

we only need to consider the range $\eta \leq M + \varepsilon$. In this range,

$$h_2(\eta) \geq \frac{m\eta^{2-s}}{2\left(c_s + \frac{4}{3}(M+\varepsilon)^{2-s}\right)} - c_\alpha \left(\frac{4c_\gamma R}{\eta}\right)^{\frac{\alpha}{\gamma}} \left(\log 2 + \frac{4\kappa R}{\eta}\right) := h_3(\eta).$$

If we take η^* to be the unique positive solution to the equation $h_3(\eta) = t$, then $h_2(\tilde{\eta}) = t = h_3(\eta^*) \leq h_2(\eta^*)$. Because $h_2(\eta)$ is strictly increasing on $(0, +\infty)$, we know $\tilde{\eta} \leq \eta^*$. By a basic calculation, we can bound

$$\begin{aligned}
\eta^* \leq \max\Bigg\{ &\left[6\log 2 c_\alpha (4c_\gamma)^{\frac{\alpha}{\gamma}} \left(c_s + \frac{4}{3}(M+\varepsilon)^{2-s}\right)\left(\frac{R^{\frac{\alpha}{\gamma}}}{m}\right)\right]^{\frac{1}{2-s+\alpha/\gamma}}, \\
&\left[24\kappa c_\alpha (4c_\gamma)^{\frac{\alpha}{\gamma}} \left(c_s + \frac{4}{3}(M+\varepsilon)^{2-s}\right)\left(\frac{R^{\frac{\alpha}{\gamma}+1}}{m}\right)\right]^{\frac{1}{3-s+\alpha/\gamma}}, \\
&\left[6\left(c_s + \frac{4}{3}(M+\varepsilon)^{2-s}\right)\left(\frac{t}{m}\right)\right]^{\frac{1}{2-s}} \Bigg\}.
\end{aligned}$$

This proves the proposition by taking

$$C_2 := 20\left\{\left[6\log 2c_\alpha(4c_\gamma)^{\frac{\alpha}{\gamma}}\left(c_s+\frac{4}{3}(M+\varepsilon)^{2-s}\right)\right]^{\frac{1}{2-s+\alpha/\gamma}} + \left[24\kappa c_\alpha(4c_\gamma)^{\frac{\alpha}{\gamma}}\left(c_s+\frac{4}{3}(M+\varepsilon)^{2-s}\right)\right]^{\frac{1}{3-s+\alpha/\gamma}} + \left[6\left(c_s+\frac{4}{3}(M+\varepsilon)^{2-s}\right)\right]^{\frac{1}{2-s}}\right\}. \quad \square$$

Taking $f=0$ in (1.3), we can see that

$$\lambda\|f_{z,\boldsymbol{\lambda}}\| \le \left\{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})+\lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\right\} \le \mathcal{E}_{\mathbf{z}}(0) \le M.$$

So for any $\mathbf{z}\in Z^m$, $f_{z,\boldsymbol{\lambda}}\in\mathcal{B}_R$ with $R=\frac{M}{\lambda}$. This together with Proposition 4.2 gives the following.

Corollary 4.1. *If X satisfies (3.1), K satisfies (3.2) and (4.1) holds, then for any $t>1$, with confidence at least $1-e^{-t}$, there holds*

$$S_1(\mathbf{z},\lambda) \le \frac{1}{2}\{\mathcal{E}(\pi(f_{z,\boldsymbol{\lambda}}))-\mathcal{E}(f^*)\} + C_3 t\left\{\left(\frac{1}{m\lambda^{\frac{\alpha}{\gamma}}}\right)^{\frac{1}{2-s+\alpha/\gamma}} + \left(\frac{1}{m\lambda^{\frac{\alpha}{\gamma}+1}}\right)^{\frac{1}{3-s+\alpha/\gamma}} + \left(\frac{1}{m}\right)^{\frac{1}{2-s}}\right\},$$

where $C_3 := C_2\left(M^{\frac{\alpha}{(2-s)\gamma+\alpha}} + M^{\frac{\alpha+\gamma}{(3-s)\gamma+\alpha}} + 1\right)$.

5. Learning rates

Combining the estimation in Sections 3 and 4, we can finally derive an explicit learning rate for scheme (1.3).

Theorem 5.1. *Suppose X satisfies (3.1), K satisfies (3.2), ρ_X satisfies (3.3), and assumption (2.3) and (4.1) hold, by taking $\lambda=\left(\frac{1}{m}\right)^{\min\left\{\frac{\gamma}{\tau},\frac{\gamma}{\alpha+\gamma+(3-s)\gamma\beta+\alpha\beta}\right\}}$, we have for any $0<\delta<1$, with confidence at least $1-\delta$,*

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f^*) \le C\left(\log\frac{4}{\delta}+\log m\right)^{\max\left\{1,\frac{\gamma}{\tau}\right\}}\left(\frac{1}{m}\right)^{\min\left\{\frac{\gamma\beta}{\tau},\frac{\gamma\beta}{\alpha+\gamma+(3-s)\gamma\beta+\alpha\beta}\right\}},$$

where C is a constant independent of m or δ.

Proof. Putting Theorem 3.1, Proposition 4.1, Corollary 4.1 and assumption (2.3) into Proposition 2.1, we find for any $t>1$, with confidence at least $1-4e^{-t}$,

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f^*) \le \frac{1}{2}\{\mathcal{E}(\pi(f_{z,\boldsymbol{\lambda}}))-\mathcal{E}(f^*)\}$$

$$+ C_3 t \left\{ \left(\frac{1}{m\lambda^{\frac{\alpha}{\gamma}}} \right)^{\frac{1}{2-s+\alpha/\gamma}} + \left(\frac{1}{m\lambda^{\frac{\alpha}{\gamma}+1}} \right)^{\frac{1}{3-s+\alpha/\gamma}} + \left(\frac{1}{m} \right)^{\frac{1}{2-s}} \right\}$$
$$+ \left(\frac{7\kappa c_\beta}{6m\lambda^{1-\beta}} + \frac{8(M+\varepsilon)}{3m} + \left(\frac{2c_s}{m} \right)^{\frac{1}{2-s}} \right) t$$
$$+ C_1 c_\beta \left(\frac{\log m + t}{m} \right)^{\frac{\gamma}{\tau}} \lambda^{\beta-1} + 2c_\beta \lambda^\beta.$$

Therefore,

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*)$$
$$\leq C_4 (\log m + t)^{\max\{1, \frac{\gamma}{\tau}\}} \left\{ \left(\frac{1}{m\lambda^{\frac{\alpha}{\gamma}}} \right)^{\frac{1}{2-s+\alpha/\gamma}} + \left(\frac{1}{m\lambda^{\frac{\alpha}{\gamma}+1}} \right)^{\frac{1}{3-s+\alpha/\gamma}} + \left(\frac{1}{m} \right)^{\frac{1}{2-s}} \right\}$$
$$+ \left\{ \frac{1}{m\lambda^{1-\beta}} + \frac{1}{m} + \left(\frac{1}{m} \right)^{\frac{\gamma}{\tau}} \lambda^{\beta-1} + \lambda^\beta \right\},$$

where $C_4 := 2\left(C_3 + \frac{7\kappa c_\beta}{6} + \frac{8(M+\varepsilon)}{3} + (2c_s)^{\frac{1}{2-s}} + C_1 c_\beta + 2c_\beta \right\}$.

By the choice of λ, we can easily check that

$$\left(\frac{1}{m\lambda^{\frac{\alpha}{\gamma}}} \right)^{\frac{1}{2-s+\alpha/\gamma}} \leq \lambda^\beta, \qquad \left(\frac{1}{m\lambda^{\frac{\alpha}{\gamma}+1}} \right)^{\frac{1}{3-s+\alpha/\gamma}} \leq \lambda^\beta,$$
$$\frac{1}{m} \leq \left(\frac{1}{m} \right)^{\frac{1}{2-s}} \leq \lambda^\beta, \qquad \frac{1}{m\lambda^{1-\beta}} \leq \lambda^\beta, \qquad \left(\frac{1}{m} \right)^{\frac{\gamma}{\tau}} \lambda^{\beta-1} \leq \lambda^\beta.$$

So our theorem follows by taking $C = 7C_4$ and $t = \log \frac{4}{\delta}$. □

Acknowledgments

The authors thank the referees for their valuable comments and helpful suggestions.

References

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[2] E.J. Candès, T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory 51 (2005) 4203–4215.
[3] D.R. Chen, Q. Wu, Y.M. Ying, D.X. Zhou, Support vector machine soft margin classifiers: error analysis, J. Mach. Learn. Res. 5 (2004) 1143–1175.
[4] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000.
[5] F. Cucker, S. Smale, On the mathematical foundations of learing theory, Bull. Amer. Math. Soc. 39 (2001) 1–49.
[6] F. Cucker, D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2007.
[7] I. Daubechies, M. Defrise, C. Demol, An iterative thresholding algorithm for linear inverse problems with sparsity constraint, Comm. Pure Appl. Math. 57 (2004) 1413–1541.
[8] E. De Vito, A. Caponnetto, L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. 5 (2005) 59–85.
[9] D. Donoho, For most large underdetermined systems of linear equations, the minimal l^1-norm solution is also the sparsest solution, Technical report, Stanford University, 2004.
[10] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.

[11] M. Pontil, S. Mukherjee, F. Girosi, On the noise model of support vector machine regression, A. I. Memo 1651, MIT Artificial Intelligence Lab., 1998.
[12] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (2003) 17–41.
[13] H.W. Sun, Q. Wu, Least square regression with indefinite kernels and coefficient regularization, Appl. Comput. Harmonic Anal. 30 (2011) 96–109.
[14] B. Tarigan, S.A. Van de Geer, Classifiers of support vector machine type with l_1 complexity regularization, Bernoulli 12 (6) (2006) 1045–1076.
[15] H.Z. Tong, D.R. Chen, L.Z. Peng, Analysis of support vector machines regression, Found. Comput. Math. 9 (2009) 243–257.
[16] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[17] Q. Wu, Y. Ying, D.X. Zhou, Learning rates of least-square regularized regression, Found. Comput. Math. 6 (2006) 171–192.
[18] Q. Wu, D.X. Zhou, SVM soft margin classifiers: linear programming versus quadratic programming, Neural Comput. 17 (2005) 1160–1187.
[19] Q. Wu, D.X. Zhou, Learning with sample dependent hypothesis spaces, Comput. Math. Appl. 56 (2008) 2896–2907.
[20] Q.W. Xiao, D.X. Zhou, Learning by nonsymmetric kernels with data dependent spaces and l^1-regularizer, Taiwanese J. Math. 4 (2010) 1821–1836.
[21] J. Zhu, S. Rosset, T. Hastie, R. Tibshirani, 1-norm support vector machines, Adv. Neural Inf. Process. Syst. 16 (2004) 49–56.