Found Comput Math
DOI 10.1007/s10208-008-9026-0

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Analysis of Support Vector Machines Regression

Hongzhi Tong · Di-Rong Chen · Lizhong Peng

Received: 2 April 2007 / Revised: 28 December 2007 / Accepted: 29 January 2008

Abstract Support vector machines regression (SVMR) is a regularized learning algorithm in reproducing kernel Hilbert spaces with a loss function called the ε-insensitive loss function. Compared with the well-understood least square regression, the study of SVMR is not satisfactory, especially the quantitative estimates of the convergence of this algorithm. This paper provides an error analysis for SVMR, and introduces some recently developed methods for analysis of classification algorithms such as the projection operator and the iteration technique. The main result is an explicit learning rate for the SVMR algorithm under some assumptions.

Keywords Support vector machines regression · Regularization · Learning rates · Reproducing kernel Hilbert spaces · Excess error

Mathematics Subject Classification (2000) 68T05 · 62J02

Communicated by Felipe Cucker.

Research supported by NNSF of China No. 10471002, No. 10571010 and RFDP of China No. 20060001010.

H. Tong (✉) · L. Peng
LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China
e-mail: tonghz@math.pku.edu.cn

L. Peng
e-mail: lzpeng@pku.edu.cn

D.-R. Chen
Department of Mathematics, and LMIB, Beijing University of Aeronautics and Astronautics, Beijing 100083, People's Republic of China
e-mail: drchen@buaa.edu.cn

1 Introduction

Support vector machines regression (SVMR) [10, 19] has a foundation in the framework of statistical leaning theory and classical regularization theory for function approximation. The main difference between SVMR and the classical least square regression (LSR) [8, 10, 22] is that SVMR uses the ε-insensitive loss function (ILF) to measure the empirical error. Compared to quadratic loss function (see Fig. 1(a)) used in LSR, ILF is more robust and sparse. In addition, [13] shows that under the assumption the noise is additive and Gaussian where the variance and the mean of the Gaussian are random variables, use of the ILF is more justified. In this paper, we provide a mathematical analysis for SVMR. Our target is to determine a satisfactory learning rate for this algorithm.

Let us recall some basic concepts of statistical learning theory in the regression setting (see [6, 10, 15] and references therein for details).

From now on, we assume X is a compact subset of $\mathbb{R}^n$, Y is contained in $[-M, M]$ for some $M > 0$. The relation between the input $x \in X$ and the output $y \in Y$ is described by a probability distribution $\rho(x, y) = \rho(y|x)\rho_X(x)$ on $Z := X \times Y$, where $\rho(y|x)$ is the conditional probability of y given x and $\rho_X(x)$ is the marginal probability of x. The distribution 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 good estimate of y when a new input x is given.

The error for a measurable function f is the so-called expected risk

$$\mathcal{E}(f) := \int_Z V\big(y, f(x)\big)\, d\rho = \mathbb{E}V\big(y, f(x)\big),$$

where $V(y, f(x))$ is the loss function which measures the cost paid by replacing the true y with the estimate $f(x)$. We will denote by f^* the function which minimizes the error

$$f^* := \arg\min \mathcal{E}(f). \tag{1.1}$$

Obviously, f^* is our ideal estimator and it is often called the target function.

In this paper, we investigate the ε-insensitive loss function

$$V\big(y, f(x)\big) = \big|y - f(x)\big|_\varepsilon = \begin{cases} 0, & \text{if } |y - f(x)| < \varepsilon, \\ |y - f(x)| - \varepsilon, & \text{otherwise.} \end{cases} \tag{1.2}$$

ILF (see Fig. 1(b)) is similar to some of the functions used in robust statistics [11] (see Fig. 1(c)). It assigns zero cost to the errors smaller than ε.

SVMR depends on a reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel. Let $K : X \times X \to \mathbb{R}$ be 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 $(K(x_i, x_j))_{i,j=1}^l$ is positive semidefinite. Such a kernel is called a Mercer kernel.

The RKHS $\mathcal{H}_K$ associated with the 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

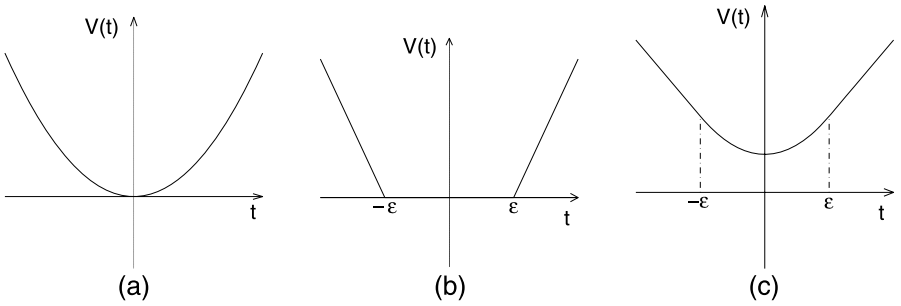

Fig. 1 (**a**) quadratic loss function $V_1(t) = t^2$; (**b**) ILF $V_1(t) = \begin{cases} |t|-\varepsilon, & |t|>\varepsilon \\ 0, & \text{otherwise;} \end{cases}$ (**c**) robust loss function $V_1(t) = \begin{cases} \frac{t^2}{2}+\frac{\varepsilon}{2} & |t|<\varepsilon, \\ |t| & \text{otherwise.} \end{cases}$ Here, $t := y - f(x)$ and $V_1(t) := V(y, f(x))$

$\langle \cdot, \cdot \rangle_K$ satisfying

$$\langle K_x, K_y \rangle_K = K(x, y).$$

The reproducing property is given by

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

Denote $C(X)$ as the space of continuous functions on X with the norm $\| \cdot \|_\infty$. Because of the continuity of K and compactness of X, we have

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

So, the above reproducing property tells us

$$\| f \|_\infty \le \kappa \| f \|_K, \quad \forall f \in \mathcal{H}_K. \tag{1.3}$$

Now the SVMR learning algorithm is defined to be the minimizer of the following Tikhonov regularization scheme in RKHS $\mathcal{H}_K$ with Mercer kernel K:

$$f_{\mathbf{z},\lambda} := \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} V\big(y_i, f(x_i)\big) + \lambda \| f \|_K^2 \right\}. \tag{1.4}$$

Where V is ILF defined by (1.2), λ is a positive constant called regularization parameter; it depends on m: $\lambda = \lambda(m)$, and usually $\lambda(m) \to 0$ as m becomes large.

A data free limit of (1.4) is

$$f_\lambda := \arg \min_{f \in \mathcal{H}_K} \big\{ \mathcal{E}(f) + \lambda \| f \|_K^2 \big\}. \tag{1.5}$$

If we set the empirical error with respect to the random samples $\mathbf{z}$ as

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} V\big(y_i, f(x_i)\big),$$

then the scheme (1.4) can be rewritten as

$$f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathcal{H}_K} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right\}. \tag{1.6}$$

Our main goal is to estimate the excess risk

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

for the scheme (1.6), where $\pi(\cdot)$ is a projection operator defined in Sect. 2.

There is a vast literature of error analysis for LSR, e.g., [6, 8, 16, 17, 22]. For general convex loss functions, a probabilistic bound on the sample error (see Sect. 3 for the definition) was considered in [14], but their results were founded on Ivanov regularization, that is, the hypothesis space is the ball of radius R in the RKHS $\mathcal{H}_K$. This setting is different from Tikhonov regularization (1.4) where the choice of regularization parameter $\lambda = \lambda(m)$ is essentially difficult even when f^* lies in $\mathcal{H}_K$. Recently, [5] established the consistency for a broad class of kernel-based regression methods. Although it did not give any explicit convergence rates, a further analysis implied the learning rates were slower than $m^{-\frac{1}{2}}$. In this paper, we provide an analysis tailored to SVMR by choosing some special values of the parameters, we can derive a learning rate arbitrarily close to m^{-1}. To our knowledge, there are no learning rates which can exceed m^{-1} in classification or LSR. So, it is reasonable to believe that our learning rate for SVMR is satisfactory. We have not investigated whether the results presented in this paper can be extended to the more general loss functions for regression. This problem will be considered in another work.

The rest of the paper is organized as follows: In Sect. 2, we discuss the target function with respect to ILF and introduce a projection operator. In Sect. 3, we give some definitions and assumptions which are necessary for the proof of the main results in Sect. 4.

2 Target Function and Projection Operator

It is well known (see [6, 12, 16, 22]) that the target function of quadratic loss is the regression function of ρ, which is defined by

$$f_\rho(x) = \int_Y y\, d\rho(y|x), \quad \forall x \in X.$$

It can be regarded as the average of the y coordinate $\{x\} \times Y$. But what is f which minimizes the expected ε-insensitive loss?

Theorem 2.1 *If f^* is the target function in (1.1) with respect to ILF, then it satisfies*

$$\int_{-M}^{f^*(x)-\varepsilon} d\rho(y|x) = \int_{f^*(x)+\varepsilon}^{M} d\rho(y|x), \quad \forall x \in X. \tag{2.1}$$

Proof Note that since

$$\mathcal{E}(f) = \mathbb{E}V\big(y, f(x)\big) = \mathbb{E}\big(\mathbb{E}V\big(y, f(x)\big)|x\big),$$

we can minimize $\mathcal{E}(f)$ by minimizing the conditional expectation $\mathbb{E}(V(y, f(x))|x)$.

For fixed x,

$$\begin{aligned}\mathbb{E}\big(V\big(y, f(x)\big)\big|x\big) &= \int_Y \big|y - f(x)\big|_\varepsilon \, d\rho(y|x) \\ &= \int_{-M}^{f(x)-\varepsilon} \big(f(x) - y - \varepsilon\big)\, d\rho(y|x) \\ &\quad + \int_{f(x)+\varepsilon}^{M} \big(y - f(x) - \varepsilon\big)\, d\rho(y|x).\end{aligned}$$

Here, we use the fact that y is supported on $[-M, M]$.

Let

$$F(\omega) = \int_{-M}^{\omega-\varepsilon} (\omega - y - \varepsilon)\, d\rho(y|x) + \int_{\omega+\varepsilon}^{M} (y - \omega - \varepsilon)\, d\rho(y|x),$$

by taking the derivative with respect to ω and setting it to 0, we obtain

$$\int_{-M}^{\omega-\varepsilon} d\rho(y|x) - \int_{\omega+\varepsilon}^{M} d\rho(y|x) = 0.$$

Hence, $f^*(x)$ the minimizer of $\mathbb{E}(\mathbb{E}V(y, f(x))|x)$ satisfies (2.1). □

Remark When $\varepsilon = 0$, the target function is the median function of ρ, which can be regarded as the median of y coordinate $\{x\} \times Y$.

Corollary 2.1 *If $f^*$is the target function with respect to ILF, then for any $x \in X$,*

$$\big|f^*(x)\big| \le M + \varepsilon. \tag{2.2}$$

Proof If the conclusion is not true, we have $f^*(x) > M + \varepsilon$ or $f^*(x) < -M - \varepsilon$ for some $x \in X$. In the first case,

$$\int_{-M}^{f^*(x)-\varepsilon} d\rho(y|x) \ge \int_{-M}^{M} d\rho(y|x) = 1.$$

But

$$\int_{f^*(x)+\varepsilon}^{M} d\rho(y|x) = 0.$$

This is a contradiction to (2.1). The other case is similar. □

Corollary 2.1 tells us $f^*(x) \in [-M - \varepsilon, M + \varepsilon]$, so it is natural to restrict the approximating functions to those also contained in $[-M - \varepsilon, M + \varepsilon]$. The idea of the projection operator was introduced in classification algorithm (see, e.g., [2, 4, 21]).

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

$$\pi(f)(x) = \begin{cases} M+\varepsilon, & \text{if } f(x) > M+\varepsilon, \\ -M-\varepsilon, & \text{if } f(x) < -M-\varepsilon, \\ f(x), & \text{if } -M-\varepsilon \le f(x) \le M+\varepsilon. \end{cases} \tag{2.3}$$

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

$$\mathcal{E}\big(\pi(f)\big) \le \mathcal{E}(f), \qquad \mathcal{E}_{\mathbf{z}}\big(\pi(f)\big) \le \mathcal{E}_{\mathbf{z}}(f). \tag{2.4}$$

Therefore, it is more accurate to estimate f^* by $\pi(f)$ than f. By virtue of it, we take $\pi(f_{\mathbf{z},\lambda})$ instead of $f_{\mathbf{z},\lambda}$ as our empirical target function and analyze the related learning rates.

3 Definitions and Assumptions

Estimating the excess error

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

for the SVMR scheme (1.6) is our goal. To this end, we present the following error decomposition which leads to bounds of the excess error.

Proposition 3.1 *Let f^*, $f_{\mathbf{z},\lambda}$, and f_λ be defined by (1.1), (1.4), and (1.5), respectively. Then $\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) \le \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2$ which can be bounded by*

$$\big\{\mathcal{E}(f_\lambda) - \mathcal{E}(f^*) + \lambda\|f_\lambda\|_K^2\big\} + \big\{\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big) - \mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z},\lambda})\big) + \mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\big\}. \tag{3.1}$$

Proof Write $\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2$ as

$$\begin{aligned} &\big\{\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big) - \mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z},\lambda})\big)\big\} + \big\{\big(\mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z},\lambda})\big) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2\big) - \big(\mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\|_K^2\big)\big\} \\ &\quad + \big\{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\big\} + \big\{\mathcal{E}(f_\lambda) - \mathcal{E}(f^*) + \lambda\|f_\lambda\|_K^2\big\}. \end{aligned}$$

By (2.4) and (1.6), the second term is at most zero. This proves (3.1). □

The first term of (3.1) is called the regularization error. It is independent of samples $\mathbf{z}$ and measures the approximation power of $\mathcal{H}_K$ for f^*.

Definition 3.1 The regularization error of scheme (1.6) is defined as

$$D(\lambda) := \inf_{f \in \mathcal{H}_K} \big\{\mathcal{E}(f) - \mathcal{E}(f^*) + \lambda\|f\|_K^2\big\}.$$

By (1.5), we know that $D(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f^*) + \lambda \|f_\lambda\|_K^2$. It is easy to see that $\mathcal{E}(f) - \mathcal{E}(f^*) \le \|f - f^*\|_{L^1_{\rho_X}}$. Hence, the regularization error concerns the approximation of f^* in $L^1_{\rho_X}$ by functions from $\mathcal{H}_K$; it can be characterized by requiring f^* to lie in some interpolation spaces of the pair $(L^1_{\rho_X}, \mathcal{H}_K)$, as done in [4] for classification.

Definition 3.2 We say the target function f^* can be approximated by $\mathcal{H}_K$ with exponent $0 < \beta \le 1$ if there exists a constant c_β, such that

$$\text{A1} \quad D(\lambda) \le c_\beta \lambda^\beta, \quad \forall \lambda > 0.$$

The second term of (3.1) is called the sample error. It has been well understood in learning theory by some concentration inequalities (see, e.g., [3, 6, 9, 20]). If some information about K and ρ is available, the sample error could be improved.

The information we need about K is the capacity measured by covering number.

Definition 3.3 Let $\mathcal{F}$ be a subset of a metric space. For any $\eta > 0$, the covering number $\mathcal{N}(\mathcal{F}, \eta)$ is defined to be the minimal integer $l \in \mathbb{N}$ such that there exist l balls with radius η covering $\mathcal{F}$.

In this paper, we use only the uniform covering number. Let $\mathcal{B}_R = \{f \in \mathcal{H}_K : \|f\|_K \le R\}$. It is a subset of $C(X)$, and the covering number is well defined. We denote the covering number of the unit ball $\mathcal{B}_1$ as

$$\mathcal{N}(\eta) := \mathcal{N}(\mathcal{B}_1, \eta), \quad \forall \eta > 0.$$

Definition 3.4 The RKHS $\mathcal{H}_K$ is said to have polynomial complexity exponent $s > 0$ if there is some constant $c_s > 0$, such that

$$\text{A2} \quad \log \mathcal{N}(\eta) \le c_s (1/\eta)^s, \quad \forall \eta > 0.$$

In learning theory, the uniform covering number has been extensively studied (see, e.g., [6, 23, 24]). It was shown in [24] that Assumption A2 holds if K is C^r with $r > 0$.

The information we need about ρ is the following variance-expectation bound condition (see [25]).

There exists some $\alpha \in [0, 1]$ and a constant c_α, such that

$$\text{A3} \quad \mathbb{E}\big\{V\big(y, f(x)\big) - V\big(y, f^*(x)\big)\big\}^2 \le c_\alpha \big\{\mathcal{E}(f) - \mathcal{E}(f^*)\big\}^\alpha, \quad \forall \|f\|_\infty \le M + \varepsilon.$$

It is easy to see that Assumption A3 always holds for $\alpha = 0$ and $c_\alpha = 4(M + \varepsilon)^2$.

4 Main Results

In this section, we discuss the estimation of the sample error in connection with Assumption A1, we derive the learning rates. The learning rates will be stated in

terms of the sample size m with proper choice of the regularization parameter $\lambda = \lambda(m) \to 0$.

Write the sample error in (3.1) as

$$\begin{aligned}&\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big) - \mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z},\lambda})\big) + \mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\\ &\quad = \big\{\big(\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big) - \mathcal{E}(f^*)\big) - \big(\mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z},\lambda})\big) - \mathcal{E}_{\mathbf{z}}(f^*)\big)\big\}\\ &\qquad + \big\{\big(\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f^*)\big) - \big(\mathcal{E}(f_\lambda) - \mathcal{E}(f^*)\big)\big\}. \end{aligned} \tag{4.1}$$

To bound the last term of (4.1), we need the following one-sided Bernstein inequality [6, 21].

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 $\tau > 0$

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

Proposition 4.1 *For any $t > 1$, under Assumption* A3, *with confidence $1 - 2e^{-t}$, we have*

$$\begin{aligned}&\big(\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f^*)\big) - \big(\mathcal{E}(f_\lambda) - \mathcal{E}(f^*)\big)\\ &\quad \le \frac{7\kappa t}{6m}\sqrt{\frac{D(\lambda)}{\lambda}} + \frac{8(M+\varepsilon)t}{3m} + \left(\frac{2c_\alpha t}{m}\right)^{\frac{1}{2-\alpha}} + D(\lambda). \end{aligned} \tag{4.2}$$

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

$$\begin{aligned}&\big(\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f^*)\big) - \big(\mathcal{E}(f_\lambda) - \mathcal{E}(f^*)\big)\\ &\quad = \left\{\frac{1}{m}\sum_{i=1}^m \xi_1(z_i) - \mathbb{E}\xi_1\right\} + \left\{\frac{1}{m}\sum_{i=1}^m \xi_2(z_i) - \mathbb{E}\xi_2\right\}. \end{aligned}$$

By Definition 3.1, we have

$$\lambda\|f_\lambda\|_K^2 \le \mathcal{E}(f_\lambda) - \mathcal{E}(f^*) + \lambda\|f_\lambda\|_K^2 = D(\lambda).$$

It follows from (1.3) that

$$\|f_\lambda\|_\infty \le \kappa\|f_\lambda\|_K \le \kappa\sqrt{\frac{D(\lambda)}{\lambda}}. \tag{4.3}$$

So, we can easily check that $0 \le \xi_1 \le \kappa\sqrt{\frac{D(\lambda)}{\lambda}}$ and $\sigma^2(\xi_1) \le \kappa\sqrt{\frac{D(\lambda)}{\lambda}}\mathbb{E}\xi_1$. By the one-sided Bernstein inequality, we see that with confidence at least $1 - e^{-t}$ that

$$\frac{1}{m}\sum_{i=1}^m \xi_1(z_i) - \mathbb{E}\xi_1 \le \frac{2\kappa t}{3m}\sqrt{\frac{D(\lambda)}{\lambda}} + \sqrt{\frac{2t\sigma^2(\xi_1)}{m}}$$

$$\leq \frac{2\kappa t}{3m}\sqrt{\frac{D(\lambda)}{\lambda}} + \frac{\kappa t}{2m}\sqrt{\frac{D(\lambda)}{\lambda}} + \mathbb{E}\xi_1$$
$$= \frac{7\kappa t}{6m}\sqrt{\frac{D(\lambda)}{\lambda}} + \mathbb{E}\xi_1.$$

For ξ_2, noting that both $\pi(f_\lambda)(x)$ and $f^*(x)$ are contained in $[-M-\varepsilon, M+\varepsilon]$, we know from Assumption A3

$$\sigma^2(\xi_2) \leq c_\alpha\big(\mathbb{E}(\xi_2)\big)^\alpha, \quad |\xi_2| \leq \big|\pi(f_\lambda)(x) - f^*(x)\big| \leq 2(M+\varepsilon).$$

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

$$\frac{1}{m}\sum_{i=1}^m \xi_2(z_i) - \mathbb{E}\xi_2 \leq \frac{8(M+\varepsilon)t}{3m} + \sqrt{\frac{2t\sigma^2(\xi_2)}{m}}$$
$$\leq \frac{8(M+\varepsilon)t}{3m} + \sqrt{\frac{2c_\alpha t(\mathbb{E}\xi_2)^\alpha}{m}}.$$

Recall an elementary inequality

$$\frac{1}{p}+\frac{1}{q}=1, \text{ with } p,q>1 \Rightarrow ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q, \ \forall a,b>0. \tag{4.4}$$

Using it with $a = (\mathbb{E}\xi_2)^{\frac{\alpha}{2}}$, $b = \big(\frac{2c_\alpha t}{m}\big)^{\frac{1}{2}}$, and $p = \frac{2}{\alpha}$, we have

$$\sqrt{\frac{2c_\alpha t(\mathbb{E}\xi_2)^\alpha}{m}} \leq \frac{\alpha}{2}\mathbb{E}\xi_2 + \left(1-\frac{\alpha}{2}\right)\left(\frac{2c_\alpha t}{m}\right)^{\frac{1}{2-\alpha}}.$$

Hence,

$$\frac{1}{m}\sum_{i=1}^m \xi_2(z_i) - \mathbb{E}\xi_2 \leq \frac{8(M+\varepsilon)t}{3m} + \left(\frac{2c_\alpha t}{m}\right)^{\frac{1}{2-\alpha}} + \mathbb{E}\xi_2.$$

Combining the above estimate for ξ_1 and ξ_2 with the fact $\mathbb{E}\xi_1 + \mathbb{E}\xi_2 = \mathcal{E}(f_\lambda) - \mathcal{E}(f^*) \leq D(\lambda)$, we prove the conclusion. □

The first term of (4.1) involves the samples $\mathbf{z}$, and thus runs over a set of functions. So, we need a probability inequality concerning the uniform convergence. The following lemma has been proved in [21].

Lemma 4.1 *Let $0 \leq \alpha \leq 1$, $B>0$, $c \geq 0$, and $\mathcal{G}$ be a set of functions on Z, such that for every $g \in \mathcal{G}$, $\mathbb{E}g \geq 0$, $|\mathbb{E}g - g| \leq B$ and $\mathbb{E}g^2 \leq c(\mathbb{E}g)^\alpha$. Then for any $\tau > 0$,*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}} \frac{\mathbb{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{(\mathbb{E}g)^\alpha + \tau^\alpha}} > 4\tau^{1-\frac{\alpha}{2}}\right\} \leq \mathcal{N}(\mathcal{G},\tau)\exp\left\{\frac{-m\tau^{2-\alpha}}{2(c+\frac{1}{3}B\tau^{1-\alpha})}\right\}.$$

We define the function set $\mathcal{F}_R$ with $R > 0$, by

$$\mathcal{F}_R := \big\{ V\big(y, \pi(f)(x)\big) - V\big(y, f^*(x)\big) : f \in \mathcal{B}_R \big\}.$$

Proposition 4.2 *Let $R > 0$, under Assumptions* A2 *and* A3, *for any $t > 1$, with confidence at least $1 - e^{-t}$, we have*

$$\big\{\mathcal{E}\big(\pi(f)\big) - \mathcal{E}(f^*)\big\} - \big\{\mathcal{E}_{\mathbf{z}}\big(\pi(f)\big) - \mathcal{E}_{\mathbf{z}}(f^*)\big\} \le 4\tau + 4\tau^{1-\frac{\alpha}{2}} \big\{\mathcal{E}\big(\pi(f)\big) - \mathcal{E}(f^*)\big\}^{\frac{\alpha}{2}}$$

for all $f \in \mathcal{B}_R$, where τ is given by

$$\tau := \left\{ 4\left(c_\alpha + \frac{4}{3}(M+\varepsilon)^{2-\alpha} \right)^{\frac{1}{2-\alpha}} + 1 \right\} \left(\left(\frac{t}{m} \right)^{\frac{1}{2-\alpha}} + \left(\frac{c_s R^s}{m} \right)^{\frac{1}{2-\alpha+s}} \right). \tag{4.5}$$

Proof Each function $g \in \mathcal{F}_R$ has the form $g(x, y) = V(y, \pi(f)(x)) - V(y, f^*(x))$ with some $f \in \mathcal{B}_R$. We can easily see that $\|g\|_\infty \le 2(M + \varepsilon)$, and thus $|g - \mathbb{E}g| \le B := 4(M + \varepsilon)$, $\mathbb{E}g = \mathcal{E}(\pi(f)) - \mathcal{E}(f^*) \ge 0$, $\frac{1}{m}\sum_{i=1}^m g(z_i) = \mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f^*)$. The Assumption A3 tells us $\mathbb{E}g^2 \le c(\mathbb{E}g)^\alpha$ for $c = c_\alpha$. So applying Lemma 4.1 to function set $\mathcal{F}_R$, we have

$$\begin{aligned} &\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{ \sup_{f\in\mathcal{B}_R} \frac{\{\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)\} - \{\mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f^*)\}}{\sqrt{(\mathcal{E}(\pi(f)) - \mathcal{E}(f^*))^\alpha + \tau^\alpha}} > 4\tau^{1-\frac{\alpha}{2}} \right\} \\ &\le \mathcal{N}(\mathcal{F}_R, \tau) \exp\left\{ \frac{-m\tau^{2-\alpha}}{2(c_\alpha + \frac{4}{3}(M+\varepsilon)\tau^{1-\alpha})} \right\}. \end{aligned}$$

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

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

We know that any τ-covering of $\mathcal{B}_R$ is also a τ-covering of $\mathcal{F}_R$, together with Assumption A2, we have

$$\log \mathcal{N}(\mathcal{F}_R, \tau) \le \log \mathcal{N}(\mathcal{B}_R, \tau) = \log \mathcal{N}\left(\frac{\tau}{R} \right) \le c_s \left(\frac{R}{\tau} \right)^s.$$

Therefore, if we set $\tilde{\tau}$ is the unique positive solution of the equation

$$\frac{m\tau^{2-\alpha}}{2(c_\alpha + \frac{4}{3}(M+\varepsilon)\tau^{1-\alpha})} - c_s \left(\frac{R}{\tau} \right)^s = t,$$

then with confidence at least $1 - e^{-t}$, we have

$$\begin{aligned} \big\{\mathcal{E}\big(\pi(f)\big) - \mathcal{E}(f^*)\big\} - \big\{\mathcal{E}_{\mathbf{z}}\big(\pi(f)\big) - \mathcal{E}_{\mathbf{z}}(f^*)\big\} &\le 4\tilde{\tau}^{1-\frac{\alpha}{2}} \sqrt{\big(\mathcal{E}\big(\pi(f)\big) - \mathcal{E}(f^*)\big)^\alpha + \tilde{\tau}^\alpha} \\ &\le 4\tilde{\tau} + 4\tilde{\tau}^{1-\frac{\alpha}{2}} \big\{\mathcal{E}\big(\pi(f)\big) - \mathcal{E}(f^*)\big\}^{\frac{\alpha}{2}}. \end{aligned}$$

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

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

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

$$\frac{m\tau^{2-\alpha}}{2(c_\alpha+\frac{4}{3}(M+\varepsilon)\tau^{1-\alpha})}-c_s\left(\frac{R}{\tau}\right)^s\ge\frac{m\tau^{2-\alpha}}{2(c_\alpha+\frac{4}{3}(M+\varepsilon)^{2-\alpha})}-c_s\left(\frac{R}{\tau}\right)^s=:h(\tau).$$

Because $h(\tau)$ is strictly increasing in $(0,+\infty)$, we know $\tilde{\tau}\le\tau^*$, where τ^* is the unique positive solution of the equation $h(\tau)=t$. By Lemma 7 from [7],

$$\tau^*\le\left(\frac{4(c_\alpha+\frac{4}{3}(M+\varepsilon)^{2-\alpha})t}{m}\right)^{\frac{1}{2-\alpha}}+\left(\frac{4(c_\alpha+\frac{4}{3}(M+\varepsilon)^{2-\alpha})c_sR^s}{m}\right)^{\frac{1}{2-\alpha+s}}.$$

This proves (4.5) and the proposition follows. □

Putting the above two estimates into Proposition 3.1, we can derive the error bounds. For $R>1$, denote

$$\mathcal{W}(R)=\big\{\mathbf{z}\in Z^m:\|f_{\mathbf{z},\lambda}\|_K\le R\big\}. \tag{4.6}$$

Proposition 4.3 *For all $t>1$, under Assumptions* A2 *and* A3, *there exists a set $V_R\subseteq Z^m$ with $\rho(V_R)\le 3e^{-t}$ such that, for all $\mathbf{z}\in\mathcal{W}(R)\backslash V_R$,*

$$\begin{aligned}&\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big)-\mathcal{E}(f^*)+\lambda\|f_{\mathbf{z},\lambda}\|_K^2\\&\le\frac{7\kappa t}{3m}\sqrt{\frac{D(\lambda)}{\lambda}}+\frac{16(M+\varepsilon)t}{3m}+2\left(\frac{2c_\alpha t}{m}\right)^{\frac{1}{2-\alpha}}+72\tau+4D(\lambda),\end{aligned}$$

where τ is given by (4.5).

Proof Proposition 3.1, 4.1, and 4.2 tell us that for any $t>1$, there exists a set $V_R\subseteq Z^m$ with measure at most $3e^{-t}$, such that for every $\mathbf{z}\in\mathcal{W}(R)\backslash V_R$,

$$\begin{aligned}&\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big)-\mathcal{E}(f^*)+\lambda\|f_{\mathbf{z},\lambda}\|_K^2\\&\le\frac{7\kappa t}{6m}\sqrt{\frac{D(\lambda)}{\lambda}}+\frac{8(M+\varepsilon)t}{3m}+\left(\frac{2c_\alpha t}{m}\right)^{\frac{1}{2-\alpha}}+4\tau\\&\quad+4\tau^{1-\frac{\alpha}{2}}\big\{\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big)-\mathcal{E}(f^*)\big\}^{\frac{\alpha}{2}}+2D(\lambda).\end{aligned}$$

Recall another elementary inequality

$$x\le ax^\nu+b,\ \ a,b,x>0\Rightarrow x\le\max\big\{(2a)^{\frac{1}{1-\nu}},2b\big\}.$$

Applying it with $x=\mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f^*)+\lambda\|f_{\mathbf{z},\lambda}\|_K^2$ and $\nu=\frac{\alpha}{2}$, we can derive the conclusion. □

We also need an R satisfying $\mathcal{W}(R) = Z^m$. By taking $f = 0$ in (1.4), we can see

Lemma 4.2 *For all $\lambda > 0, \mathbf{z} \in Z^m$, there holds*

$$\|f_{\mathbf{z},\lambda}\|_K \le \sqrt{\frac{M}{\lambda}}.$$

So, some learning rates in weak form can be obtained from Proposition 4.3, Lemma 4.2 and Assumption A1.

Corollary 4.1 *Under Assumptions* A1, A2, *and* A3, *for any $0 < \delta < 1$, by taking $\lambda = (\frac{1}{m})^{\min\{\frac{2}{1+\beta}, \frac{2}{2\beta(2-\alpha+s)+s}\}}$, with confidence $1-\delta$, we have*

$$\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big) - \mathcal{E}(f^*) \le \tilde{c_2} \log\left(\frac{3}{\delta}\right)\left(\frac{1}{m}\right)^{\min\{\frac{2\beta}{1+\beta}, \frac{2\beta}{2\beta(2-\alpha+s)+s}\}},$$

where $\tilde{c_2}$ is a constant independent of m and δ.

Proof Taking $R = \sqrt{\frac{M}{\lambda}}$, then Lemma 4.2 implies $f_{\mathbf{z},\lambda} \in \mathcal{B}_R$ for all $\mathbf{z} \in Z^m$. Hence, from Proposition 4.3, Assumption A1 and (4.5), we can see that for any $t > 1$ with confidence at least $1 - 3e^{-t}$

$$\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big) - \mathcal{E}(f^*) \le \tilde{c_1} t \left\{ \left(\frac{1}{m\lambda^{s/2}}\right)^{\frac{1}{2-\alpha+s}} + \frac{1}{m}\lambda^{\frac{\beta-1}{2}} + \lambda^{\beta} \right\},$$

where $\tilde{c_1}$ is a constant independent of m and t. By the choice of λ, we have

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

Therefore, the corollary follows by taking $\tilde{c_2} = 3\tilde{c_1}$ and $t = \log\frac{3}{\delta}$. □

The learning rate given in Corollary 4.1 is weak because we use the bound $\|f_{\mathbf{z},\lambda}\|_K \le \sqrt{\frac{M}{\lambda}}$ shown in Lemma 4.2. It is worse than the bound for f_λ derived from the proof of Proposition 4.1, namely, $\|f_\lambda\|_K \le \sqrt{\frac{D(\lambda)}{\lambda}}$. In order to improve the bound, we shall use a iteration technique which was introduced in [18] for a support vector machines classification algorithm. In particular, we can see $\|f_{\mathbf{z},\lambda}\|_K$ could be bounded essentially by $\sqrt{\frac{D(\lambda)}{\lambda}}$ with high confidence. Now, we present our main result.

Theorem 4.1 *Assume Assumptions* A1, A2, *and* A3. *Take $\lambda = (\frac{1}{m})^{\gamma}$, for any $\zeta > 0$ and $0 < \delta < 1$, there exists a constant $\tilde{c}$ independent of m such that with confidence $1-\delta$,*

$$\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big) - \mathcal{E}(f^*) \le \tilde{c}\left(\frac{1}{m}\right)^{\theta},$$

where

$$\gamma = \min\left\{\frac{2}{1+\beta}, \frac{2}{2\beta(2-\alpha+s)+(1-\beta)s}\right\},$$
$$\theta = \min\left\{\frac{2\beta}{1+\beta}, \frac{2\beta}{2\beta(2-\alpha+s)+(1-\beta)s} - \zeta\right\}.$$

Proof Denote $\Delta_{\mathbf{z}} = \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2$. We know from Proposition 4.3 and Assumption A1 that for any $t > 1$, there exists a set $V_R \subseteq Z^m$ with measure at most $3e^{-t}$, such that for every $\mathbf{z} \in \mathcal{W}(R)\backslash V_R$,

$$\Delta_{\mathbf{z}} \le \tilde{c}_3 t\left\{\left(\frac{R^s}{m}\right)^{\frac{1}{2-\alpha+s}} + \frac{1}{m}\lambda^{\frac{\beta-1}{2}} + \lambda^{\beta}\right\}, \tag{4.7}$$

where $\tilde{c}_3 > 1$ is a constant independent of m and t. Choosing $\lambda = (\frac{1}{m})^{\gamma}$, we can easily check that

$$\frac{1}{m}\lambda^{\frac{\beta-1}{2}} \le \lambda^{\beta}, \qquad \left(\frac{1}{m}\right)^{\frac{1}{2-\alpha+s}} \le \lambda^{\beta+\frac{(1-\beta)s}{2(2-\alpha+s)}}.$$

So (4.7) implies that

$$\Delta_{\mathbf{z}} \le \lambda^{\beta}\left\{\tilde{c}_3 t\left(\lambda^{\frac{1-\beta}{2}} R\right)^{\frac{s}{2-\alpha+s}} + 2\tilde{c}_3 t\right\}, \quad \forall \mathbf{z} \in \mathcal{W}(R)\backslash V_R. \tag{4.8}$$

Since $\|f_{\mathbf{z},\lambda}\|_K \le \sqrt{\frac{\Delta_{\mathbf{z}}}{\lambda}}$, by using (4.8) iteratively, we can find a small ball $\mathcal{B}_R$ that contains $f_{\mathbf{z},\lambda}$ with high confidence.

Start with $R = R^{(0)} = \sqrt{\frac{M}{\lambda}}$. Lemma 4.2 verifies $\mathcal{W}(R^{(0)}) = Z^m$. By (4.8), we know $Z^m = \mathcal{W}(R^{(0)}) \subseteq \mathcal{W}(R^{(1)}) \cup V_{R^{(0)}}$, where

$$R^{(1)} := \left\{\lambda^{\beta-1}\left(\tilde{c}_3 t\left(\lambda^{\frac{1-\beta}{2}}\sqrt{\frac{M}{\lambda}}\right)^{\frac{s}{2-\alpha+s}} + 2\tilde{c}_3 t\right)\right\}^{\frac{1}{2}}$$
$$\le \lambda^{\frac{\beta-1}{2}}\left(\tilde{c}_3(\sqrt{M}+1)t\lambda^{\frac{-\beta s}{4(2-\alpha+s)}} + 2\tilde{c}_3 t\right).$$

From (4.8), for $k = 2, 3, \ldots$, we iteratively derive

$$Z^m = \mathcal{W}\left(R^{(0)}\right) \subseteq \mathcal{W}\left(R^{(1)}\right) \cup V_{R^{(0)}} \subseteq \cdots \subseteq \mathcal{W}\left(R^{(k)}\right) \cup \left(\bigcup_{j=0}^{k-1} V_{R^{(j)}}\right),$$

where each $V_{R^{(j)}}$ has measure at most $3e^{-t}$ and $R^{(k)}$ is given by

$$R^{(k)} = \lambda^{\frac{\beta-1}{2}}\left\{\tilde{c}_3(\sqrt{M}+1)t\lambda^{-\frac{\beta}{2}(\frac{s}{2(2-\alpha+s)})^k} + 2k\tilde{c}_3 t\right\}. \tag{4.9}$$

For $\zeta > 0$, choose $k_0 \in \mathbb{N}$, such that

$$\left(\frac{s}{2(2-\alpha+s)}\right)^{k_0+1} \leq \frac{2\beta(2-\alpha+s)+(1-\beta)s}{2\beta}\zeta.$$

Take $k = k_0$ in (4.9), we know for $\mathbf{z} \in \mathcal{W}(R^{(k_0)})$

$$\|f_{\mathbf{z},\lambda}\|_K \leq \lambda^{\frac{\beta-1}{2}}\left\{\tilde{c}_3(\sqrt{M}+1)t\lambda^{-\frac{\beta}{2}(\frac{s}{2(2-\alpha+s)})^{k_0}} + 2k_0\tilde{c}_3 t\right\}.$$

This together with (4.7) gives

$$\mathcal{E}\big(\pi(f_{\mathbf{z},\lambda})\big) - \mathcal{E}(f^*) \leq \Delta_{\mathbf{z}} \leq \tilde{c}\left(\frac{1}{m}\right)^{\theta}, \quad \forall \mathbf{z} \in \mathcal{W}\big(R^{(k_0)}\big)\backslash V_{R^{(k_0)}}.$$

Since $\bigcup_{j=0}^{k_0} V_{R^{(j)}}$ has measure at most $3(k_0+1)e^{-t}$, taking $t = \log(\frac{3(k_0+1)}{\delta})$, the measure of $\mathcal{W}(R^{(k_0)})\backslash V_{R^{(k_0)}}$ is at least $1-\delta$. Then Theorem 4.1 is proved. □

Remark When $\beta = 1$ and $\alpha > s$, we have $\theta > \frac{1}{2}$ (up to a ζ). In particular, when $\beta = 1, \alpha = 1, s \to 0$, θ is arbitrarily close to 1.

Acknowledgements The first author would like to thank Professor Ding-Xuan Zhou for his encouragement and helpful suggestions. The authors thank the anonymous referees for their careful review and valuable comments and suggestions.

References

1. N. Aronszajn, Theory of reproducing kernels, *Trans. Am. Math. Soc.* **68**, 337–404 (1950).
2. P.L. Bartlett, The sample complexity of pattern classification with neural networks: The size of the weights is more import than the size of the network, *IEEE Trans. Inf. Theory* **44**, 525–536 (1998).
3. O. Bousquet, A. Elisseeff, Stability and generalization, *J. Mach. Learn. Res.* **2**, 499–526 (2002).
4. D.R. Chen, Q. Wu, Y. Ying, D.X. Zhou, Support vector machine soft margin classifiers: error analysis, *J. Mach. Learn. Res.* **5**, 1143–1175 (2004).
5. A. Christmann, I. Steinwart, Consistency and robustness of kernel-based regression in convex risk minimization, *Bernoulli* **13**, 799–819 (2007).
6. F. Cucker, S. Smale, On the mathematical foundations of learning theory, *Bull. Am. Math. Soc.* **39**, 1–49 (2001).
7. F. Cucker, S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, *Found. Comput. Math.* **2**, 413–428 (2002).
8. E. De Vito, A. Caponnetto, L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, *Found. Comput. Math.* **5**, 59–85 (2005).
9. L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition* (Springer, New York, 1997).
10. T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.* **13**, 1–50 (2000).
11. P.J. Huber, *Robust Statistics* (Wiley, New York, 1981).
12. T. Poggio, S. Smale, The mathematics of learning: Deal with data, *Not. Am. Math. Soc.* **50**, 537–544 (2003).
13. M. Pontil, S. Mukherjee, F. Girosi, On the noise model of support vector machine regression, A.I. Memo 1651, MIT Artificial Intelligence Lab., 1998.
14. L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, A. Verri, Are loss functions all the same? *Neural Comput.* **16**, 1063–1076 (2004).

15. B. Scholkopf, A.J. Smola, *Learning with Kernel* (MIT Press, Cambridge, 2002).
16. S. Smale, D.X. Zhou, Shannon sampling II. Connections to learning theory, *Appl. Comput. Harmon. Anal.* **19**, 285–302 (2005).
17. S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their applications, *Constr. Approx.* **26**, 153–172 (2007).
18. C. Scovel, I. Steinwart, Fast rates for support vector machine, in *Proceedings of the Conference on Learning Theory* (COLT-2005), pp. 279–294.
19. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995).
20. V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).
21. Q. Wu, D.X. Zhou, SVM soft margin classifiers: Linear programming versus quadratic programming, *Neural Comput.* **17**, 1160–1187 (2005).
22. Q. Wu, Y. Ying, D.X. Zhou, Learning rates of least-square regularized regression, *Found. Comput. Math.* **6**, 171–192 (2006).
23. D.X. Zhou, The covering number in learning theory, *J. Complex.* **18**, 739–767 (2002).
24. D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Inf. Theory* **49**, 1743–1752 (2003).
25. D.X. Zhou, K. Jetter, Approximation with polynomial kernels and SVM classifiers, *Adv. Comput. Math.* **25**, 323–344 (2006).