Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2014, Article ID 206015, 8 pages
http://dx.doi.org/10.1155/2014/206015

Research Article

A Simpler Approach to Coefficient Regularized Support Vector Machines Regression

Hongzhi Tong,[1] Di-Rong Chen,[2] and Fenghong Yang[3]

[1] *School of Statistics, University of International Business and Economics, Beijing 100029, China*
[2] *College of Mathematics & Computer Science, Wuhan Textile University, Wuhan 430200, China*
[3] *School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China*

Correspondence should be addressed to Hongzhi Tong; tonghz@uibe.edu.cn

Received 1 March 2014; Accepted 19 May 2014; Published 27 May 2014

Academic Editor: Gerd Teschke

We consider a kind of support vector machines regression (SVMR) algorithms associated with l^q $(1 \leq q < \infty)$ coefficient-based regularization and data-dependent hypothesis space. Compared with former literature, we provide here a simpler convergence analysis for those algorithms. The novelty of our analysis lies in the estimation of the hypothesis error, which is implemented by setting a stepping stone between the coefficient regularized SVMR and the classical SVMR. An explicit learning rate is then derived under very mild conditions.

1. Introduction

Recall the regression setting in learning theory, and let X be a compact subset of $\mathbb{R}^n$, $Y \subset [-M, M]$, for some $M > 0$. ρ is an unknown probability distribution endowed on $Z := X \times Y$, and $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ is a set of samples independently drawn according to ρ. Given samples $\mathbf{z}$, the regression problem aims to find 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) is a kind of kernel-based regression algorithms with the ε-insensitive loss defined by $V(y, t) = \max\{0, |y - t| - \varepsilon\}$ for some fixed $\varepsilon \geq 0$. A function $K : X \times X \to \mathbb{R}$ is called a Mercer kernel if it is continuous, symmetric, and positive semidefinite; that is, 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. The reproducing kernel Hilbert space (RKHS) $\mathscr{H}_K$ associated with a Mercer kernel K is defined (see [1]) to be the completion 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

$$\langle K_x, K_u \rangle_K = K(x, u), \tag{1}$$

and the reproducing property is given by

$$\langle K_x, f \rangle_K = f(x), \quad \forall x \in X, \ f \in \mathscr{H}_K. \tag{2}$$

Let

$$\kappa := \sup_{x \in X} \sqrt{K(x, x)} < \infty, \tag{3}$$

and then the reproducing property tells us the following:

$$\|f\|_\infty \leq \kappa \|f\|_K, \quad \forall f \in \mathscr{H}_K. \tag{4}$$

The classical SVMR proposed by Vapnik and his coworkers [2, 3] is given by the following regularization scheme:

$$\tilde{f}_{\mathbf{z},\mu} := \arg\min_{f \in \mathscr{H}_K} \left\{ \mathscr{E}_{\mathbf{z}}(f) + \mu \|f\|_K^2 \right\}, \tag{5}$$

where $\mathscr{E}_{\mathbf{z}}(f) := (1/m) \sum_{i=1}^m V(y_i, f(x_i))$ is the empirical error with respect to $\mathbf{z}$ and μ is a regularization parameter. It is well known, see for example, [4, Proposition 6.21], that the solution is of the form

$$\tilde{f}_{\mathbf{z},\mu} = \sum_{i=1}^m \tilde{\alpha}_i K(x_i, \cdot), \tag{6}$$

where the coefficients $\tilde{\alpha}_i$ are a solution of the optimization problem

$$\begin{aligned}&\text{maximize} \quad \sum_{i=1}^{m}\alpha_i y_i - \varepsilon\sum_{i=1}^{m}|\alpha_i| - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j K(x_i, x_j)\\ &\text{subject to} \quad |\alpha_i| \le \frac{1}{2\mu m} \quad \forall i = 1, 2, \ldots, m.\end{aligned} \tag{7}$$

Remark 1. The equality constraint $\sum_{i=1}^{m}\alpha_i = 0$ needed in [4, Proposition 6.21] is superfluous since we do not include an offset term b in the primal problem (5).

The mathematical analysis of algorithm (5) has been well understood with various techniques in extensive literature; see, for example, [5–7]. In this paper, we are interested in a different regularized SVMR algorithm. In our setting, the regularizer is not the RKHS norm but an l^q-norm of the coefficients in the kernel ensembles.

Definition 2. For $1 \le q < \infty$, let

$$\begin{aligned}\mathscr{H}_{K,\mathbf{z}} &:= \left\{\sum_{i=1}^{m}\alpha_i K_{x_i} : \alpha_i \in \mathbb{R},\ i = 1, 2, \ldots, m\right\},\\ \Omega_{\mathbf{z}}(f) &= \inf\left\{\sum_{i=1}^{m}|\alpha_i|^q : f = \sum_{i=1}^{m}\alpha_i K_{x_i}\right\}.\end{aligned} \tag{8}$$

Then, the SVMR with l^q-coefficient regularization learning algorithm that we study in this paper takes the form

$$f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathscr{H}_{K,\mathbf{z}}}\{\mathscr{E}_{\mathbf{z}}(f) + \lambda\Omega_{\mathbf{z}}(f)\}. \tag{9}$$

Remark 3. The regularization parameter λ in (9) may be different from μ in scheme (5), but a relationship between λ and μ will be given in Section 3 as we derive the learning rate of algorithm (9).

Learning with coefficient-based regularization has attracted a considerable amount of attention in recent years, on both theoretical analysis and applications. It was pointed out in [8] that by taking $q < 2$, and especially for the limit value $q = 1$, the proposed minimization procedure in (9) can promote the sparsity of the solution; that is, it tends to result in a solution with a few nonzero coefficients [9]. This phenomenon has been also observed in the LASSO algorithm [10] and the literature of compressed sensing [11].

However, it should be noticed that there are essential differences between the learning schemes (9) and (5). On one hand, the regularizer $\Omega_{\mathbf{z}}(f)$ is not a Hilbert space norm, which causes a technical difficulty for mathematical analysis. On the other hand, both hypothesis space $\mathscr{H}_{K,\mathbf{z}}$ and regularizer $\Omega_{\mathbf{z}}(f)$ depend on samples $\mathbf{z}$, and this increases the flexibility and adaptivity of algorithm (9) but causes the standard error analysis methods for scheme (5) which are not appropriate to scheme (9) any longer. To overcome these difficulties, [12] introduces a Banach space $\mathscr{H}$ of all functions of the form

$$f(x) = \sum_{j=1}^{\infty}\alpha_j K(u_j, x), \quad u_j \in X, \tag{10}$$

with norm

$$\|f\| = \inf\left\{\sum_{j=1}^{\infty}|\alpha_j| : f = \sum_{j=1}^{\infty}\alpha_j K_{u_j}\right\}. \tag{11}$$

An error analysis framework is established and then a series of papers start to investigate the performance of kernel learning scheme with coefficient regularization (see [13–16]). In those literatures, an L_τ condition imposed on the marginal distribution of ρ on X plays a critical role in the error analysis. A probability measure ρ_X on X is said to satisfy L_τ condition if there exist some $\tau > 0$ and $c_\tau > 0$ such that

$$\rho_X(\{u \in X : |u - x| < r\}) \ge c_\tau r^\tau, \quad \forall x \in X,\ 0 < r \le 1. \tag{12}$$

In general, the index τ is hard to estimate. If X satisfies some regularity conditions (such as an interior cone condition) and ρ_X is uniform distribution on X, then (12) holds with $\tau = n$. It leads to a low convergence rate and depends on n, the dimension of the input space X, which is often large in learning problem.

In this paper, we succeed to remove L_τ condition (12) and provide a simpler error analysis for scheme (9). The novelty of our analysis is a stepping stone technique applied to bound the hypothesis error. As a result, we derive an explicit learning rate of (9) under very mild conditions.

2. Error Decomposition and Hypothesis Error

The main purpose of this paper is to provide a convergence analysis of the learning scheme (9). With respect to the ε-insensitive loss V, the prediction ability of a measurable function f is measured by the following generalization error:

$$\begin{aligned}\mathscr{E}(f) &:= \int_Z V(y, f(x))\,d\rho\\ &= \int_X\int_Y V(y, f(x))\,d\rho(y \mid x)\,d\rho_X(x),\end{aligned} \tag{13}$$

where ρ_X is the marginal distribution on X and $\rho(\cdot \mid x)$ is the conditional probability measure at x induced by ρ. Let f^* be a minimizer of $\mathscr{E}(f)$ among all measurable functions on X. It was proved in [6] that $|f^*(x)| \le M+\varepsilon$ for almost every $x \in X$. To make full use of the feature of the target function f^*, one can introduce a projection operator, which was extensively used to the error analysis of learning algorithm; see, for example, [17, 18].

Definition 4. The projection operator $\pi = \pi_{M+\varepsilon}$ is defined on the space of measurable functions $f : X \to \mathbb{R}$ as

$$\pi(f)(x) = \begin{cases} M + \varepsilon, & \text{if } f(x) > M + \varepsilon,\\ -M - \varepsilon, & \text{if } f(x) < -M - \varepsilon,\\ f(x), & \text{if } -M - \varepsilon \le f(x) \le M + \varepsilon.\end{cases} \tag{14}$$

It is easy to see that $V(y, \pi(f)(x)) \leq V(y, f(x))$, so

$$\mathscr{E}(\pi(f)) \leq \mathscr{E}(f), \qquad \mathscr{E}_{\mathbf{z}}(\pi(f)) \leq \mathscr{E}_{\mathbf{z}}(f). \tag{15}$$

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

2.1. Error Decomposition.

The error decomposition is a useful approach to the error analysis for the regularized learning schemes. With sample-dependent hypothesis space $\mathscr{H}_{K,\mathbf{z}}$, [12] proposes a modified error decomposition with an extra hypothesis error term, by introducing a regularization function as

$$f_\mu := \arg\min_{f \in \mathscr{H}_K} \left\{ \mathscr{E}(f) + \mu \|f\|_K^2 \right\}, \quad \mu > 0. \tag{16}$$

We can conduct the error decomposition for scheme (9) with the same underlying idea of [12].

Proposition 5. *Let $f_{\mathbf{z},\lambda}$, f_μ be defined in* (9) *and* (16). *Then,*

$$\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}(f^*) \leq S(\mathbf{z}, \lambda, \mu) + P(\mathbf{z}, \lambda, \mu) + D(\mu). \tag{17}$$

Here,

$$\begin{aligned} S(\mathbf{z}, \lambda, \mu) &:= \{\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\} \\ &\quad + \left\{\mathscr{E}_{\mathbf{z}}(f_\mu) - \mathscr{E}(f_\mu)\right\}, \\ P(\mathbf{z}, \lambda, \mu) &:= \{\mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} \\ &\quad - \left\{\mathscr{E}_{\mathbf{z}}(f_\mu) + \mu \|f_\mu\|_K^2\right\}, \\ D(\mu) &:= \mathscr{E}(f_\mu) - \mathscr{E}(f^*) + \mu \|f_\mu\|_K^2 \\ &= \inf_{f \in \mathscr{H}_K} \left\{\mathscr{E}(f) - \mathscr{E}(f^*) + \mu \|f\|_K^2\right\}. \end{aligned} \tag{18}$$

Proof. A direct computation shows that

$$\begin{aligned} &\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}(f^*) \\ &\quad \leq \mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}(f^*) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \\ &\quad = \{\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\} \\ &\qquad + \{\mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} \\ &\qquad + \left\{\mathscr{E}_{\mathbf{z}}(f_\mu) - \mathscr{E}(f_\mu)\right\} - \left\{\mathscr{E}_{\mathbf{z}}(f_\mu) + \mu \|f_\mu\|_K^2\right\} \\ &\qquad + \left\{\mathscr{E}(f_\mu) - \mathscr{E}(f^*) + \mu \|f_\mu\|_K^2\right\} \\ &\quad = S(\mathbf{z}, \lambda, \mu) + P(\mathbf{z}, \lambda, \mu) + D(\mu). \end{aligned} \tag{19}$$

This proves the proposition. □

$S(\mathbf{z}, \lambda, \mu)$ is usually called the sample error; it will be estimated by some concentration inequality in the next section. $D(\mu)$ is independent of the sample and is often called the approximation error, the decay of $D(\mu)$, as $\mu \to 0$ characterizes the approximation ability of $\mathscr{H}_K$. We will assume that, for some $0 < \beta \leq 1$ and $c_\beta > 0$,

$$D(\mu) \leq c_\beta \mu^\beta, \quad \forall \mu > 0. \tag{20}$$

Remark 6. Since $\mathscr{E}(f) - \mathscr{E}(f^*) \leq \|f - f^*\|_{L^1_{\rho_X}}$, $D(\mu)$ concerns the approximation of f^* in $L^1_{\rho_X}$ by functions from $\mathscr{H}_K$. In fact, (20) can be satisfied when f^* is in some interpolation spaces of the pair $(L^1_{\rho_X}, \mathscr{H}_K)$ (see, e.g., [19, 20]).

$P(\mathbf{z}, \lambda, \mu)$ is called hypothesis error since the regularization function f_μ may not be in the hypothesis space $\mathscr{H}_{K,\mathbf{z}}$. The major contribution we make in this paper is to give a simpler estimation of $P(\mathbf{z}, \lambda, \mu)$ by a stepping stone between $f_{\mathbf{z},\lambda}$ and $\tilde{f}_{\mathbf{z},\mu}$.

2.2. Hypothesis Error Estimate.

The solution $f_{\mathbf{z},\lambda}$ of scheme (9) has a representation similar to $\tilde{f}_{\mathbf{z},\mu}$ in scheme (5); it is reasonable to expect close relations between the two schemes. So, the latter may play roles in the analysis of the former.

Theorem 7. *Let λ, $\mu > 0$, $1 \leq q < \infty$, and then*

$$P(\mathbf{z}, \lambda, \mu) \leq \frac{m\lambda}{(2m\mu)^q}. \tag{21}$$

Proof. Let $\tilde{f}_{\mathbf{z},\mu} = \sum_{i=1}^m \tilde{\alpha}_i K_{x_i}$ be the solution to (5). By (7), we have

$$\Omega_{\mathbf{z}}(\tilde{f}_{\mathbf{z},\mu}) \leq \sum_{i=1}^m |\tilde{\alpha}_i|^q \leq \frac{m}{(2m\mu)^q}. \tag{22}$$

Noting that $\tilde{f}_{\mathbf{z},\mu} \in \mathscr{H}_{K,\mathbf{z}}$, it can be derived from (15), (9), and (22) that

$$\begin{aligned} \mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) &\leq \mathscr{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \\ &\leq \mathscr{E}_{\mathbf{z}}(\tilde{f}_{\mathbf{z},\mu}) + \lambda \Omega_{\mathbf{z}}(\tilde{f}_{\mathbf{z},\mu}) \\ &\leq \mathscr{E}_{\mathbf{z}}(\tilde{f}_{\mathbf{z},\mu}) + \frac{m\lambda}{(2m\mu)^q} \\ &\leq \mathscr{E}_{\mathbf{z}}(\tilde{f}_{\mathbf{z},\mu}) + \mu \|\tilde{f}_{\mathbf{z},\mu}\|_K^2 + \frac{m\lambda}{(2m\mu)^q}. \end{aligned} \tag{23}$$

By taking $f = f_\mu$ in (5), one can get

$$\mathscr{E}_{\mathbf{z}}(\tilde{f}_{\mathbf{z},\mu}) + \mu \|\tilde{f}_{\mathbf{z},\mu}\|_K^2 \leq \mathscr{E}_{\mathbf{z}}(f_\mu) + \mu \|f_\mu\|_K^2. \tag{24}$$

Putting (24) into (23), one then has

$$\mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \leq \mathscr{E}_{\mathbf{z}}(f_\mu) + \mu \|f_\mu\|_K^2 + \frac{m\lambda}{(2m\mu)^q}. \tag{25}$$

This proves the theorem. □

Remark 8. The stepping stone method was first introduced in [21] to the error analysis of linear programming SVM classifiers. This technique was also used in [22, 23] to study l^q-coefficient regularized least square regression with $q \in [1, 2]$. While, in this paper, we extend the index q to a large range $[1, +\infty)$, it will be helpful to improve the understanding of those coefficient-based regularized algorithms.

Remark 9. Theorem 7 presents a simpler approach for estimating the hypothesis error. Different from the former literature (see, e.g., [13–16]), we conduct the estimation without imposing any assumptions on the input space X, kernel K, and marginal distribution ρ_X.

3. Sample Error and Learning Rate

This section is devoted to estimating the sample error $S(\mathbf{z}, \lambda, \mu)$ and deriving the learning rate of algorithm (9).

3.1. Sample Error Estimate.

3.1. Sample Error Estimate. We will adopt some results from the literature to estimate the sample error. To this end, we need some definitions and assumptions. For a measurable function $f : Z \to \mathbb{R}$, denote $\mathbb{E} f := \int_Z f(z) d\rho$.

Definition 10. A variance power s of the pair (V, ρ) is a number in $[0, 1]$ such that for any $f : X \to [-M - \varepsilon, M + \varepsilon]$, there exists some constant $c_s > 0$ satisfying

$$\mathbb{E}[V(y, f(x)) - V(y, f^*(x))]^2 \le c_s [\mathscr{E}(f) - \mathscr{E}(f^*)]^s. \tag{26}$$

Equation (26) is usually called a variance-expectation condition for the pair (V, ρ). It is easy to see that (26) always holds for $s = 0$ and $c_s = 4(M + \varepsilon)^2$. When $\varepsilon = 0$, the target function f^* becomes the median $f_{\rho,1/2}$ of ρ (see, [6]). In this case, as it points out in [24], if ρ has a median $f_{\rho,1/2}$ of a-average type b for some $a \in (0, \infty]$ and $b \in [1, \infty)$, then (26) can be satisfied with $s = \min\{2/b, a/(a + 1)\} \in (0, 1]$. Here, we say ρ has a median $f_{\rho,1/2}$ of a-average type b if, for every $x \in X$, there exist constants $c_x \in (0, 2]$ and $d_x > 0$ such that, for all $v \in [0, c_x]$,

$$\begin{aligned} \rho\left(\left\{y \in \left(f_{\rho,1/2}(x) - v, f_{\rho,1/2}(x)\right) \mid x\right\}\right) &\ge d_x v^{b-1}, \\ \rho\left(\left\{y \in \left(f_{\rho,1/2}(x), f_{\rho,1/2}(x) + v\right) \mid x\right\}\right) &\ge d_x v^{b-1}, \end{aligned} \tag{27}$$

and that the function on X taking value $(d_x c_x^{b-1})^{-1}$ at $x \in X$ lies in $L^a_{\rho_X}$. But for $\varepsilon > 0$, as we know, it is still open to find a meaningful condition for ρ to guarantee that (26) holds with a positive index $s > 0$.

Definition 11. Let $\mathscr{F}$ be a class of functions on Z and let $\mathbf{z} = \{z_i\}_{i=1}^m \in Z^m$. The l^2-metric $d_{2,\mathbf{z}}$ is defined on $\mathscr{F}$ by

$$d_{2,\mathbf{z}}(f, g) := \left\{\frac{1}{m}\sum_{i=1}^m \left|f(z_i) - g(z_i)\right|^2\right\}^{1/2}, \quad \forall f, g \in \mathscr{F}. \tag{28}$$

For every $\eta > 0$, the covering number of $\mathscr{F}$ with respect to $d_{2,\mathbf{z}}$ is

$$\mathscr{N}_{2,\mathbf{z}}(\mathscr{F}, \eta) := \inf\left\{l \in \mathbb{N} : \exists \{f_i\}_{i=1}^l \text{ such that } \mathscr{F} = \bigcup_{i=1}^l \{f \in \mathscr{F} : d_{2,\mathbf{z}}(f, f_i) \le \eta\}\right\}. \tag{29}$$

Let $\mathscr{B}_R := \{f \in \mathscr{H}_K : \|f\|_K \le R\}$. The l^2-empirical covering number of the unit ball $\mathscr{B}_1$ is defined as

$$\mathscr{N}(\mathscr{B}_1, \eta) := \sup_{m \in \mathbb{N}} \sup_{\mathbf{x} \in X^m} \mathscr{N}_{2,\mathbf{x}}(\mathscr{B}_1, \eta). \tag{30}$$

We assume that $\mathscr{H}_K$ satisfies the following capacity assumption.

There exists an exponent p with $0 < p < 2$ and a constant $c_p > 0$ such that

$$\log \mathscr{N}(\mathscr{B}_1, \eta) \le c_p \eta^{-p}, \quad \forall \eta > 0. \tag{31}$$

We now set out to bound the sample error. Write $S(\mathbf{z}, \lambda, \mu)$ as

$$\begin{aligned} S(\mathbf{z}, \lambda, \mu) = {} & \left\{\left[\mathscr{E}_{\mathbf{z}}(f_\mu) - \mathscr{E}_{\mathbf{z}}(f^*)\right] - \left[\mathscr{E}(f_\mu) - \mathscr{E}(f^*)\right]\right\} \\ & + \left\{\left[\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}(f^*)\right] \right. \\ & \left. - \left[\mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}_{\mathbf{z}}(f^*)\right]\right\} \\ =: {} & S_1(\mathbf{z}, \mu) + S_2(\mathbf{z}, \lambda). \end{aligned} \tag{32}$$

Applying [6, Proposition 4.1], we yield the following estimation for $S_1(\mathbf{z}, \mu)$.

Lemma 12. *For any $t > 0$, under the assumption (26), with confidence $1 - 2e^{-t}$, one has*

$$\begin{aligned} S_1(\mathbf{z}, \mu) \le {} & \frac{7\kappa t}{6m}\sqrt{\frac{D(\mu)}{\mu}} + \frac{8(M + \varepsilon)t}{3m} \\ & + \left(\frac{2c_s t}{m}\right)^{1/(2-s)} + D(\mu). \end{aligned} \tag{33}$$

The estimation for $S_2(\mathbf{z}, \lambda)$ is based on the following concentration inequality which can be found in [25].

Lemma 13. *Let $\mathscr{F}$ be a set of measurable functions on Z, and let $B, c > 0$, and $s \in [0, 1]$ be constants such that each $f \in \mathscr{F}$ satisfies $\|f\|_\infty \le B$ and $\mathbb{E}(f^2) \le c(\mathbb{E} f)^s$. If, for some $A > 0$ and $p \in (0, 2)$,*

$$\sup_{m \in \mathbb{N}} \sup_{\mathbf{z} \in Z^m} \log \mathscr{N}_{2,\mathbf{z}}(\mathscr{F}, \eta) \le A\eta^{-p}, \quad \forall \eta > 0, \tag{34}$$

then there exists a constant c_p' depending only on p such that for any $t > 0$, with probability $1 - e^{-t}$, there holds

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \leq \frac{1}{2}\theta^{1-s}(\mathbb{E}f)^s + c_p'\theta + 2\left(\frac{ct}{m}\right)^{1/(2-s)} + \frac{18Bt}{m}, \quad \forall f \in \mathcal{F}, \tag{35}$$

where $\theta = \max\{c^{(2-p)/(4-2s+ps)}(A/m)^{2/(4-2s+ps)}, B^{(2-p)/(2+p)}(A/m)^{2/(2+p)}\}$.

We may apply Lemma 13 to a set of functions $\mathcal{F}_R$ with $R > 0$, where

$$\mathcal{F}_R := \{V(y, \pi(f)(x)) - V(y, f^*(x)) : f \in \mathcal{B}_R\}. \tag{36}$$

Proposition 14. *If assumptions* (26) *and* (31) *are satisfied, then for any $t > 0$, with confidence $1 - e^{-t}$, there holds*

$$\begin{aligned}&[\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)] - [\mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f^*)] \\ &\quad\leq \frac{1}{2}[\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)] + \left(\frac{1}{2} + c_p'\right)\theta_R \\ &\qquad + 2\left(\frac{c_s t}{m}\right)^{1/(2-s)} + \frac{36(M+\varepsilon)t}{m}\end{aligned} \tag{37}$$

for all $f \in \mathcal{B}_R$, where

$$\theta_R = \max\left\{c_s^{(2-p)/(4-2s+ps)}\left(\frac{c_p R^p}{m}\right)^{2/(4-2s+ps)}, [2(M+\varepsilon)]^{(2-p)/(2+p)}\left(\frac{c_p R^p}{m}\right)^{2/(2+p)}\right\}. \tag{38}$$

Proof. Each function $g \in \mathcal{F}_R$ has a form

$$g(z) = V(y, \pi(f)(x)) - V(y, f^*(x)) \tag{39}$$

with some $f \in \mathcal{B}_R$. We can easily see that $\|g\|_\infty \leq 2(M+\varepsilon)$ and

$$\begin{aligned}\mathbb{E}g &= \mathcal{E}(\pi(f)) - \mathcal{E}(f^*), \\ \frac{1}{m}\sum_{i=1}^{m} g(z_i) &= \mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f^*).\end{aligned} \tag{40}$$

The assumption (26) tells us that $\mathbb{E}(g^2) \leq c(\mathbb{E}g)^s$ with $c = c_s$. Moreover, for any $f_1, f_2 \in \mathcal{B}_R$, and $(x, y) \in Z$,

$$\begin{aligned}&|V(y, \pi(f_1)(x)) - V(y, \pi(f_2)(x))| \\ &\quad\leq |\pi(f_1)(x) - \pi(f_2)(x)| \leq |f_1(x) - f_2(x)|,\end{aligned} \tag{41}$$

we get

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{F}_R, \eta) \leq \mathcal{N}_{2,\mathbf{x}}(\mathcal{B}_R, \eta) = \mathcal{N}_{2,\mathbf{x}}\left(\mathcal{B}_1, \frac{\eta}{R}\right). \tag{42}$$

This together with (31) implies

$$\sup_{m\in\mathbb{N}}\sup_{\mathbf{z}\in Z^m} \log\mathcal{N}_{2,\mathbf{z}}(\mathcal{F}_R, \eta) \leq \log\mathcal{N}\left(\mathcal{B}_1, \frac{\eta}{R}\right) \leq c_p R^p \eta^{-p}. \tag{43}$$

Hence, all the conditions in Lemma 13 hold, and we know that, for any $t > 0$, with confidence $1 - e^{-t}$, there holds, for every $f \in \mathcal{B}_R$,

$$\begin{aligned}&[\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)] - [\mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f^*)] \\ &\quad\leq \frac{1}{2}\theta_R^{1-s}[\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)]^s + c_p'\theta_R \\ &\qquad + 2\left(\frac{c_s t}{m}\right)^{1/(2-s)} + \frac{36(M+\varepsilon)t}{m}.\end{aligned} \tag{44}$$

Here,

$$\theta_R = \max\left\{c_s^{(2-p)/(4-2s+ps)}\left(\frac{c_p R^p}{m}\right)^{2/(4-2s+ps)}, [2(M+\varepsilon)]^{(2-p)/(2+p)}\left(\frac{c_p R^p}{m}\right)^{2/(2+p)}\right\}. \tag{45}$$

Recall an elementary inequality

$$\begin{aligned}&\frac{1}{\iota} + \frac{1}{\nu} = 1, \quad \text{with } \iota, \nu > 1 \\ &\Longrightarrow \psi\omega \leq \frac{1}{\iota}\psi^\iota + \frac{1}{\nu}\omega^\nu, \quad \forall\psi, \omega > 0.\end{aligned} \tag{46}$$

Applying it with $\psi = [\mathcal{E}(\pi(f)) - \mathcal{E}(f^*)]^s$, $\omega = \theta_R^{1-s}$, $\iota = 1/s$ to the first term of (44), we can derive the conclusion. □

It remains to find a ball containing $f_{\mathbf{z},\lambda}$ for all $\mathbf{z} \in Z^m$.

Lemma 15. *Let $1 \leq q < \infty$, $f_{\mathbf{z},\lambda}$ be defined by* (9). *Then, for any $\mathbf{z} \in Z^m$, one has*

$$\|f_{\mathbf{z},\lambda}\|_K \leq \kappa m^{1-(1/q)}\left(\frac{M}{\lambda}\right)^{1/q}. \tag{47}$$

Proof. For any $\zeta > 0$, there exists $\{\alpha_i\}_{i=1}^m \in \mathbb{R}^m$, such that $f_{\mathbf{z},\lambda} = \sum_{i=1}^m \alpha_i K_{x_i}$ and

$$\sum_{i=1}^{m}|\alpha_i|^q \leq \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \zeta. \tag{48}$$

Taking $f = 0$ in (9), we can see that

$$\lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \leq \mathcal{E}_{\mathbf{z}}(0) \leq M. \tag{49}$$

It follows that

$$\sum_{i=1}^{m}|\alpha_i|^q \leq \frac{M}{\lambda} + \zeta. \tag{50}$$

When $q > 1$, by the Hölder inequality, we can see that

$$\begin{aligned}\|f_{\mathbf{z},\lambda}\|_K &= \left\|\sum_{i=1}^{m}\alpha_i K_{x_i}\right\|_K \\ &\leq \kappa\sum_{i=1}^{m}|\alpha_i| \\ &\leq \kappa m^{1-(1/q)}\left\{\sum_{i=1}^{m}|\alpha_i|^q\right\}^{1/q} \\ &\leq \kappa m^{1-(1/q)}\left\{\frac{M}{\lambda}+\zeta\right\}^{1/q}.\end{aligned} \tag{51}$$

It is easy to check that (51) still holds for $q = 1$. Let $\zeta \to 0$; we then get the assertion. □

From Lemma 15 and Proposition 14, we can get the following.

Corollary 16. *If assumptions* (26) *and* (31) *hold, then, for any* $t > 1$, *with confidence* $1 - e^{-t}$, *there holds*

$$\begin{aligned}S_2(\mathbf{z},\lambda) \leq{}& \frac{1}{2}\left[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*)\right] \\ &+ C_1 t\left\{\left(m^{p-(p/q)-1}\lambda^{-p/q}\right)^{2/(4-2s+ps)}\right. \\ &\left.+\left(m^{p-(p/q)-1}\lambda^{-p/q}\right)^{2/(2+p)} + m^{1/(s-2)}\right\},\end{aligned} \tag{52}$$

where

$$\begin{aligned}C_1 :={}& \left(\frac{1}{2}+c_p'\right)\left\{c_s^{(2-p)/(4-2s+ps)}\left(c_p\kappa^p M^{p/q}\right)^{2/(4-2s+ps)}\right. \\ &\left.+\left[2(M+\varepsilon)\right]^{(2-p)/(2+p)}\left(c_p\kappa^p M^{p/q}\right)^{2/(2+p)}\right\} \\ &+ 2c_s^{1/(2-s)} + 36(M+\varepsilon).\end{aligned} \tag{53}$$

3.2. Deriving Learning Rates. Combining the estimation in Sections 2.2 and 3.1, we can derive an explicit learning rate for scheme (9) by suitably selecting the regularization parameters λ and μ.

Theorem 17. *Suppose that assumptions* (20), (26), *and* (31) *are satisfied, for any* $0 < \delta < 1$, *by taking* $\lambda = m^{q-1}\mu^{\beta+q}$, $\mu = (1/m)^{\min\{2/(1+\beta),2q/((4-2s+ps)\beta q+2p(\beta+q))\}}$, *and we have, with confidence* $1-\delta$,

$$\begin{aligned}&\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) \\ &\quad\leq C\log\frac{3}{\delta}\left(\frac{1}{m}\right)^{\min\{2\beta/(1+\beta),2q\beta/((4-2s+ps)\beta q+2p(\beta+q))\}},\end{aligned} \tag{54}$$

where C is a constant independent of m or δ.

Proof. Putting Theorem 7, Lemma 12, Corollary 16, and assumption (20) into Proposition 5, by taking $\lambda = m^{q-1}\mu^{\beta+q}$, we find that, for any $t > 1$, with confidence $1 - 3e^{-t}$,

$$\begin{aligned}&\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) \\ &\quad\leq 2\left\{C_1 t\left[\left(m^{-1}\mu^{-(p\beta/q)-p}\right)^{2/(4-2s+ps)}\right.\right. \\ &\qquad\left.+\left(m^{-1}\mu^{-(p\beta/q)-p}\right)^{2/(2+p)} + m^{1/(s-2)}\right] \\ &\qquad+\frac{7\kappa t\sqrt{c_\beta}}{6m}\mu^{(\beta-1)/2} + \frac{8(M+\varepsilon)t}{3m} \\ &\qquad\left.+\left(\frac{2c_s t}{m}\right)^{1/(2-s)} + 2c_\beta\mu^\beta + \frac{\mu^\beta}{2^q}\right\} \\ &\quad\leq C_2 t\left\{\left(\frac{1}{m^q\mu^{p(\beta+q)}}\right)^{2/(4-2s+ps)q}\right. \\ &\qquad+\left(\frac{1}{m^q\mu^{p(\beta+q)}}\right)^{2/(2+p)q} \\ &\qquad\left.+\left(\frac{1}{m}\right)^{1/(2-s)} + \frac{\mu^{(\beta-1)/2}}{m} + \mu^\beta\right\}.\end{aligned} \tag{55}$$

Here, $C_2 := 2[C_1 + (7/6)\kappa\sqrt{c_\beta} + 8(M+\varepsilon)/3 + (2c_s)^{1/(2-s)} + 2c_\beta + 2^{-q}]$. According to the choice of μ, we can easily check that

$$\begin{gathered}\left(\frac{1}{m}\right)^{1/(2-s)} \leq \mu^\beta, \qquad \frac{\mu^{(\beta-1)/2}}{m} \leq \mu^\beta, \\ \left(\frac{1}{m^q\mu^{p(\beta+q)}}\right)^{2/(2+p)q} \leq \left(\frac{1}{m^q\mu^{p(\beta+q)}}\right)^{2/(4-2s+ps)q} \leq \mu^\beta.\end{gathered} \tag{56}$$

So, our theorem follows by taking $C = 5C_2$ and $t = \log(3/\delta)$. □

Remark 18. Theorem 17 provides an explicit learning rate for l^q $(1 \leq q < \infty)$ coefficient-based regularized SVMR. This learning rate is independent of the dimension n of the input space X. We do not require the marginal distribution ρ_X and the kernel K to satisfy any additional regularity condition, such as the L_τ condition.

Remark 19. Another advantage of coefficient-based regularization scheme is its flexibility in choosing the kernel. For instance, [26, 27] consider the least square regression with indefinite kernels and an l^2-coefficient regularization, where they relax the requirement of the kernel to be only continuous and uniformly bounded bivariate function on X. It will be a very interesting topic in future work to extend the method in this paper to the indefinite kernel setting.

Let us end this paper by comparing our result with the learning rate presented in [7] in a special case $\varepsilon = 0$. To this end, we reformulate [7, Theorem 2.3] for $\varepsilon = 0$ as follows.

Proposition 20. *If $K \in C^\infty(X \times X)$, $f_{\rho,1/2} \in \mathscr{H}_K$, and $f_{\rho,1/2}$ is of a-average of type 2 for some $a \in (0,\infty]$, taking $\mu = m^{-(a+1)/(a+2)}$, $0 < \varsigma < (a+1)/2(a+2)$, then, with $a^* = 2a/(a+1)$, for any $0 < \delta < 1$, with confidence $1-\delta$, one has*

$$\left\|\pi(\widetilde{f}_{\mathbf{z},\mu}) - f_{\rho,1/2}\right\|_{L^{a^*}_{\rho_X}} \le \widetilde{C}_1 \log\frac{3}{\delta} m^{\varsigma-(a+1)/2(a+2)}, \qquad (57)$$

where $\widetilde{C}_1$ is a constant independent of m or δ.

By Theorem 17, we can see the following.

Corollary 21. *Under the same conditions of Proposition 20, for $\varepsilon = 0$, by taking $\lambda = m^{(q-2a-3)/(a+2)}$, one has, for any $0 < \delta < 1$, with confidence $1-\delta$,*

$$\left\|\pi(f_{\mathbf{z},\lambda}) - f_{\rho,1/2}\right\|_{L^{a^*}_{\rho_X}} \le \widetilde{C}_2 \sqrt{\log\frac{3}{\delta}} m^{-(a+1)/2(a+2)}, \qquad (58)$$

where $\widetilde{C}_2$ is a constant independent of m or δ.

Proof. Note that $f^* = f_{\rho,1/2}$ when $\varepsilon = 0$. From [24, Theorem 2.7], we know that

$$\left\|\pi(f_{\mathbf{z},\lambda}) - f_{\rho,1/2}\right\|_{L^{a^*}_{\rho_X}} \le c_\rho \left(\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}(f_{\rho,1/2})\right)^{1/2}, \quad (59)$$

and here c_ρ is a constant independent of m or δ.

Since $f_{\rho,1/2} \in \mathscr{H}_K$, we know that (20) holds with $\beta = 1$. $f_{\rho,1/2}$ is a-average type 2 which implies that (26) is satisfied with $s = a/(a+1)$. Since $X \subset \mathbb{R}^n$ and $K \in C^\infty(X \times X)$, we know from [28] that (31) holds true for any $p > 0$. Therefore, let $p \to 0$; according to Theorem 17, by taking $\lambda = m^{q-1}\mu^{1+q}$, $\mu = m^{-(a+1)/(a+2)}$, we have, for any $0 < \delta < 1$, with confidence $1-\delta$,

$$\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right) - \mathscr{E}\left(f_{\rho,1/2}\right) \le C \log\frac{3}{\delta}\left(\frac{1}{m}\right)^{(a+1)/(a+2)}. \qquad (60)$$

This together with (59) proves the corollary with the constant $\widetilde{C}_2 = c_\rho\sqrt{C}$. □

Corollary 21 shows us that the learning rate presented in Theorem 17 for l^q-coefficient regularized SVMR is faster than the one given in [7] for RKHS norm regularized learning schemes at least in the case of $\varepsilon = 0$.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank the referees for their valuable comments and helpful suggestions which greatly improved this work. The research is supported by NSF of China under Grants 11171014.

References

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

[2] V. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in *Advances in Neural Information Proceeding Systems*, M. Mozer, M. Jordan, and T. Petsche, Eds., vol. 9, pp. 81–287, MIT Press, Cambridge, Mass, USA, 1997.

[3] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[4] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

[5] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, New York, NY, USA, 2008.

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

[7] D.-H. Xiang, T. Hu, and D.-X. Zhou, "Approximation analysis of learning algorithms for support vector regression and quantile regression," *Journal of Applied Mathematics*, vol. 2012, Article ID 902139, 17 pages, 2012.

[8] I. Daubechies, M. Defrise, and C. Demol, "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.

[9] D. Donoho, "For most large underdetermined systerms of linear equations, the minimal l^1-norm solution is also the sparsest solution," Tech. Rep., Stanford University, 2004.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[11] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[12] Q. Wu and D.-X. Zhou, "Learning with sample dependent hypothesis spaces," *Computers and Mathematics with Applications*, vol. 56, no. 11, pp. 2896–2907, 2008.

[13] 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.

[14] L. Shi, Y.-L. Feng, and D.-X. Zhou, "Concentration estimates for learning with l^1-regularizer and data dependent hypothesis spaces," *Applied and Computational Harmonic Analysis*, vol. 31, no. 2, pp. 286–302, 2011.

[15] 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.

[16] H.-Y. Wang, Q.-W. Xiao, and D.-X. Zhou, "An approximation theory approach to learning with l^1 regularization," *Journal of Approximation Theory*, vol. 167, pp. 240–258, 2013.

[17] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 525–536, 1998.

[18] I. Steinwart, D. Hush, and C. Scovel, "An oracle inequality for clipped regularized risk minimizers," in *Advances in Neural Information Proceeding Systems*, B. Scholköpf, J. Platt, and T. Hoffman, Eds., vol. 19, pp. 1321–1328, MIT Press, Cambridge, Mass, USA, 2007.

[19] S. Smale and D. X. Zhou, “Estimating the approximation error in learning theory,” *Analysis and Applications*, vol. 1, no. 1, pp. 17–41, 2003.

[20] D.-X. Zhou, “Density problem and approximation error in learning theory,” *Abstract and Applied Analysis*, vol. 2013, Article ID 715683, 13 pages, 2013.

[21] Q. Wu and D.-X. Zhou, “SVM soft margin classifiers: linear programming versus quadratic programming,” *Neural Computation*, vol. 17, no. 5, pp. 1160–1187, 2005.

[22] H. Z. Tong, D.-R. Chen, and F. H. Yang, “Least square regression with l^p-coefficient regularization,” *Neural Computation*, vol. 22, no. 12, pp. 3221–3235, 2010.

[23] Y.-L. Feng and S.-G. Lv, “Unified approach to coefficient-based regularized regression,” *Computers and Mathematics with Applications*, vol. 62, no. 1, pp. 506–515, 2011.

[24] I. Steinwart and A. Christmann, “Estimating conditional quantiles with the help of the pinball loss,” *Bernoulli*, vol. 17, no. 1, pp. 211–225, 2011.

[25] Q. Wu, Y. Ying, and D.-X. Zhou, “Multi-kernel regularized classifiers,” *Journal of Complexity*, vol. 23, no. 1, pp. 108–134, 2007.

[26] H. W. Sun and Q. Wu, “Least square regression with indefinite kernels and coefficient regularization,” *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 96–109, 2011.

[27] H. W. Sun and Q. Wu, “Indefinite kernel network with dependent sampling,” *Analysis and Applications*, vol. 11, no. 5, Article ID 1350020, 15 pages, 2013.

[28] D.-X. Zhou, “Capacity of reproducing kernel spaces in learning theory,” *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1743–1752, 2003.