Multiscale support vector regression method in Sobolev spaces on bounded domains

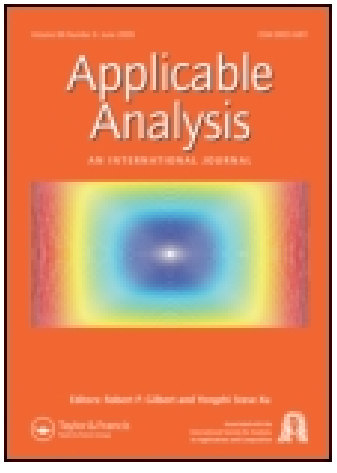
Article *in* Applicable Analysis · December 2014
DOI: 10.1080/00036811.2014.918261

3 authors:

Boxi Xu
Fudan University
4 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Shuai Lu
Fudan University
39 PUBLICATIONS **376** CITATIONS
SEE PROFILE

Min Zhong
Southeast University (China)
9 PUBLICATIONS **40** CITATIONS
SEE PROFILE



Applicable Analysis: An International Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gapa20

Multiscale support vector regression method in Sobolev spaces on bounded domains

Boxi Xu[a], Shuai Lu[a] & Min Zhong[a]

[a] School of Mathematical Sciences, Fudan University, 200433 Shanghai, P.R. China.

Published online: 19 May 2014.

To cite this article: Boxi Xu, Shuai Lu & Min Zhong (2015) Multiscale support vector regression method in Sobolev spaces on bounded domains, Applicable Analysis: An International Journal, 94:3, 548-569, DOI: 10.1080/00036811.2014.918261

To link to this article: http://dx.doi.org/10.1080/00036811.2014.918261

Applicable Analysis, 2015
Vol. 94, No. 3, 548–569, http://dx.doi.org/10.1080/00036811.2014.918261

Multiscale support vector regression method in Sobolev spaces on bounded domains

Boxi Xu, Shuai Lu and Min Zhong*

School of Mathematical Sciences, Fudan University, 200433 Shanghai, P.R. China

Communicated by J. Cheng

(*Received 26 February 2014; accepted 18 April 2014*)

In this paper, we investigate the multiscale support vector regression (SVR) method for approximation of functions in Sobolev spaces on bounded domains. The Vapnik ϵ-intensive loss function, which has been developed well in learning theory, is introduced to replace the standard l^2 loss function in multiscale least squares methods. Convergence analysis is presented to verify the validity of the multiscale SVR method with scaled versions of compactly supported radial basis functions. Error estimates on noisy observation data are also derived to show the robustness of our proposed algorithm. Numerical simulations support the theoretical predictions.

Keywords: multiscale analysis; support vector approach; radial basis functions

AMS Subject Classifications: 65J10; 46E35

1. Introduction

The radial basis function (RBF) has formed an efficient mesh-less approximating tool and been applied to varies of problems, for instance, in multivariate scattered data interpolation, spline smoothing, numerical methods on partial differential equations, and inverse problems (see [1–10] and references therein).

In modern approximation theory, RBF is firstly implemented to approximate an unknown function defined on arbitrary dimensions and with arbitrary smoothness, see, for instance, [10]. However, one has to consider the trade-off between approximating accuracy and computational cost in practice.[11] The later introduced compactly supported radial basis functions (CSRBFs) in [12–14], which reduce the computational cost by building sparse collocation matrices but the approximate solution becomes radius dependent. More precisely, by choosing a small support radius, one can build a well-conditioned sparse collocation system thus with low computational cost but, nevertheless, the approximation is not accurate. On the other hand, a large support radius yields good approximate solutions at the expense of extremely ill-conditioned collocation matrix.

To balance the approximation accuracy and the computational cost with CSRBFs, a natural idea is varying the support radius [12,15,16] and consequently multiple scales are

*Corresponding author. Email: 09110180007@fudan.edu.cn
Dedicated to Professor Bernd Hofmann for his 60th birthday.

introduced which yield multilevel or multiscale type schemes.[17] Among the latter choices, Wendland [17] developed a multiscale algorithm based on the least squares methods by implementing CSRBFs for multivariate interpolation. For the first time, rigorous proofs demonstrate that the multiscale approach obtains an excellent trade-off between superior convergence rate and low computational costs. Later on, such an idea has been successfully utilized for approximating functions on a sphere [18,19]; solving elliptic PDEs on a sphere [20]; solving second-order elliptic PDEs by Galerkin method on bounded domains, and Stokes problems [21,22] with zero boundary conditions.[23] Its performance on linear ill-posed problems is also investigated in [24,25].

In this manuscript, based on the multiscale least squares method in [17], we extend the multiscale analysis to the support vector regression (SVR) method. The SVR method, which is also a variant of the support vector machine, has been developed in learning theory by Boser et al. in [26] and further investigated by Vapnik in [27–29]. The SVR method usually accomplishes with a cut-off parameter ϵ and the so-called Vapnik $\epsilon-$intensive loss function defined by

$$|x|_\epsilon = \begin{cases} 0, & \text{for } |x| \le \epsilon; \\ |x| - \epsilon, & \text{for } |x| > \epsilon. \end{cases}$$

By replacing the standard l^2 loss function by the $\epsilon-$intensive loss function in the data-fitting term of least squares methods, we intend to overcome over-fitting of noisy data which destroys the generalization characteristic of the target functions. An overview of SVR methods and the classification of the Vapnik $\epsilon-$intensive loss function can be found in [30–33] and references therein. Inspired by the recent results in [34] concerning *a priori* parameter choice strategies for SVR method on single-level data-set, our manuscript extends both the SVR method and its error estimates to multiscale analysis, which provides a better performance in practice.

The manuscript is organized as follows. Section 2 contains an overview of the preliminary knowledge and presents the multiscale SVR approximation algorithm. Convergence analyses for the approximation algorithm are carried out in Sections 3 and 4 for smooth and rough functions, respectively. Section 5 devotes to the error estimate for noisy observation. Finally, some numerical simulations in Section 6 show the efficiency of our proposed multiscale algorithm.

2. Preliminary and Algorithm

2.1. *RBFs and RKHS*

A RBF Φ is a scalar function acting $R^d \to R$. By defining a bounded domain Ω and choosing a sampling data-set $X = \{x_j \in \Omega \subset R^d, 1 \le j \le N\}$ with respect to two measures (the fill distance h_X and the separation distance q_X)

$$h_X := \sup_{x\in\Omega} \min_{x_j\in X} \|x - x_j\|,$$
$$q_X := \min_{j\neq k} \|x_j - x_k\|,$$

we can define a closed and finite dimensional approximation space

$$W_X := \text{span}\{\Phi(\cdot - x_j), 1 \le j \le N\}.$$

Table 1. [10, Table 9.1] SPD$\left(R^d\right)$ Wendland kernel functions with minimal degrees.

Dimension	Index ℓ	Kernel functions $p_{d,\ell}$	Smoothness $C^{2\ell}$
$d = 1$	0	$(1-r)_+$	C^0
	1	$(1-r)_+^3(3r+1)$	C^2
	2	$(1-r)_+^5(24r^2+15r+3)$	C^4
$d = 2, 3$	0	$(1-r)_+^2$	C^0
	1	$(1-r)_+^4(4r+1)$	C^2
	2	$(1-r)_+^6(35r^2+18r+3)$	C^4
	3	$(1-r)_+^8(32r^3+25r^2+8r+1)$	C^6

The discrete observation data-set $g|_X := \{g_j\}_{j=1}^N$ is associated with the sampling data-set X such that $g_j = f^*(x_j)$ given a target function f^*.

Typical choices of RBFs include thin-plate splines, Sobolev splines,[35] multiquadrics,[36] Matern functions, and Wendland functions.[13,37] Numerical algorithms can be easily realized on these kernel functions, especially when Φ is compactly supported (for instance, Matern and Wendland kernels). In current work, we concentrate on the Wendland kernels which are compactly supported strictly positive definite radial basis functions (CS-SPD-RBFs).[13,14] These kernels are of the form

$$\Phi_{d,\ell} = \begin{cases} p_{d,\ell}(r), & 0 \le r \le 1; \\ 0, & 1 < r, \end{cases}$$

with a univariate polynomial $p_{d,\ell}$ of the minimal degree $\lfloor \frac{d}{2} \rfloor + 3\ell + 1$. In Table 1, we list some kernel functions with respect to the dimensionality and smoothness.

Taking the Fourier transform on Φ

$$\widehat{\Phi}(\omega) = (2\pi)^{-d/2} \int_{R^d} \Phi(x) e^{-ix^T\omega} dx,$$

we observe that these Fourier coefficients satisfy

$$c_1 \left(1 + \|\omega\|^2\right)^{-\tau} \le \widehat{\Phi}(\omega) \le c_2 \left(1 + \|\omega\|^2\right)^{-\tau} \tag{1}$$

with two constants $c_2 \ge c_1 > 0$ and $\tau = d/2 + \ell + 1/2$. Referring to the following lemma in [10], there is a intrinsic relation between reproducing kernel Hilbert spaces (RKHS) and the standard Sobolev spaces $H^\tau\left(R^d\right)$, provided that $\tau > \frac{d}{2}$.

LEMMA 2.1 [10,13] *Suppose that the Fourier transform of an integrable function $\Phi : R^d \to R$ satisfies (1) with an arbitrary $\tau > \frac{d}{2}$. Then, the kernel function Φ is a reproducing kernel of the Hilbert space (native space) $\mathcal{N}_\Phi\left(R^d\right)$, with the inner-product*

$$(f, g)_\Phi = \int_{R^d} \frac{\widehat{f}(\omega)\overline{\widehat{g}(\omega)}}{\widehat{\Phi}(\omega)} d\omega.$$

Furthermore, norms in $\mathcal{N}_\Phi\left(R^d\right)$ *and* $H^\tau\left(R^d\right)$ *are equivalent, i.e.*

$$c_1^{1/2}\|f\|_\Phi \le \|f\|_{H^\tau(R^d)} \le c_2^{1/2}\|f\|_\Phi. \tag{2}$$

Here and subsequently, we denote $\|\cdot\|_\Phi$ *as* $\|\cdot\|_{\mathcal{N}_\Phi(R^d)}$ *for simplicity.*

In view of the preceding lemma, the RBFs $\Phi_{d,\ell}$ referring to Table 1 are reproducing kernels of Hilbert spaces which are isomorphic to the standard Sobolev spaces $H^{d/2+\ell+1/2}\left(R^d\right)$.

2.2. *Function approximation by SVR method*

In the sequel, we will confine our interests on the cases such that $\Phi := \Phi_{d,\ell}$ is a CS-SPD-RBF. The goal to approximate or learn a function is to approximate its value from some known discrete (maybe noisy) training observation data on one hand, and predict it on the unknown data-set on the other hand.[32,38,39] Concerning the regression problem with the Vapnik ϵ-intensive loss function, we introduce the following cost functional

$$\frac{1}{N}\sum_{j=1}^{N}|f(x_j)-g_j|_\epsilon + \gamma\|f\|_\Phi^2,$$

where there exists a minimizer (referring to [40]) with a regularization parameter $\gamma > 0$. This minimizer can be represented by a linear combination of basic functions $\Phi(\cdot - x_j)$, i.e.

$$f = \sum_{j=1}^{N}\alpha_j\Phi(\cdot - x_j).$$

To numerically calculate the approximate solution, we introduce slack variables $a_j, b_j \ge 0$ for $j = 1, 2, \ldots, N$. Thus, the minimization problem is equivalent to the following constrained quadratic optimization problem, see, for instance, [41],

$$\min_{\alpha,a,b\in R^N, a,b\ge 0}\left(\frac{1}{N}\sum_{j=1}^{N}\left(a_j+b_j\right)+\gamma\alpha^T\Phi_X\alpha\right) \tag{3}$$

$$\begin{aligned}&\text{s.t. } (\Phi_X)_j\,\alpha - g_j \le \epsilon + a_j, \quad j = 1,\cdots,N,\\ &-\left[(\Phi_X)_j\,\alpha - g_j\right] \le \epsilon + b_j, \quad j = 1,\cdots,N,\end{aligned}$$

where $\Phi_X \in R^{N\times N}$ is the *symmetric collocation matrix* with entries

$$(\Phi_X)_{ji} = \Phi(x_i - x_j), \qquad 1 \le i, j \le N.$$

Denoting $1_N = (1,\ldots,1)^T \in R^N$ and $z = (\alpha, a, b)^T \in R^{3N}$, the quadratic optimization problem (3) can be written in the matrix form

$$\begin{aligned}&\min_{z\in R^{3N}}\left(\frac{1}{2}z^THz + c^Tz\right)\\ &\text{s.t. } Cz \le b, \quad z_{N+1},\cdots,z_{3N} \ge 0\end{aligned}$$

with

$$H = 2\begin{pmatrix} \gamma\Phi_X & \mathcal{O}_{N,2N} \\ \mathcal{O}_{2N,N} & \mathcal{O}_{2N,2N} \end{pmatrix} \in R^{3N\times 3N}, \quad c = \frac{1}{N}\begin{pmatrix} \mathcal{O}_N \\ 1_{2N} \end{pmatrix} \in R^{3N},$$

$$C = \begin{pmatrix} \Phi_X & -\mathcal{I}_{N,N} & \mathcal{O}_{N,N} \\ -\Phi_X & \mathcal{O}_{N,N} & -\mathcal{I}_{N,N} \end{pmatrix} \in R^{2N\times 3N}, \quad b = \begin{pmatrix} g|_X + \epsilon 1_N \\ -g|_X + \epsilon 1_N \end{pmatrix} \in R^{2N}.$$

2.3. *Multiscale SVR approximation algorithm*

Main ingredients in the multiscale analysis are the multiple radius of the kernel functions.[17] More precisely, instead of a single sampling data-set, we are dealing with a sequence of sampling data-sets $X_1, X_2, \ldots$, where $X_k := \{x_j^{(k)}\}_{j=1}^{N_k}$ and their fill distances $h_k := h_{X_k}$ monotonically decrease with a pre-specified constant $\mu \in (0, 1)$, i.e. $h_{k+1} = \mu h_k$. Observation data are realized on the nested sampling data-sets and denoted by $g|_{X_1}, g|_{X_2}, \ldots$. For noisy data, we take the form of $g^\delta|_{X_1}, g^\delta|_{X_2}, \ldots$ with a superscript δ.

We choose an integrable CS-SPD-RBF $\Phi : R^d \to R$ with a basic support radii in a unit ball $B(0, 1)$ and define a sequence of kernel functions $\Phi_k : R^d \to R$ in a scaled version:

$$\Phi_k(x, y) := \eta_k^{-d}\Phi\left(\frac{x - y}{\eta_k}\right), \tag{4}$$

where η_k is called the scaling parameter. For the sake of simplicity, we will assume $\eta_k = \nu h_k$ with a constant $\nu > 0$. In Figure 1, we illustrate two sets of scaled kernel functions whose radius will decrease when the level increases.

Furthermore, assume that the basic kernel function Φ satisfies the condition (1), the Fourier coefficients of its scaled version do not inherit such nice properties. Direct calculation shows that the following inequality is satisfied for the scaled kernel function Φ_k

$$c_1\left(1 + \eta_k^2\|\omega\|^2\right)^{-\tau} \le \widehat{\Phi_k}(\omega) \le c_2\left(1 + \eta_k^2\|\omega\|^2\right)^{-\tau}, \quad \omega \in R^d. \tag{5}$$

The inequality (5) thus yields

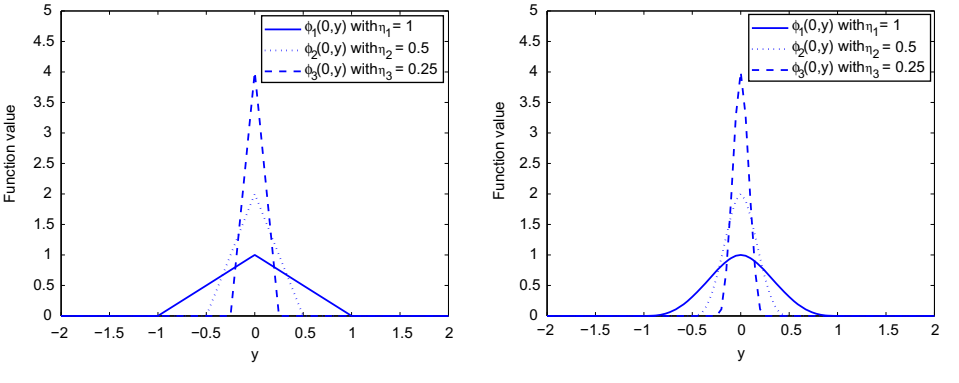


Figure 1. 1-D scaled C^0 (left) and C^2 (right) Wendland function $\Phi_k(0, y)$ for $k = 1, 2, 3$.

LEMMA 2.2 [17, Lemma 1] *For every $\eta_k \in (0,1]$, Φ_k is a kernel function and generates a RKHS $\mathcal{N}_{\Phi_k}(R^d)$ which is equivalent to $H^\tau(R^d)$. For all $f \in H^\tau(R^d)$, we have the norm equivalency*

$$c_1^{\frac{1}{2}} \|f\|_{\Phi_k} \le \|f\|_{H^\tau(R^d)} \le c_2^{\frac{1}{2}} \eta_k^{-\tau} \|f\|_{\Phi_k}.$$

Here and subsequently, we denote $\|\cdot\|_{\Phi_k}$ as $\|\cdot\|_{\mathcal{N}_{\Phi_k}(R^d)}$ for simplicity.

Our proposed multiscale SVR method is summarized as the following algorithm.

Algorithm 1 Multiscale SVR approximation algorithm

Input: Number of levels n, right-hand side $g^\delta|_{X_1}, \ldots, g^\delta|_{X_n}$
Output: Approximate solution $f_n^\delta = s_1^\delta + \cdots + s_n^\delta$

1: Set $f_0^\delta = 0, e_0^\delta = g^\delta$
2: **for** $k = 1, 2, \ldots, n$ **do**
3: Determine a local approximate solution by

$$s_k^\delta = \arg \min_{s \in H^\tau(R^d)} \left(\frac{1}{N_k} \sum_{j=1}^{N_k} \left| e_{k-1}^\delta\left(x_j^{(k)}\right) - s\left(x_j^{(k)}\right) \right|_{\epsilon_k} + \gamma_k \|s\|_{\Phi_k}^2 \right)$$

4: Set $f_k^\delta = f_{k-1}^\delta + s_k^\delta$
5: Set $e_k^\delta = e_{k-1}^\delta - s_k^\delta|_\Omega$
6: **end for**

This algorithm can be viewed as an extension of the original multiscale least squares method in [17]. Defining

$$W_k = \operatorname{span}\left\{ \Phi_k(\cdot, x_j^{(k)}),\ j = 1, 2, \ldots, N_k \right\},$$
$$V_k = W_1 + \cdots + W_k,$$

the local approximate solution s_k^δ belongs to W_k and k-level approximate solution $f_k^\delta \in V_k$. Meanwhile, the residual between the target function and the coarse mesh approximate solution exists on the next scale. A finer data-set and a smaller support radius are chosen to obtain another local approximate solution with even finer details. The sum of these local approximate solutions is interpolated to yield a better approximation on the finer grid. The whole process continues iteratively until a pre-specified level number is reached. Additionally, referring to [17], for any fixed $k \ge 1$, V_k is a closed and finite dimensional subspace of $L^2(R^d)$ satisfying $\overline{\cup_{j=1}^\infty V_j} = L^2(R^d)$.

3. Multiscale SVR method for smooth functions

In this section, the target function f^* is assumed to be smooth enough such that $f^* \in H^\tau(\Omega)$. Recall Algorithm 1, the local approximate solutions s_k are obtained along the scaled kernels Φ_k defined in (4) with the noise-free observation data $g|_{X_k}$ (i.e. $\delta_k = 0$). Notice that the local approximate solution s_k satisfies

$$s_k = \arg \min_{s \in H^\tau(R^d)} \left(\frac{1}{N_k} \sum_{j=1}^{N_k} \left| e_{k-1}\left(x_j^{(k)}\right) - s\left(x_j^{(k)}\right) \right|_{\epsilon_k} + \gamma_k \|s\|_{\Phi_k}^2 \right) \tag{6}$$

and is defined on the whole domain R^d. Meanwhile, the target function f^* is defined on a bounded domain Ω and we have to coordinate the difference between both functions. Assuming that the bounded domain Ω owns a Lipschitz boundary, the following lemma then holds true.

LEMMA 3.1 [42] *Suppose that the bounded domain $\Omega \subset R^d$ is open and has a Lipschitz boundary. Let $\tau \geq 0$, there exists a bounded linear operator $E : H^\tau(\Omega) \to H^\tau\left(R^d\right)$ such that, for all $f \in H^\tau(\Omega)$, $Ef|_\Omega = f$ and*

$$\|f\|_{H^\tau(\Omega)} \leq \|Ef\|_{H^\tau(R^d)} \leq C_\tau \|f\|_{H^\tau(\Omega)},$$

where the constant C_τ only depends on τ. Thus, the same operator E can be utilized for any fixed $\tau \geq 0$.

The following sampling inequality in [43] is also important.

LEMMA 3.2 [43] *Let $\Omega \subset R^d$ be a bounded domain with a Lipschitz boundary and constants θ, σ satisfy $\theta > \sigma + d/2$. For any sampling data-set $X \subseteq \Omega$ with a sufficiently small fill distance h, there exists a constant $C_d := C_d(\sigma, d, \Omega) > 0$ such that $\forall f \in H^\theta(\Omega)$*

$$\|f\|_{H^\sigma(\Omega)} \leq C_d \left(h^{\theta-\sigma}\|f\|_{H^\theta(\Omega)} + h^{-\sigma}\|f\|_{l^\infty(X)}\right).$$

In order to obtain the convergence analysis of our proposed algorithm, we summarize the main assumptions as follows.

ASSUMPTION 3.2

- A1 *Let $\Omega \subset R^d$ be a bounded domain with a Lipschitz boundary*;
- A2 *$\{X_1, X_2, \dots\}$ forms a sequence of sampling data-sets. Each X_k has N_k sampling points where the fill distance h_1 is sufficiently small and $h_{k+1} = \mu h_k$ for $k = 1, 2, \dots$ with a fixed $\mu \in (0, 1)$. Moreover, these sets are quasi-uniform and $N_k \leq c h_k^{-d}$ with a constant $c > 0$;*
- A3 *The basic kernel function Φ generates a RKHS equivalent to the Sobolev space $H^\tau(R^d)$, i.e. Φ satisfies (1) and Φ_k is its scaled version defined by (4) with $\eta_k = \nu h_k$ for a fixed constant $\nu > 0$.*

We establish the main theorem in this section for the convergence analysis of the multiscale SVR approximation algorithm for a target function $f^* \in H^\tau(\Omega)$ with $\tau > d/2$.

THEOREM 3.3 *Let Assumption 3.2 be valid. Furthermore, assume that the constants μ and ν in A2-A3 satisfy $1/h_1 \geq \nu \geq T/\mu$ with a constant $T > 0$. In addition, the observation data are noise free and the target function f^* belongs to $H^\tau(\Omega)$, $\tau > d/2$ satisfying $\|f^*\|_{H^\tau(\Omega)} < c_1^{\frac{1}{2}}/C_\tau$ with constants c_1 in (1) and C_τ in Lemma 3.1.*

At each level k, if we choose the cut-off parameter $\epsilon_k = 0$ and the regularization parameter γ_k

$$\gamma_k \leq \left(\frac{1}{\nu}\right)^\tau h_k^d,$$

then there exists a constant $C_1 > 0$, such that by additionally assuming $p := C_1\mu^\tau < 1$,

$$\|Ee_k\|_{\Phi_{k+1}} < p\,\|Ee_{k-1}\|_{\Phi_k} < 1, \quad k = 1, 2, \ldots, \tag{7}$$

and another constant $C > 0$

$$\left\|f^* - f_n\right\|_{L^2(\Omega)} < Cp^n \left\|f^*\right\|_{H^\tau(\Omega)}.$$

Proof Firstly, by above assumptions, there holds

$$\|Ee_0\|_{\Phi_1} = \left\|Ef^*\right\|_{\Phi_1} \le c_1^{-\frac{1}{2}} C_\tau \left\|f^*\right\|_{H^\tau(\Omega)} < 1.$$

Assuming the induction hypothesis $\|Ee_{k-1}\|_{\Phi_k} < 1$, we obtain the following estimate for $\|Ee_k\|_{\Phi_{k+1}}$ by implementing (5),

$$\begin{aligned}
\|Ee_k\|^2_{\Phi_{k+1}} &= \int_{R^d} \frac{|\widehat{Ee_k}(\omega)|^2}{\widehat{\Phi_{k+1}}(\omega)} d\omega \le \frac{1}{c_1}\int_{R^d} |\widehat{Ee_k}(\omega)|^2 (1 + \eta_{k+1}^2 \|\omega\|^2)^\tau d\omega \\
&= \frac{1}{c_1}\left(\int_{\|\omega\| \le \frac{1}{\eta_{k+1}}} + \int_{\|\omega\| \ge \frac{1}{\eta_{k+1}}}\right) |\widehat{Ee_k}(\omega)|^2 (1 + \eta_{k+1}^2 \|\omega\|^2)^\tau d\omega \\
&:= \frac{1}{c_1}(I_1 + I_2)
\end{aligned}$$

or similarly $\|Ee_k\|_{\Phi_{k+1}} \le c_1^{-\frac{1}{2}}(\sqrt{I_1} + \sqrt{I_2})$. We then estimate error estimates for these two terms $\sqrt{I_1}$, $\sqrt{I_2}$ separately.

Using Lemmas 3.1 and 3.2, $\sqrt{I_1}$ can be firstly estimated in the following sense

$$\begin{aligned}
\sqrt{I_1} &\le 2^{\frac{\tau}{2}} \|Ee_k\|_{L^2(R^d)} \le 2^{\frac{\tau}{2}} C_0 \|e_k\|_{L^2(\Omega)} \\
&\le 2^{\frac{\tau}{2}} C_0 C_d \left(h_k^\tau \|e_k\|_{H^\tau(\Omega)} + \|e_k\|_{l^\infty(X_k)}\right).
\end{aligned}$$

Considering the minimization problem (6) and taking $Ee_{k-1} = Ef^* - s_1 - \ldots - s_{k-1} \in H^\tau\left(R^d\right)$ as a candidate, we obtain

$$e_{k-1}|_{X_k} = Ee_{k-1}|_{X_k},$$

and the minimality of the cost function implies that

$$\begin{aligned}
&\frac{1}{N_k}\sum_{j=1}^{N_k} \left|e_{k-1}\left(x_j^{(k)}\right) - s_k\left(x_j^{(k)}\right)\right|_{\epsilon_k} + \gamma_k \|s_k\|^2_{\Phi_k} \\
&\le \frac{1}{N_k}\sum_{j=1}^{N_k} \left|e_{k-1}\left(x_j^{(k)}\right) - Ee_{k-1}\left(x_j^{(k)}\right)\right|_{\epsilon_k} + \gamma_k \|Ee_{k-1}\|^2_{\Phi_k} \\
&= \gamma_k \|Ee_{k-1}\|^2_{\Phi_k}.
\end{aligned}$$

Thus, the induction hypothesis and Assumption 3.2 yield

$$\begin{aligned}\|e_k\|_{l^\infty(X_k)} &= \|e_{k-1} - s_k\|_{l^\infty(X_k)} = \max_{j=1,\ldots,N_k} \left|e_{k-1}\left(x_j^{(k)}\right) - s_k\left(x_j^{(k)}\right)\right| \\ &\le \max_{j=1,\ldots,N_k} \left|e_{k-1}\left(x_j^{(k)}\right) - s_k\left(x_j^{(k)}\right)\right|_{\epsilon_k} + \epsilon_k \\ &\le \sum_{j=1}^{N_k} \left|e_{k-1}\left(x_j^{(k)}\right) - s_k\left(x_j^{(k)}\right)\right|_{\epsilon_k} + \epsilon_k \\ &\le N_k\gamma_k \|Ee_{k-1}\|_{\Phi_k}^2 + \epsilon_k \le ch_k^{-d}\gamma_k \|Ee_{k-1}\|_{\Phi_k}^2 + \epsilon_k \\ &< ch_k^{-d}\gamma_k \|Ee_{k-1}\|_{\Phi_k} + \epsilon_k, \end{aligned} \tag{8}$$

and

$$\|s_k\|_{\Phi_k} \le \|Ee_{k-1}\|_{\Phi_k}.$$

With $\eta_{k+1} < 1$, Lemma 2.2 leads to

$$\begin{aligned}\|e_k\|_{H^\tau(\Omega)} &= \|e_{k-1} - s_k\|_{H^\tau(\Omega)} \\ &\le c_2^{\frac{1}{2}}\eta_k^{-\tau} \|Ee_{k-1} - s_k\|_{\Phi_k} \\ &\le 2c_2^{\frac{1}{2}}\eta_k^{-\tau} \|Ee_{k-1}\|_{\Phi_k}. \end{aligned} \tag{9}$$

Combining inequalities (8) and (9) and the choices of parameters ϵ_k and γ_k, we obtain the estimate for the term $\sqrt{I_1}$,

$$\begin{aligned}\sqrt{I_1} &< 2^{\frac{\tau}{2}} C_0 C_d \left(2c_2^{\frac{1}{2}}\left(\frac{1}{\nu}\right)^\tau \|Ee_{k-1}\|_{\Phi_k} + ch_k^{-d}\gamma_k \|Ee_{k-1}\|_{\Phi_k} + \epsilon_k\right) \\ &\le 2^{\frac{\tau}{2}} C_0 C_d \left(2c_2^{\frac{1}{2}} + c\right)\left(\frac{1}{\nu}\right)^\tau \|Ee_{k-1}\|_{\Phi_k} \\ &\le \frac{2^{\frac{\tau}{2}} C_0 C_d \left(2c_2^{\frac{1}{2}} + c\right)}{T^\tau}\mu^\tau \|Ee_{k-1}\|_{\Phi_k}. \end{aligned}$$

Noticing that $\eta_{k+1}\|\omega\| \ge 1$ implies

$$\left(1 + \eta_{k+1}^2\|\omega\|^2\right)^\tau \le 2^\tau \eta_{k+1}^{2\tau}\|\omega\|^{2\tau} \le 2^\tau \eta_{k+1}^{2\tau}\left(1 + \|\omega\|^2\right)^\tau,$$

we then derive the estimate for the second term $\sqrt{I_2}$, by the inequality (9),

$$\begin{aligned}\sqrt{I_2} &\le 2^{\frac{\tau}{2}}\eta_{k+1}^\tau \|Ee_k\|_{H^\tau(R^d)} \le 2^{\frac{\tau}{2}} C_\tau \eta_{k+1}^\tau \|e_k\|_{H^\tau(\Omega)} \\ &\le 2^{\frac{\tau}{2}+1} c_2^{\frac{1}{2}} C_\tau \left(\frac{\eta_{k+1}}{\eta_k}\right)^\tau \|Ee_{k-1}\|_{\Phi_k} \\ &= 2^{\frac{\tau}{2}+1} c_2^{\frac{1}{2}} C_\tau \mu^\tau \|Ee_{k-1}\|_{\Phi_k}. \end{aligned}$$

We thus obtain

$$\begin{aligned}\|Ee_k\|_{\Phi_{k+1}} &< c_1^{-\frac{1}{2}} \left(\frac{2^{\frac{\tau}{2}} C_0 C_d \left(2c_2^{\frac{1}{2}} + c\right)}{T^\tau} + 2^{\frac{\tau}{2}+1} c_2^{\frac{1}{2}} C_\tau \right) \mu^\tau \|Ee_{k-1}\|_{\Phi_k} \\ &= C_1 \mu^\tau \|Ee_{k-1}\|_{\Phi_k}.\end{aligned}$$

If we additionally assume $p := C_1\mu^\tau < 1$, it yields

$$\|Ee_k\|_{\Phi_{k+1}} < p \|Ee_{k-1}\|_{\Phi_k} < 1.$$

Moreover, the error estimate between the approximate solution and the target one can be derived by

$$\begin{aligned}\left\|f^* - f_n\right\|_{L^2(\Omega)} &= \|e_n\|_{L^2(\Omega)} \le C_d \left(h_n^\tau \|e_n\|_{H^\tau(\Omega)} + \|e_n\|_{l^\infty(X_n)}\right) \\ &< \frac{C_d \left(2c_2^{\frac{1}{2}} + c\right)}{T^\tau C_1} p \|Ee_{n-1}\|_{\Phi_n} \\ &< \cdots < \frac{C_d \left(2c_2^{\frac{1}{2}} + c\right)}{T^\tau C_1} p^n \|Ee_0\|_{\Phi_1} \le \frac{C_d \left(2c_2^{\frac{1}{2}} + c\right) C_\tau}{T^\tau C_1 c_1^{\frac{1}{2}}} p^n \left\|f^*\right\|_{H^\tau(\Omega)} \\ &= C p^n \left\|f^*\right\|_{H^\tau(\Omega)},\end{aligned}$$

which completes the proof. Thus, the multiscale SVR approximate solution f_n converges linearly to f^* in the space $L^2(\Omega)$ provided with $p < 1$. □

4. Multiscale SVR method for rough functions

Convergence analysis in the previous section highly depends on the *a priori* knowledge of the target function $f^* \in H^\tau(\Omega)$. More precisely, the regularity of the target function should be known as *a priori*, which allows us to choose an appropriate kernel function and generate a RKHS with the same regularity. For instance, Algorithm 1 is realized with the scaled kernel function Φ_k by the basic one $\Phi : R^d \to R$ satisfying (1) and generating a RKHS equivalent to $H^\tau\left(R^d\right)$.

However, in real situation, the regularity information of the target function is usually unknown beforehand. In this section, we will investigate the error estimate for Algorithm 1 when a rough target function $f^* \in H^\beta(\Omega)$ $(\frac{d}{2} < \beta \le \tau)$ is considered. We emphasize that such a case is not covered by Section 3 since $\|f^*\|_{H^\tau(\Omega)}$ might not be bounded for $f^* \in H^\beta(\Omega)$ with $\beta < \tau$.

We note that the assumption $\beta > \frac{d}{2}$ is necessary to guarantee the boundedness of point-wise evaluation, i.e. the target function is continuous. In addition, for the sake of simplicity, we assume the kernel function of $H^\beta\left(R^d\right)$ is translation invariant and denoted by $\Psi : R^d \to R$ satisfying

$$\widehat{\Psi}(\omega) = \left(1 + \|\omega\|^2\right)^{-\beta}, \quad \omega \in R^d. \tag{10}$$

Similarly, we define the scaled kernel functions with respect to Ψ in the form

$$\Psi_k(x, y) = \eta_k^{-d} \Psi\left(\frac{x - y}{\eta_k}\right).$$

It is worth to note that, since β is unknown, the basic kernel Ψ and its associated scaled versions Ψ_k only appear in theoretical analysis. Thus, their explicit forms are not necessary to be known exactly.

Let

$$\mathcal{B}_\sigma = \left\{ f \in L^2\left(R^d\right) : \operatorname{supp}(\widehat{f}) \subset B(0, \sigma) \right\},$$

where $B(0, \sigma) \subset R^d$ denotes a ball with the center 0 and a radius σ. To derive the convergence analysis for rough functions, some result concerning the band-limited function in [17] is necessary.

LEMMA 4.1 [17, Lemma 5] *Let $f \in H^\beta\left(R^d\right)$ with $\beta > \frac{d}{2}$, and $X \subset R^d$ be a finite sampling data-set with the separation distance $q_X \leq \eta$ given $\eta \in (0, 1]$. Choose $\sigma = \kappa\eta/q_X$ with a constant $\kappa \geq 1$ depending on d and β. Then there exists a function $f_{\sigma/\eta} \in \mathcal{B}_{\sigma/\eta}$ with $f_{\sigma/\eta}|_X = f|_X$ and*

$$\left\| f_{\sigma/\eta} \right\|_{\Phi_\eta} \leq C_b \sigma^{\tau-\beta} \left\| f \right\|_{\Psi_\eta}$$

where Φ_η and Ψ_η are scaled kernel functions with a scaling parameter η.

Referring to Lemma 4.1, for any fixed level k, although $f^* \in H^\beta(\Omega)$ and $Ee_{k-1} = Ef^* - s_1 - \ldots - s_{k-1} \in H^\beta\left(R^d\right)$, the element $(Ee_{k-1})_{\sigma_k/\eta_k}$ actually belongs to $H^\tau(R^d)$ if we assume $q_k := q_{X_k}$ with $q_k \simeq h_k$ or $q_k \simeq \eta_k$. Thus, we can take $(Ee_{k-1})_{\sigma_k/\eta_k}$ as a candidate of the minimization problem (6). Note that

$$(Ee_{k-1})_{\sigma_k/\eta_k}|_{X_k} = Ee_{k-1}|_{X_k} = e_{k-1}|_{X_k}.$$

Again since s_k is the minimizer of the minimization problem (6), it follows that

$$\begin{aligned} \|e_k\|_{l^\infty(X_k)} &= \|e_{k-1} - s_k\|_{l^\infty(X_k)} \\ &\leq \sum_{j=1}^{N_k} \left| e_{k-1}\left(x_j^{(k)}\right) - s_k\left(x_j^{(k)}\right) \right|_{\epsilon_k} + \epsilon_k \\ &\leq N_k \gamma_k \left\| (Ee_{k-1})_{\sigma_k/\eta_k} \right\|_{\Phi_k}^2 + \epsilon_k \\ &\leq c C_b^2 h_k^{-d} \gamma_k \sigma_k^{2\tau-2\beta} \left\| Ee_{k-1} \right\|_{\Psi_k}^2 + \epsilon_k, \end{aligned} \tag{11}$$

where the last inequality is directly from Lemma 4.1, and the following inequality holds as well

$$\|s_k\|_{\Phi_k} \leq \left\| (Ee_{k-1})_{\sigma_k/\eta_k} \right\|_{\Phi_k} \leq C_b \sigma_k^{\tau-\beta} \left\| Ee_{k-1} \right\|_{\Psi_k}.$$

Moreover, referring to (5) and (10), there holds

$$\begin{aligned}\|s_k\|_{\Psi_k}^2 &= \int_{R^d} \frac{|\widehat{s_k}(\omega)|^2}{\widehat{\Psi_k}(\omega)} d\omega = \int_{R^d} |\widehat{s_k}(\omega)|^2 \left(1+\eta_k^2\|\omega\|^2\right)^{\beta} d\omega \\ &\le \int_{R^d} |\widehat{s_k}(\omega)|^2 \left(1+\eta_k^2\|\omega\|^2\right)^{\tau} d\omega \\ &\le c_2 \int_{R^d} \frac{|\widehat{s_k}(\omega)|^2}{\widehat{\Phi_k}(\omega)} d\omega = c_2 \|s_k\|_{\Phi_k}^2 .\end{aligned}$$

Thus, the following estimate can be derived:

$$\begin{aligned}\|e_k\|_{H^\beta(\Omega)} &= \|e_{k-1}-s_k\|_{H^\beta(\Omega)} \le \|Ee_{k-1}-s_k\|_{H^\beta(R^d)} \\ &\le \eta_k^{-\beta} \|Ee_{k-1}-s_k\|_{\Psi_k} \le \eta_k^{-\beta} \left(\|Ee_{k-1}\|_{\Psi_k} + \|s_k\|_{\Psi_k}\right) \\ &\le \eta_k^{-\beta} \left(1 + c_2^{\frac{1}{2}} C_b \sigma_k^{\tau-\beta}\right) \|Ee_{k-1}\|_{\Psi_k} . \end{aligned} \tag{12}$$

Now, we are ready to prove convergence results of the multiscale SVR approximation algorithm for rough functions.

THEOREM 4.2 *Let Assumption 3.2 be valid. Furthermore assume that the constants μ and ν in* A2–A3 *satisfy $1/h_1 \ge \nu \ge T/\mu \ge 1$ with a constant $T > 0$. The fill distances $h_1, h_2, \ldots$ and separation distances $q_1, q_2, \ldots$ satisfy $q_k \le h_k \le c_q q_k$, $c_q > 1$ in each level for $k = 1, 2, \ldots$. In addition, the observation data are noise free and the target function f^* belongs to $H^\beta(\Omega)$, $d/2 < \beta \le \tau$ satisfying $\|f^*\|_{H^\beta(\Omega)} < 1/C_\beta$ with a constant C_β in Lemma 3.1.*

At each level k, if we choose the cut-off parameter $\epsilon_k = 0$ and the regularization parameter γ_k

$$\gamma_k \le \left(\frac{1}{\nu}\right)^{\beta} h_k^d,$$

then there exists a constant $C_2 > 0$, such that, by additionally assuming $p := C_2\mu^\beta < 1$,

$$\|Ee_k\|_{\Psi_{k+1}} < p \|Ee_{k-1}\|_{\Psi_k} < 1 \quad \text{for } k = 1, 2, \ldots, \tag{13}$$

and another constant $C > 0$

$$\left\|f^* - f_n\right\|_{L^2(\Omega)} < C p^n \|f^*\|_{H^\beta(\Omega)}.$$

Proof Firstly, by above assumptions, we obtain

$$\|Ee_0\|_{\Psi_1} = \|Ee_0\|_{H^\beta(R^d)} \le C_\beta \left\|f^*\right\|_{H^\beta(\Omega)} < 1.$$

To obtain the inequality (13), similar arguments of Theorem 3.3 will be utilized. In fact, we assume $\|Ee_{k-1}\|_{\Psi_k} < 1$ and estimate

$$\begin{aligned}\|Ee_k\|_{\Psi_{k+1}} &= \left(\int_{R^d} |\widehat{Ee_k}(\omega)|^2 \left(1+\eta_{k+1}^2\|\omega\|^2\right)^{\beta} d\omega\right)^{\frac{1}{2}} \\ &\le \sqrt{I_1} + \sqrt{I_2},\end{aligned}$$

where

$$I_1 := \int_{\|\omega\| \le \frac{1}{\eta_{k+1}}} |\widehat{Ee_k}(\omega)|^2 \left(1 + \eta_{k+1}^2 \|\omega\|^2\right)^\beta d\omega,$$
$$I_2 := \int_{\|\omega\| \ge \frac{1}{\eta_{k+1}}} |\widehat{Ee_k}(\omega)|^2 \left(1 + \eta_{k+1}^2 \|\omega\|^2\right)^\beta d\omega.$$

Concerning the choice of parameters ϵ_k, γ_k, Lemma 3.1, the sampling inequality and both estimates (11) and (12), we can derive the estimate for the first item

$$\begin{aligned}
\sqrt{I_1} &\le 2^{\frac{\beta}{2}} \|Ee_k\|_{L^2(R^d)} \le 2^{\frac{\beta}{2}} C_0 \|e_k\|_{L^2(\Omega)} \\
&\le 2^{\frac{\beta}{2}} C_0 C_d \left(h_k^\beta \|e_k\|_{H^\beta(\Omega)} + \|e_k\|_{l^\infty(X_k)}\right) \\
&\le 2^{\frac{\beta}{2}} C_0 C_d \left(\left(\frac{1}{\nu}\right)^\beta \left(1 + c_2^{\frac{1}{2}} C_b \sigma_k^{\tau-\beta}\right) \|Ee_{k-1}\|_{\Psi_k} \right. \\
&\qquad \left. + c C_b^2 h_k^{-d} \gamma_k \sigma_k^{2\tau-2\beta} \|Ee_{k-1}\|_{\Psi_k}^2 + \epsilon_k\right) \\
&< 2^{\frac{\beta}{2}} C_0 C_d \left(1 + c_2^{\frac{1}{2}} C_b \sigma_k^{\tau-\beta} + c C_b^2 \sigma_k^{2\tau-2\beta}\right) \left(\frac{1}{\nu}\right)^\beta \|Ee_{k-1}\|_{\Psi_k} \\
&\le \frac{2^{\frac{\beta}{2}} C_0 C_d \left(1 + c_2^{\frac{1}{2}} C_b \sigma_k^{\tau-\beta} + c C_b^2 \sigma_k^{2\tau-2\beta}\right)}{T^\beta} \mu^\beta \|Ee_{k-1}\|_{\Psi_k}.
\end{aligned}$$

The second item is estimated by (12) such that

$$\begin{aligned}
\sqrt{I_2} &\le 2^{\frac{\beta}{2}} \eta_{k+1}^\beta \|Ee_k\|_{H^\beta(R^d)} \le 2^{\frac{\beta}{2}} C_\beta \eta_{k+1}^\beta \|e_k\|_{H^\beta(\Omega)} \\
&\le 2^{\frac{\beta}{2}} C_\beta \left(\frac{\eta_{k+1}}{\eta_k}\right)^\beta \left(1 + c_2^{\frac{1}{2}} C_b \sigma_k^{\tau-\beta}\right) \|Ee_{k-1}\|_{\Psi_k} \\
&= 2^{\frac{\beta}{2}} C_\beta \left(1 + c_2^{\frac{1}{2}} C_b \sigma_k^{\tau-\beta}\right) \mu^\beta \|Ee_{k-1}\|_{\Psi_k}.
\end{aligned}$$

Combining these two items and noticing that $\sigma_k = \kappa \eta_k / q_k \le \kappa \nu c_q$, we thus have proven

$$\|Ee_k\|_{\Psi_{k+1}} < C_2 \mu^\beta \|Ee_{k-1}\|_{\Psi_k}$$

by adjusting the constant C_2 appropriately. If we additionally assume $p := C_2 \mu^\beta < 1$ and implement the sequence estimate (13), the theorem can be proven following similar arguments of Theorem 3.3. □

5. Multiscale SVR method with noisy observation data

Previous analysis depends on the fact that all the nested observation data are noise free. Nevertheless, in real function approximation, the observation data are sometimes contaminated by noise. These noisy data might not fit any target function which belongs to a certain Sobolev space $H^\tau(\Omega)$, $\tau > d/2$. Thus, Algorithm 1 becomes more sophisticated and depends on various parameters that have to be finely tuned.

In this section, we investigate the performance of Algorithm 1 towards noisy observation data of a smooth function. More precisely, assume that the target function f^* belongs to $H^\tau(\Omega)$, the nested observation data $g^\delta|_{X_1}, g^\delta|_{X_2}, \ldots$ contain some noise such that

$$\left\| g^\delta - f^* \right\|_{l^\infty(X_k)} \leq \delta_k. \tag{14}$$

Algorithm 1, at each level k, then provides the local approximate solution s_k^δ satisfying

$$s_k^\delta = \arg\min_{s \in H^\tau(R^d)} \left(\frac{1}{N_k} \sum_{j=1}^{N_k} |e_{k-1}^\delta \left(x_j^{(k)} \right) - s \left(x_j^{(k)} \right)|_{\epsilon_k} + \gamma_k \|s\|_{\Phi_k}^2 \right). \tag{15}$$

Defining a new function $Ee_{k-1} = Ef^* - s_1^\delta - \ldots - s_{k-1}^\delta$ which belongs to $H^\tau\left(R^d\right)$, we derive

$$\begin{aligned} \frac{1}{N_k} \sum_{j=1}^{N_k} \left| e_{k-1}^\delta \left(x_j^{(k)} \right) - Ee_{k-1} \left(x_j^{(k)} \right) \right|_{\epsilon_k} &= \frac{1}{N_k} \sum_{j=1}^{N_k} \left| g^\delta \left(x_j^{(k)} \right) - Ef^* \left(x_j^{(k)} \right) \right|_{\epsilon_k} \\ &\leq \frac{1}{N_k} \cdot N_k \cdot (\delta_k - \epsilon_k)_+ = (\delta_k - \epsilon_k)_+. \end{aligned}$$

Let Assumption 3.2 hold true, choose the cut-off parameters $\epsilon_k = \delta_k$ (or equivalently $(\delta_k - \epsilon_k)_+ = 0$) and use the minimality of the functional in (15), we have

$$\begin{aligned} \|e_k\|_{l^\infty(X_k)} &= \left\| e_{k-1} - s_k^\delta \right\|_{l^\infty(X_k)} \leq \left\| e_{k-1} - e_{k-1}^\delta \right\|_{l^\infty(X_k)} + \left\| e_{k-1}^\delta - s_k^\delta \right\|_{l^\infty(X_k)} \\ &\leq \delta_k + \sum_{j=1}^{N_k} \left| e_{k-1}^\delta \left(x_j^{(k)} \right) - s_k^\delta \left(x_j^{(k)} \right) \right|_{\epsilon_k} + \epsilon_k \\ &\leq \delta_k + N_k (\delta_k - \epsilon_k)_+ + N_k \gamma_k \|Ee_{k-1}\|_{\Phi_k}^2 + \epsilon_k \\ &\leq 2\delta_k + c h_k^{-d} \gamma_k \|Ee_{k-1}\|_{\Phi_k}^2, \end{aligned} \tag{16}$$

and,

$$\left\| s_k^\delta \right\|_{\Phi_k} \leq \sqrt{\frac{(\delta_k - \epsilon_k)_+}{\gamma_k}} + \|Ee_{k-1}\|_{\Phi_k} = \|Ee_{k-1}\|_{\Phi_k}.$$

The following inequality then holds true

$$\begin{aligned} \|e_k\|_{H^\tau(\Omega)} &= \left\| e_{k-1} - s_k^\delta \right\|_{H^\tau(\Omega)} \leq c_2^{\frac{1}{2}} \eta_k^{-\tau} \left\| Ee_{k-1} - s_k^\delta \right\|_{\Phi_k} \\ &\leq 2 c_2^{\frac{1}{2}} \eta_k^{-\tau} \|Ee_{k-1}\|_{\Phi_k}. \end{aligned} \tag{17}$$

With both above inequalities (16) and (17), we summarize the error estimate between the multiscale SVR approximate solution f_n^δ and the target function f^* for noisy observation data.

THEOREM 5.1 *Let Assumption 3.2 be valid. Furthermore, assume that the constants μ and ν in* A2 *and* A3 *satisfy $1/h_1 \geq \nu \geq T/\mu$ with a constant $T > 0$. In addition, assume that the target function f^* belongs to $H^\tau(\Omega)$ with $\|f^*\|_{H^\tau(\Omega)} < c_1^{\frac{1}{2}}/C_\tau$, and the observation data $g^\delta|_{X_k}$ for $k = 1, 2, \ldots$ contain some noise satisfying (14) with noise levels δ_k.*

At each level, if we choose the cut-off parameter $\epsilon_k = \delta_k$ *and the regularization parameter* γ_k

$$\gamma_k \leq \left(\frac{1}{\nu}\right)^{\tau} h_k^d,$$

then there exist constants $C_3, C_4 > 0$*, such that*

$$\|Ee_k\|_{\Phi_{k+1}} < C_3\mu^{\tau} \|Ee_{k-1}\|_{\Phi_k} + C_4\delta_k. \tag{18}$$

Denoting $p := C_3\mu^{\tau}$ *and further assuming* $p + C_4\delta_k < 1$*, then there exists a constant* C *such that*

$$\left\| f^* - f_n^{\delta} \right\|_{L^2(\Omega)} < C \left(p^n \left\| f^* \right\|_{H^{\tau}(\Omega)} + \sum_{j=0}^{n-1} p^j \delta_{n-j} \right).$$

Proof The proof is similar to that of Theorem 3.3 by splitting

$$\begin{aligned}
\|Ee_k\|_{\Phi_{k+1}} \leq\ & \frac{1}{c_1^{\frac{1}{2}}} \left[\left(\int_{\|\omega\| \leq \frac{1}{\eta_{k+1}}} |\widehat{Ee_k}(\omega)|^2 \left(1 + \eta_{k+1}^2 \|\omega\|^2\right)^{\tau} d\omega \right)^{\frac{1}{2}} \right. \\
& \left. + \left(\int_{\|\omega\| \geq \frac{1}{\eta_{k+1}}} |\widehat{Ee_k}(\omega)|^2 \left(1 + \eta_{k+1}^2 \|\omega\|^2\right)^{\tau} d\omega \right)^{\frac{1}{2}} \right] \\
:=\ & \frac{1}{c_1^{\frac{1}{2}}} \left(\sqrt{I_1} + \sqrt{I_2} \right).
\end{aligned}$$

To estimate the first term $\sqrt{I_1}$, we recall Lemma 3.1, the sampling inequality, both estimates (16) and (17) and the choice of ϵ_k and γ_k such that

$$\begin{aligned}
\sqrt{I_1} &\leq 2^{\frac{\tau}{2}} C_0 \|e_k\|_{L^2(\Omega)} \\
&\leq 2^{\frac{\tau}{2}} C_0 C_d \left(h_k^{\tau} \|e_k\|_{H^{\tau}(\Omega)} + \|e_k\|_{l^{\infty}(X_k)} \right) \\
&\leq 2^{\frac{\tau}{2}} C_0 C_d \left(2c_2^{\frac{1}{2}} \left(\frac{1}{\nu}\right)^{\tau} \|Ee_{k-1}\|_{\Phi_k} + c \left(\frac{1}{\nu}\right)^{\tau} \|Ee_{k-1}\|_{\Phi_k}^2 + 2\delta_k \right) \\
&< \frac{2^{\frac{\tau}{2}} C_0 C_d \left(2c_2^{\frac{1}{2}} + c \right)}{T^{\tau}} \mu^{\tau} \|Ee_{k-1}\|_{\Phi_k} + 2^{\frac{\tau}{2}+1} C_0 C_d \delta_k
\end{aligned}$$

with an additional hypothesis $\|Ee_{k-1}\|_{\Phi_k} < 1$.

The second term $\sqrt{I_2}$ can be estimated as follows:

$$\sqrt{I_2} \leq 2^{\frac{\tau}{2}} C_{\tau} \eta_{k+1}^{\tau} \|e_k\|_{H^{\tau}(\Omega)} \leq 2^{\frac{\tau}{2}+1} c_2^{\frac{1}{2}} C_{\tau} \mu^{\tau} \|Ee_{k-1}\|_{\Phi_k}.$$

Combining both terms, we derive

$$\begin{aligned}\|Ee_k\|_{\Phi_{k+1}} &< \frac{2^{\frac{\tau}{2}}}{c_1^{\frac{1}{2}}}\left(\frac{C_0C_d\left(2c_2^{\frac{1}{2}}+c\right)}{T^\tau}+2c_2^{\frac{1}{2}}C_\tau\right)\mu^\tau\|Ee_{k-1}\|_{\Phi_k}+2^{\frac{\tau}{2}+1}C_0C_d\delta_k\\ &:= C_3\mu^\tau\|Ee_{k-1}\|_{\Phi_k}+C_4\delta_k := p\|Ee_{k-1}\|_{\Phi_k}+C_4\delta_k\\ &< 1,\end{aligned}$$

by induction and noticing the assumption $\|Ee_0\|_{\Phi_1} \le c_1^{-\frac{1}{2}}C_\tau\|f^*\|_{H^\tau(\Omega)} < 1$.

Finally, by utilizing the above inequality (18) and further assumptions, it is easy to derive

$$\begin{aligned}\left\|f^*-f_n^\delta\right\|_{L^2(\Omega)} &= \|e_n\|_{L^2(\Omega)} \le C_d\left(h_n^\tau\|e_n\|_{H^\tau(\Omega)}+\|e_n\|_{l^\infty(X_n)}\right)\\ &< \frac{C_d\left(2c_2^{\frac{1}{2}}+c\right)}{C_3T^\tau}p\|Ee_{n-1}\|_{\Phi_n}+2C_d\delta_n\\ &< \frac{C_d\left(2c_2^{\frac{1}{2}}+c\right)}{C_3T^\tau}p\left(p\|Ee_{n-2}\|_{\Phi_{n-1}}+C_4\delta_{n-1}\right)+2C_d\delta_n\\ &< \cdots < \frac{C_d\left(2c_2^{\frac{1}{2}}+c\right)}{C_3T^\tau}\left(p^n\|Ee_0\|_{\Phi_1}+C_4\sum_{j=1}^{n-1}p^j\delta_{n-j}\right)+2C_d\delta_n\\ &< C\left(p^n\left\|f^*\right\|_{H^\tau(\Omega)}+\sum_{j=0}^{n-1}p^j\delta_{n-j}\right)\end{aligned}$$

by adjusting the constant C appropriately. □

In most applications, the known data are usually observed only once, which allows us to obtain a full observation data $g^\delta|_X$ with $\left\|g^\delta-f^*\right\|_{l^\infty(X)} \le \delta$. We then split X into nested data-sets $X_1, X_2, \cdots$ satisfying $X = \cup_{k=1}^n X_k$. In this case, we have $\delta_1 = \delta_2 = \cdots = \delta_n = \delta$ and Theorem 5.1 turns into

THEOREM 5.2 *Under all assumptions in Theorem 5.1, if $\delta_k = \delta$ for each $k = 1, 2, \ldots$, with the same constant C, we have*

$$\left\|f^*-f_n^\delta\right\|_{L^2(\Omega)} < C\left(p^n\left\|f^*\right\|_{H^\tau(\Omega)}+\frac{1-p^n}{1-p}\delta\right).$$

Thus, if we further assume $p < 1$ and let n go to infinity, the error between multiscale SVR approximate solution f_n^δ and the target function f^* in the L^2 norm can be bounded by $C\delta/(1-p)$.

6. Numerical examples

6.1. *Multiscale SVR method for smooth functions*

In this subsection, we firstly illustrate the performance of our proposed Algorithm 1 for smooth functions whose error estimate is established in Section 3. To this end, we define a bounded domain $\Omega = [-1, 1] \subset R^1$ and the target function is chosen in such a way

$$f^*(x) = \sin(\pi x) + \cos(\pi x).$$

The nested sampling data-sets are equally distributed in the bounded domain Ω with increasing discrete levels $\mathcal{N} = \{25, 50, 100, 200, 400\}$. We choose the basic kernel function $\Phi_{1,1} \in C^2(R^1)$ which generates a RKHS $H^2(R^1)$ and build its associated scaled versions.

Let $\epsilon_k = 0$ and $\gamma_k = 0.1h_k$ for $k = 1, 2, \ldots,$ Table 2 collects the quantity information including the discrete levels N_k, fill distances h_k, absolute errors in l^2 norm and l^∞ norms, respectively. As the theorem predicts, these errors are monotonically decreasing as the level increases which confirms the reliability of the multiscale SVR approximation algorithm. We note that convergence rate of the algorithm can be partially observed from the final column in Table 2.

A second example is displayed in R^2. Let $\Omega = [0, 1] \times [0, 1] \subset R^2$, the target function is the standard Franke function, which is a sum of four exponentials. This time, we choose the basic kernel function $\Phi_{2,1} \in C^2\left(R^2\right)$ which generates the RKHS $H^{2.5}(R^2)$. The nested sampling data-sets are divided into five levels and equally distributed in Ω, with discrete levels $\mathcal{N} = \{9, 25, 81, 189, 441\}$. Referring to Theorem 3.3, after choosing $\epsilon_k = 0$ and $\gamma_k = 0.1h_k^2$ for $k = 1, 2, \ldots,$ we obtain the final approximate solution f_5 which is plotted in left panel of Figure 2, where in right panel we present the associated error e_5. For comparison, we provide the approximate solution on the single mesh with the basic kernel function in Figure 3 with larger associated error.

6.2. *Multiscale SVR method for rough functions*

The performance of Algorithm 1 for rough functions is presented in this subsection. We choose the target function

$$f^*(x) = 2\min(x, 1 - x),$$

defined on $\Omega = [0, 1]$ and belongs to $H^1(\Omega)$. Let $\mathcal{N} = \{25, 50, 100, 200, 400\}$ and the basic kernel function $\Phi_{1,1} \in C^2(R^1)$ which generates a RKHS equivalent to $H^2(R^1)$. Choosing the cut-off parameter $\epsilon_k = 0$ and $\gamma_k = 0.1h_k$ for $k = 1, 2, \ldots,$ the convergence

Table 2. (Smooth function) Quantity information at each level.

level	N_k	h_k	l^2 error	l^∞ error	$\log \frac{1}{\|f_k - f^*\|_{l^2}}$
1	25	0.08	0.088408	0.030131	1.05
2	50	0.04	0.001358	0.000499	2.87
3	100	0.02	0.000061	0.000045	4.21
4	200	0.01	0.000019	0.000010	4.72
5	400	0.005	0.000015	0.000004	4.82

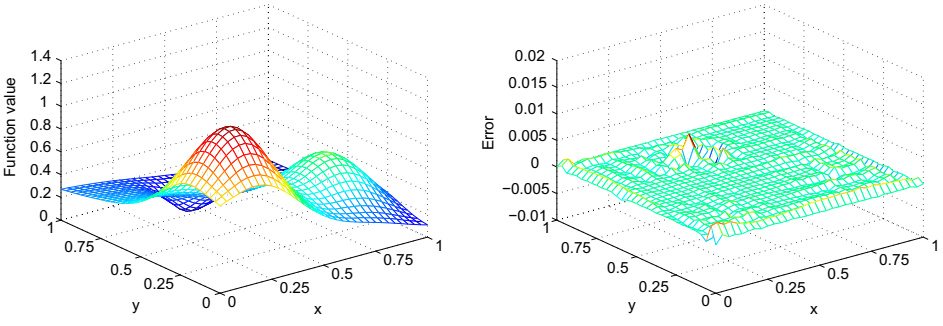


Figure 2. (Smooth function) Multiscale SVR approximate solution f_5 (left). Error between f_5 and f^* (right).

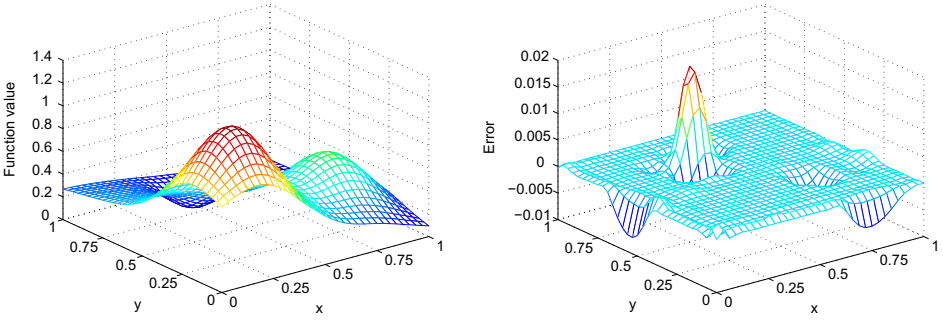


Figure 3. (Smooth function) Single mesh SVR approximate solution (left). Error between the approximate solution and f^* (right).

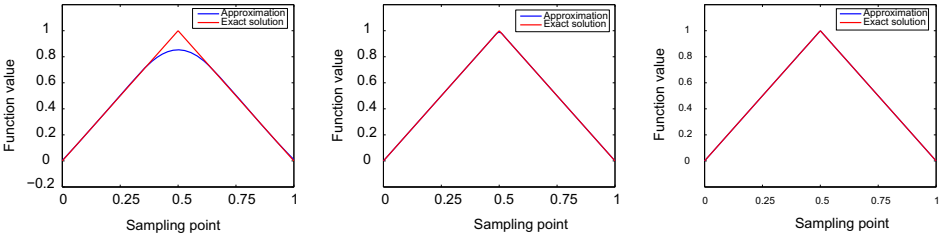


Figure 4. (Rough function) Multiscale SVR approximate solution f_k with respect to the rough target solution f^*. Left: $k = 1$; Middle: $k = 3$; Right: $k = 5$.

behavior can be observed in Figure 4, where we present the comparison between multiscale SVR approximate solutions f_k with $k = 1, 3, 5$ and the target function f^*. Table 3 collects the detailed quantity information for these approximate solutions.

For comparison, we choose another kernel function $\Phi_{1,0} \in C(R^1)$ which generates a RKHS equivalent to $H^1(R^1)$ where the target function f^* belongs. The associated l^2 error

Table 3. (Rough function) Quantity information for the approximate solution of rough functions with oversmoothing kernels.

level	N_k	l^2 and l^∞ error ($\Phi_{1,0}$)	l^2 and l^∞ error ($\Phi_{1,1}$)	l^2 and l^∞ error ($\Phi_{1,2}$)
1	25	0.2117 and 0.0789	0.7166 and 0.1438	0.7832 and 0.1525
2	50	0.0717 and 0.0402	0.1192 and 0.0419	0.1548 and 0.0498
3	100	0.0278 and 0.0203	0.0177 and 0.0100	0.0413 and 0.0194
4	200	0.0120 and 0.0103	0.0059 and 0.0028	0.0072 and 0.0043
5	400	0.0051 and 0.0051	0.0019 and 0.0017	0.0015 and 0.0010

in Table 3 shows that the oversmoothing kernel function provides better approximation. Furthermore, we choose another oversmoothing kernel function $\Phi_{1,2} \in C^4(R^1)$ which generates a RKHS equivalent to $H^3(R^1)$, but the approximate solution does not improve significantly. This observation consists with our theoretical prediction in Theorem 4.2 where convergence of the proposed algorithm depends on the regularity of the target function f^* instead of the kernel function smoothness.

6.3. *Multiscale SVR approximation algorithm for noisy observations*

Finally, we test our proposed Algorithm 1 for the noisy observation data. Choosing the same target functions in Section 6.1, we add a uniform distributed noise on the observation data with a uniform noise level $\delta = 0.02$. Cut-off parameters in the algorithm are chosen as $\epsilon_k = \delta$ and regularization parameters satisfy $\gamma_k = 0.1h_k$ $(1D)$ and $\gamma_k = 0.1h_k^2$ $(2D)$, respectively.

For comparison, we implement the multiscale least squares method in [17] with the l^2 loss function. Note that in [17], there is no specific discussion on the choices of regularization parameters γ_k for noisy data but with a suggested bound $\gamma_k \leq \kappa(h_k/\eta_k)^\tau$ of a constant level. In current subsection, we fix the regularization parameters $\gamma_k = 1$ for the algorithm in [17].

Absolute errors for our proposed multiscale SVR approximation algorithm and the one in [17] are plotted in Figure 5. The quantity information including associated l^2 and l^∞ errors for both methods are presented in Table 4. These results show that the multiscale SVR approximation algorithm outperforms the standard one in [17] in these two examples. The payment for the better approximation is that we have to realize the minimization

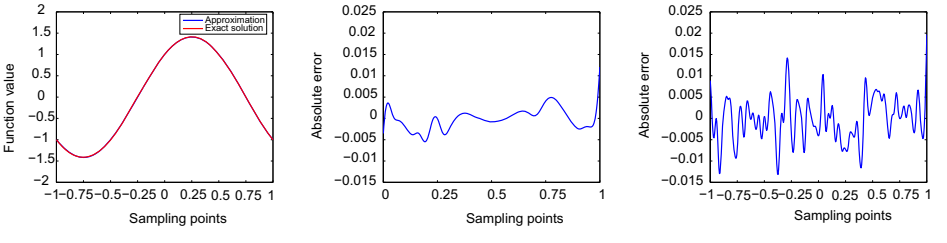


Figure 5. (Noisy observation) 1-D example, the comparison between f_Σ^δ and f^* (left), the absolute error for multiscale SVR approximation (middle) and multiscale interpolation approximation in [17] (right).

Table 4. (Noisy observation) Quantity information at each level for noisy observation. With comparison to the multiscale algorithm in [17].

level	N_k	h_k	l^2 error	l^∞ error	l^2 error ([17])	l^∞ error ([17])
1	25	0.08	0.276960	0.067304	4.355177	0.582446
2	50	0.04	0.169420	0.041949	0.541138	0.156109
3	100	0.02	0.126160	0.027322	0.142034	0.046003
4	200	0.01	0.067832	0.012057	0.085910	0.019722
5	400	0.005	0.049519	0.012057	0.100854	0.019679

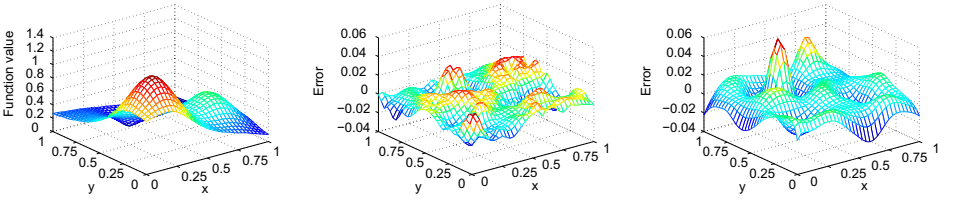


Figure 6. (Noisy observation) Multiscale SVR approximation f_5^δ (left). Error between f_5^δ and f^* (Middle). Error between single mesh SVR approximation algorithm and f^* (Right).

problem (15) at each step which consumes more computational costs compared with the standard one in [17]. Both the multiscale SVR approximate solution and the single mesh SVR approximate solution for the 2D example are presented in Figure 6, where we can observe the efficiency of our proposed algorithm.

Funding

This work was supported by National Natural Science Foundation of China [grant number 11101093]; Chinese Ministry of Education [grant number 20110071120001].

References

[1] Franke C, Schaback R. Solving partial differential equations by collocation using radial basis functions. Appl. Math. Comput. 1998;93:73–82.
[2] Hon YC, Schaback R. Solvability of partial differential equations by meshless kernel methods. Adv. Comput. Math. 2008;28:283–299.
[3] Takeuchi T, Hon YC. Discretized Tikhonov regularization by reproducing kernel Hilbert space for backward heat conduction problem. Adv. Comput. Math. 2011;34:167–183.
[4] Kansa EJ. Multiquadrics-a scattered data approximation scheme with applications to computational fluid dynamics-I. Surface approximations and partial derivative estimates. Comput. Math. Appl. 1990;19:127–145.
[5] Krebs J. Support vector regression for the solution of linear integral equations. Inverse Probl. 2011;27:065007. 23 p.
[6] Krebs J, Louis AK, Wendland H. Sobolev error estimates and a priori parameter selection for semi-discrete Tikhonov regularization. J. Inverse Ill-Posed Probl. 2009;17:845–869.

[7] Li M, Chen CS, Hon YC. A meshless method for solving nonhomogeneous Cauchy problems. Eng. Anal. Bound. Elem. 2011;35:499–506.
[8] Li M, Jiang T, Hon YC. A meshless method based on RBFs method for nonhomogeneous backward heat conduction problem. Eng. Anal. Bound. Elem. 2010;34:785–792.
[9] Takeuchi T, Yamamoto M. Tikhonov regularization by a reproducing kernel Hilbert space for the Cauchy problem for an elliptic equation. SIAM J. Sci. Comput. 2008;31:112–142.
[10] Wendland H. Scattered data approximation. Cambridge: Cambridge University Press; 2005.
[11] Schaback R. Error estimates and condition number for radial basis function interpolation. Adv. Comput. Math. 1995;3:251–264.
[12] Schaback R. Creating surfaces from scattered data using radial basis functions. In: Daehlen M, Lyche T, Schumaker LL, editors. Mathematical methods for curves and surfaces. Nashville: Vanderbilt University Press; 1995. p. 477–496.
[13] Wendland H. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Adv. Comput. Math. 1995;4:389–396.
[14] Wu Z. Compactly supported positive definite radial functions. Adv. Comput. Math. 1995;4: 283–292.
[15] Floater MS, Iske A. Multistep scattered data interpolation using compactly supported radial basis functions. J. Comput. Appl. Math. 1996;73:65–78.
[16] Schaback R. On the efficiency of interpolation by radial basis functions. In: Le Méhauté A, Rabut C, Schumaker LL, editors. Surface fitting and multiresolution methods. Nashville: Vanderbilt University Press; 1997. p. 309–318.
[17] Wendland H. Multiscale analysis in Sobolev space on bounded domains. Numer. Math. 2010;116:493–517.
[18] Le Gia QT, Sloan I, Wendland H. Multiscale analysis in Sobolev spaces on the sphere. SIAM J. Numer. Anal. 2010;48:2065–2090.
[19] Le Gia QT, Sloan I, Wendland H. Multiscale approximation for functions in arbitrary Sobolev spaces by scaled radial basis functions on the unit sphere. Appl. Comput. Harmon. Anal. 2012;32:401–412.
[20] Le Gia QT, Sloan I, Wendland H. Multiscale RBF collocation for solving PDEs on spheres. Numer. Math. 2012;121:99–125.
[21] Chernih A, Le Gia QT. Multiscale methods with compactly supported radial basis functions for Galerkin approximation of elliptic PDEs. IMA J. Numer. Anal. 2014;34:569–591.
[22] Chernih A, Le Gia QT. Multiscale methods with compactly supported radial basis functions for the Stokes problem on bounded domains. ANZIAM J. Electron. Suppl. 2012;54:C137–C152.
[23] Townsend A, Wendland H. Multiscale analysis in Sobolev spaces on bounded domains with zero boundary values. IMA J. Numer. Anal. 2013;33:1095–1114.
[24] Zhong M, Hon YC, Lu S. Multiscale analysis for ill-posed problem with support vector approach. Forthcoming.
[25] Zhong M, Lu S, Cheng J. Multiscale analysis for ill-posed problems with semi-discrete Tikhonov regularization. Inverse Probl. 2012;28:065019. 19 p.
[26] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. New York (NY): ACM; 1992. p. 144–152.
[27] Cortes C, Vapnik V. Support-vector networks. Mach. Learn. 1995;20:273–297.
[28] Vapnik V. Statistical learning theory. New York (NY): Wiley; 1998.
[29] Vapnik V. The nature of statistical learning theory. 2nd ed. New York (NY): Springer; 2000.
[30] Christianini N, Shawe-Taylor J. Support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
[31] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York (NY): Springer; 2009.
[32] Schölkopf B, Smola AJ. Learning with kernels. Cambridge: MIT Press; 2002.

[33] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat. Comput. 2004;14:199–222.
[34] Rieger C, Zwicknagel B. Deterministic error analysis of support vector regression and related regularized kernel methods. J. Mach. Learn. Res. 2009;10:2115–2132.
[35] Wu Z, Schaback R. Local error estimates for radial basis function interpolation of scattered data. IMA J. Numer. Anal. 1993;13:13–27.
[36] Madych WR, Nelson SA. Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation. Numer. Math. 2005;101:643–662.
[37] Wendland H. Error estimates for interpolation by compactly supported radial basis functions of minimal degree. J. Approx. Theory. 1998;93:258–272.
[38] De Vito E, Rosasco L, Caponnetto A, Giovannini UD, Odone F. Learning from examples as an inverse problem. J. Mach. Learn. Res. 2005;6:883–904.
[39] Poggio T, Smale S. The mathematics of learning: dealing with data. Not. Am. Math. Soc. 2003;50:537–544.
[40] Micchelli CA, Pontil M. Learning the kernel function via regulation. J. Mach. Learn. Res. 2005;6:1099–1125.
[41] Schölkopf B, Wiliamson RC, Bartlett PL. New support vector algorithms. Neural Comput. 2000;12:1207–1245.
[42] Stein EM. Singular integrals and differentiability properties of functions. Priceton: Priceton University Press; 1971.
[43] Wendland H, Rieger C. Approximate interpolation with applications to selecting smoothing parameters. Numer. Math. 2005;101:643–662.