SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

March 2012 Vol. 55 No. 3: 701–713
doi: 10.1007/s11432-011-4314-8

Estimation of convergence rate for multi-regression learning algorithm

XU ZongBen[1,2], ZHANG YongQuan[3,1,2] & CAO FeiLong[3*]

[1]*Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an* 710049*, China;*
[2]*MOE Key Labratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an* 710049*, China;*
[3]*Department of Information and Mathematics Sciences, China Jiliang University, Hangzhou* 310018*, China*

Received December 4, 2009; accepted May 10, 2010; published online August 11, 2011

Abstract In many applications, the pre-information on regression function is always unknown. Therefore, it is necessary to learn regression function by means of some valid tools. In this paper we investigate the regression problem in learning theory, i.e., convergence rate of regression learning algorithm with least square schemes in multi-dimensional polynomial space. Our main aim is to analyze the generalization error for multi-regression problems in learning theory. By using the famous Jackson operators in approximation theory, covering number, entropy number and relative probability inequalities, we obtain the estimates of upper and lower bounds for the convergence rate of learning algorithm. In particular, it is shown that for multi-variable smooth regression functions, the estimates are able to achieve almost optimal rate of convergence except for a logarithmic factor. Our results are significant for the research of convergence, stability and complexity of regression learning algorithm.

Keywords learning theory, covering number, rate of convergence, entropy number

Citation Xu Z B, Zhang Y Q, Cao F L. Estimation of convergence rate for multi-regression learning algorithm. Sci China Inf Sci, 2012, 55: 701–713, doi: 10.1007/s11432-011-4314-8

1 Introduction

It is well known that regression and classification are two important basic issues in learning theory. Regression is a hot problem in the research of learning theory, and the convergence of regression is one of the core problems. As early as in 1998, Vapnik [1] systematically investigated the convergence of classification and regression algorithm in statistical learning theory via VC-dimension. Meanwhile, he studied the consistency of Structure Risk Minimization principle algorithm and obtained a series of relative results of convergence. In [2], Shawe-Taylor et al. showed the generalization performance for the Structure Risk Minimization algorithm via VC-dimension. In 2001, Cuker and Smale [3] gave the basic framework of learning theory from mathematical view and pointed out that regression was one of the most important problems in learning theory. They then studied the approximation questions in learning theory by using the method of covering number. In 2002, Cuker and Smale [4] engaged in how to construct the approximations of regression functions and chose regularization parameters to get a better order. In 2006, using the classical Bernstein's inequality in probability theory, Wu et al. [5] studied the convergence

[*] Corresponding author (email: feilongcao@gmail.com)

for least squares regularized regression learning algorithm in reproducing kernel Hilbert space. In 2007, Smale and Zhou [6] investigated the approximation problems of regression learning algorithm by using the integral operators of reproducing kernels. Caponnetto and DeVito [7] showed the generalization performance for the least squares regularized regression algorithm in reproducing kernel Hilbert space. In particular, they established the estimate of convergence rate for the vector-valued objective functions. Temlyakov [8] made use of the methods and techniques in function approximation theory to study the approximation problems in learning algorithm. For research on the convergence rate please refer to [9–11].

These estimations cannot, however, completely characterize the convergence capability of learning algorithms in general, because an established upper estimation might be too loose to reflect their inherent convergence capability. Hence, in order to present the inherent convergence rates of learning algorithm accurately, we need to estimate not only the upper bound for convergence rate of learning algorithm but also the lower bound. Naturally, the estimate for lower bound is difficult but significant. In [12], Temlyakov studied the convergence rate of learning algorithm by the theory of optimal approximation in approximation theory and showed the estimate of lower bound for the convergence rate under some assumptions of regression functions.

A series of research referred above focuses on the convergence performance and rate of learning algorithm, especially the estimates of upper and lower bounds for the convergence rate of learning algorithm which have important effects on the studies of performance, stability and complexity of learning algorithm.

However, for the general integral operators corresponding to kernels, calculations of their feature values are very difficult. Moreover, we know that the covering number is widely used as the measurement of complexity [13–17]. Hence, in this paper we first establish the upper bound for the convergence rate using the covering number of reproducing kernel Hilbert space as the measurement of complexity. Then we introduce the entropy number and show the lower bound of learning rate by using this tool. The upper and lower bounds that we have obtained have the same order except for a logarithmic factor.

2 Regression learning algorithm and upper bound for convergence rate

Let $\mathcal{R}^s$ be s-dimensional Euclidean space and $[-1,1]^s$ be the cube in $\mathcal{R}^s$ where s is an integer. In this paper, we consider a function set consisting of polynomial functions on $X=[-1,1]^s$ as a hypothesis space, over which we minimize a least squares risk. In regression analysis, an $\mathcal{R}^s \times \mathcal{R}$-valued random vector $(\mathcal{X},\mathcal{Y})$ with $\boldsymbol{E}\mathcal{Y}^2<\infty$ is considered and the dependency of $\mathcal{Y}$ on the value of $\mathcal{X}$ is of interest. The goal is to find a function $f:\mathcal{R}^s\to\mathcal{R}$ such that $f(\mathcal{X})$ is a good approximation of $\mathcal{Y}$. In the sequel, the main aim of the analysis is to minimize the mean squared prediction error or L_2 risk:

$$\mathcal{E}(f)=\boldsymbol{E}\{|f(\boldsymbol{x})-y|^2\}.$$

The function that minimizes the error is called the regression function, given by

$$m(\boldsymbol{x})=\boldsymbol{E}\{\mathcal{Y}|\mathcal{X}=\boldsymbol{x}\},\quad \boldsymbol{x}\in\mathcal{R}^s.$$

Indeed, let $f:\mathcal{R}^s\to\mathcal{R}$ be an arbitrary measurable function on $\mathcal{R}^s$. We denote the distribution of X by ν. The well-known relation [9,18]

$$\boldsymbol{E}\{|f(\boldsymbol{x})-y|^2\}=\boldsymbol{E}\{|m(\boldsymbol{x})-y|^2\}+\int_{\mathcal{R}^s}(f(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(d\boldsymbol{x})$$

implies that the regression function is the optimal predictor in view of minimization of the L_2 risk:

$$\boldsymbol{E}\{|m(\boldsymbol{x})-y|^2\}=\min_{f:\ \mathcal{R}^s\to\mathcal{R}}\boldsymbol{E}\{|f(\boldsymbol{x})-y|^2\}.$$

In addition, any measurable function f is a good predictor with its L_2 risk close to the optimal value, if and only if

$$\mathcal{E}(f)=\boldsymbol{E}\{|f(\boldsymbol{x})-y|^2\}\tag{1}$$

is small. This motivates us to measure the error caused by using the function f instead of the regression function by error (1).

In many applications, the distribution of the sample is usually unknown. Hence the regression function is unknown. But often it is possible to observe a sample chosen according to the distribution. This leads to the regression estimation problem. Let $\boldsymbol{z} = \{z_i\}_{i=1}^n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be independent and identically distributed random vectors drawn on $X \times Y$. Our goal is to construct an estimate

$$f_{\boldsymbol{z}}(\cdot) = f(\cdot, \boldsymbol{z})$$

of the regression function such that the L_2 error

$$\int_X (f_{\boldsymbol{z}}(\boldsymbol{x}) - m(\boldsymbol{x}))^2 \nu(d\boldsymbol{x})$$

is small.

Throughout this paper, we assume that $|m(\boldsymbol{x})| \leqslant M/4$ for some $M \in \mathcal{R}_+$. Here it is necessary to impose smoothness conditions on the regression function. Since learning processes do not take place in a vacuum, and some structure needs to be at the beginning of the process, this structure (which is called hypothesis space) usually takes the forms of functions (e.g., a space of polynomials, continuous function space, etc.). A familiar hypothesis space is polynomial function space, which has been used in [18, 19]. The goal of learning process will thus be to find the best approximation of the regression function $m(\boldsymbol{x})$ within the hypothesis space.

In the sequel, we will introduce the polynomial functions [20] on $X = [-1, 1]^s$. Let $\mathcal{H}_d$ be the space of all functions

$$f : \mathcal{R}^s \to \mathcal{R},\ f(\boldsymbol{x}) = \sum_{0 \leqslant k_1 \leqslant d} \cdots \sum_{0 \leqslant k_s \leqslant d} a_{k_1,\ldots,k_s} a_{k_1,\ldots,k_s}, x_1^{k_1} x_2^{k_2} \cdots x_s^{k_s},\ \boldsymbol{x} \in [-1, 1]^s,$$

where $a_{k_1,\ldots,k_s} \in \mathcal{R}$ for $|k_1| \leqslant d, \ldots, |k_s| \leqslant d$. By the definition of $\mathcal{H}_d$, $\dim \mathcal{H}_d = (2d)^s$.

In the paper, we consider the set $\mathcal{F}_d = \{f \in \mathcal{H}_d : \ |f(\boldsymbol{x})| \leqslant M/4, \boldsymbol{x} \in [-1, 1]^s\}$ as the hypothesis space. The estimate $f_{\boldsymbol{z}}$ is defined by

$$f_{\boldsymbol{z}} = \arg \min_{f \in \mathcal{F}_d} \frac{1}{n} \sum_{i=1}^n (f(\boldsymbol{x}_i) - y_i)^2, \tag{2}$$

where $f(\boldsymbol{x}) = \sum_{0 \leqslant k_1 \leqslant d} \cdots \sum_{0 \leqslant k_s \leqslant d} a_{k_1,\ldots,k_s} x_1^{k_1} x_2^{k_2} \cdots x_s^{k_s}, \boldsymbol{x} \in [-1, 1]^s$.

We will analyze the rate of convergence of this least squares estimate $f_{\boldsymbol{z}}$. The efficiency of the algorithm (2) is measured by the difference between $f_{\boldsymbol{z}}$ and the regression function $m(\boldsymbol{x})$. According to the definition of $m(\boldsymbol{x})$, we know

$$\int_X (f_{\boldsymbol{z}}(\boldsymbol{x}) - m(\boldsymbol{x}))^2 \nu(d\boldsymbol{x}) = \mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m).$$

Our goal is to estimate the above error for algorithm (2) by means of properties of μ and the functions set $\mathcal{F}_d$.

Set the empirical error at

$$\mathcal{E}_{\boldsymbol{z}}(f) = \frac{1}{n} \sum_{i=1}^n (f(\boldsymbol{x}_i) - y_i)^2.$$

It is a discretization of the error $\mathcal{E}(f)$. Therefore, $f_{\boldsymbol{z}}$ can also be written as

$$f_{\boldsymbol{z}} = \arg \min_{f \in \mathcal{F}_d} \mathcal{E}_{\boldsymbol{z}}(f).$$

Our first theorem gives an upper bound for the expected L_2 error of our estimate.

Theorem 1. Let $\mathcal{F}_d$ be a hypothesis space. Then, for the estimate $f_{\boldsymbol{z}}$ defined by (2), we have

$$\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m) \leqslant \frac{204M^2}{n} \log \frac{2}{\delta} + \frac{64M^2(2d)^s \log(4M^2 n)}{n} + 3(\mathcal{E}(Q_d(m)) - \mathcal{E}(m))$$

with confidence at least $1-\delta$.

Together with the approximation result this theorem implies the next corollary, which considers the rate of convergence of the estimate. Here it is necessary to impose smoothness condition on the regression function. Denote all continuous functions on the set $X=[-1,1]^s$ by $C_{[-1,1]^s}$. For $f\in C_{[-1,1]^s}$ and positive integer r, define the difference of f:

$$\Delta_{\boldsymbol{t}}^1 f(\boldsymbol{x})=f(\boldsymbol{x}+\boldsymbol{t})-f(\boldsymbol{x}),\ \ \Delta_{\boldsymbol{t}}^r f(\boldsymbol{x})=\Delta_{\boldsymbol{t}}\Delta_{\boldsymbol{t}}^{r-1}f(\boldsymbol{x}).$$

It is well known that

$$\Delta_{\boldsymbol{t}}^r f(\boldsymbol{x})=\sum_{j=0}^{r}(-1)^j\binom{r}{j}f(\boldsymbol{x}+j\boldsymbol{t}).$$

Let $\|\boldsymbol{t}\|_2=(t_1^2+\cdots+t_s^2)^{1/2}$, the usual Euclidean norm of $(t_1,\ldots,t_s)$. The modulus of smoothness of a continuous function f is then defined as

$$\omega_r(f,h)=\sup_{\|\boldsymbol{t}\|_2\leqslant h}\|\Delta_{\boldsymbol{t}}f\|_\infty,$$

where $\|f\|_\infty=\max_{\boldsymbol{x}\in[-1,1]^s}|f(x)|$.

The modulus of smoothness is a tool used in approximation theory, and it is also usually used to describe smoothness of functions and approximation error and has the following properties:

Proposition 1 [21]**.** There hold the following inferences.

1) For $\lambda>0$, $\omega_r(f,\lambda h)\leqslant(1+\lambda)^r\omega_r(f,h)$.

2) Let ∂_i denote the partial derivative with respect to t_i and $\partial^k=\partial_1^{k_1}\cdots\partial_s^{k_s}$. Then

$$\omega_r(f,h)\leqslant h^r\sum_{k_1+\cdots+k_s=r}\frac{r!}{k_1!\cdots k_s!}\|\partial^k f\|_\infty.$$

In the sequel, we give the definition of partial derivative of the function that belongs to Lipschitz class.

Definition 1. Suppose k is a natural number, $0<\beta\leqslant 1$ and $C>0$. Let $\alpha=(\alpha_1,\ldots,\alpha_s)$, $\alpha_i\in\mathbb{N}$, $\sum_{j=1}^s\alpha_j=k$. Let $f:[-1,1]^s\to\mathcal{R}$. Then we have partial derivative:

$$\frac{\partial^k f}{\partial x_1^{\alpha_1}\cdots\partial x_s^{\alpha_s}}.$$

If for all $\boldsymbol{x},\boldsymbol{z}\in[-1,1]^s$, there exists a constant $C>0$ such that

$$\left|\frac{\partial^k f(\boldsymbol{x})}{\partial x_1^{\alpha_1}\cdots\partial x_s^{\alpha_s}}-\frac{\partial^k f(\boldsymbol{z})}{\partial z_1^{\alpha_1}\cdots\partial z_s^{\alpha_s}}\right|\leqslant C\|\boldsymbol{x}-\boldsymbol{z}\|^\beta,$$

then we say that the kth order of partial derivative of function f belongs to the class of $\mathrm{Lip}_C\beta$, i.e., $\partial^k f\in\mathrm{Lip}_C\beta$.

From the above definition, we obtain the following corollary based on Theorem 1.

Corollary 1. Let $\mathcal{F}_d$ be a hypothesis space. Suppose k is a natural number, $0<\beta\leqslant 1$. If the kth order of partial derivative of regression function $m(\boldsymbol{x})$ belongs to $\mathrm{Lip}_C 1$, and $d=[n^{1/(k+1)}]$, then we have

$$\boldsymbol{E}\int_X(f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(d\boldsymbol{x})\leqslant\frac{408M^2\log 2}{n}+2C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}},$$

where $C_{K,s}=2(108M^2 2^s)^{\frac{2k}{2k+s}}+2(3C_K^{'})^{\frac{s}{2k+s}}$, $[a]$ denotes the integer part of real number a.

3 Estimation of lower bound for convergence rate of learning algorithm

In this section, we will present the estimate of lower bound for convergence rate of learning algorithm, and show that the upper bound obtained by Corollary 1 is almost optimal. In order to give the lower bound for convergence rate, we first introduce entropy number of a set.

Definition 2 [12]. Let E be a Banach space, and $F \subset E$ be a bounded set. For $i \geqslant 1$, the ith entropy number $e_i(F, E)$ of F is defined to be the infimum over all $\varepsilon > 0$ such that there exist $x_1, x_2, \ldots, x_{2^{i-1}} \in F$ with

$$F \subset \cup_{j=1}^{2^{i-1}} (x_j + \varepsilon B_E),$$

where B_E denotes the closed unit ball of E.

The following theorem gives the lower bounds of learning rates.

Theorem 2. Let ν be the distribution of X. Θ is a compact subset of $L_2(\nu)$ such that $\Theta \subset \frac{1}{4}U(\mathcal{C}(X))$. Assume that there exists a positive integer k, $c_1, c_2 > 0$ such that

$$c_1 i^{-k/s} \leqslant e_i(\Theta, L_2(\nu)) \leqslant c_2 i^{-k/s}.$$

Then for all algorithms $\mathcal{A}$ defined by (2) there exists a distribution P on $X \times [-M, M]$ satisfying $P_X = \nu$ and $f_\rho \in \Theta$ such that

$$\boldsymbol{E} \int_X (\pi_M(f_{\boldsymbol{z},q})(x) - f_\rho(x))^2 d\rho_X \geqslant C_1 \left(\frac{1}{m}\right)^{\frac{2k}{2k+s}},$$

where C_1 is a constant. The proof of Theorem 2 is based on the following Lemma 1.

Lemma 1 [12]. Let ν be a distribution on X, and $\Theta \subset L_2(\nu)$ such that $\|f\|_\infty \leqslant M/4$ for all $f \in \Theta$ and some $M > 0$. In addition, assume that there exists an $r > 0$ such that

$$e_i(\Theta, L_2(\nu)) \sim i^{-1/r}.$$

Then there exist constants $\delta_0, c_1, c_2 > 0$ and a sequence $\{\varepsilon_m\}$ with

$$\varepsilon_m \sim m^{-\frac{2}{2+r}},$$

such that for all learning methods $\mathcal{A}$ defined by (2) there exists a distribution P on $X \times Y$ satisfying $P_X = \nu$ and $f_\rho \in \Theta$ such that for all $\varepsilon > 0$ and $m \geqslant 1$:

$$P^m(\boldsymbol{z} : \mathcal{E}(\pi_M(f_{\boldsymbol{z},q})) - \mathcal{E}(f_\rho) \geqslant \varepsilon) \geqslant \begin{cases} \delta_0, & \text{if } \varepsilon < \varepsilon_m, \\ c_1 e^{-c_2 \varepsilon m}, & \text{if } \varepsilon \geqslant \varepsilon_m, \end{cases}$$

where $f_{\boldsymbol{z}}$ is the the decision function produced by $\mathcal{A}$ for a given training set D.

Our next goal is to apply Lemma 1 to the proof of Theorem 2.

Proof of Theorem 2. Since the set Θ satisfies

$$c_1 i^{-k/s} \leqslant e_i(\Theta, L_2(\nu)) \leqslant c_2 i^{-k/s},$$

we apply Lemma A3 (see Appendix) with $r = \frac{s}{k}$, and know that there exists a sequence $\{\varepsilon_m\}$ with

$$\varepsilon_m \sim m^{-\frac{2k}{2k+s}},$$

such that for $m \in \Theta$,

$$P^m(\boldsymbol{z} : \mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m) \geqslant \varepsilon) \geqslant \begin{cases} \delta_0, & \text{if } \varepsilon < \varepsilon_m, \\ c_1 e^{-c_2 \varepsilon m}, & \text{if } \varepsilon \geqslant \varepsilon_m. \end{cases}$$

Using the above inequality, we get

$$\boldsymbol{E} \int_X (f_{\boldsymbol{z}}(\boldsymbol{x}) - m(\boldsymbol{x})))^2 d\rho_X = \int_0^\infty P^m(\boldsymbol{z} : \mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m) \geqslant \varepsilon) d\varepsilon$$

$$\geqslant \int_0^{\varepsilon_m} \varepsilon d\varepsilon + c_1 \int_{\varepsilon_m}^{\infty} e^{-c_2\varepsilon m} d\varepsilon = \delta_0 \varepsilon_m + \frac{c_1}{mc_2} e^{-c_2 m \varepsilon_m} \geqslant c_1 \left(\frac{1}{m}\right)^{\frac{2k}{2k+s}}.$$

The proof of Theorem 2 is finished.

4 Conclusions

In this paper we have investigated the upper and lower bounds of the convergence rate for least squares regularized learning algorithm on polynomial space. The obtained upper and lower bounds have the same order except for a logarithmic factor, which are significant for the studies of convergence, stability and adaption of regression learning algorithm.

We have introduced and investigated deeply the multi-variate Jackson operators, especially estimated the approximation order of these operators by the rth modulus of smoothness. Meanwhile, combining the covering number in Hilbert space and inequalities in probability theory, we have established the upper bound of convergence rate for the regression learning algorithm. Particularly, we have obtained the better order of convergence when regression functions satisfy some smoothness conditions. In order to get the lower bound for the convergence rate of the learning algorithm, we have obtained the lower bound by entropy number which satisfies some conditions.

It is well known that the performance of an algorithm is frequently determined by such factors as characteristics of convergence and complexity. This paper has not only demonstrated the algorithm studied in convergence, but also given the function relation among algorithm, samples and hypothesis space. The upper and lower bounds obtained in this paper for the generalization error of least squares regularized algorithm are almost optimal. Although [22] got better upper and lower bounds, the complexity of hypothesis space was measured by feature values of integral operators. However, it is hard to calculate feature values of general integral operators. We have obtained the upper bound for learning rate via covering number of reproducing kernels Hilbert space, and given the lower bound by the entropy number of the set. Covering number and entropy number have been widely used as the measurement of complexity in learning theory.

According to Corollary 1 and Theorem 2, when the kth order of partial derivative of regression function $m(\boldsymbol{x})$ belongs to $\mathrm{Lip}_C 1$, it is easy to get the following inequality:

$$c_1 \left(\frac{1}{n}\right)^{\frac{2k}{2k+s}} \leqslant \int_X (f(\boldsymbol{x}) - m(\boldsymbol{x}))^2 \nu(d\boldsymbol{x}) \leqslant c_2 \left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}.$$

The above inequality implies that the obtained estimates of convergence rate are optimal except for a logarithmic factor when the regression functions satisfy some smoothness assumptions. Naturally, we expect that the upper and lower 'orders' are the same. We try to characterize exactly the essential order of convergence of learning algorithm. This is an important problem and deserves further study. We guess that the upper bound can be improved into $c_1(\frac{1}{n})^{\frac{2k}{2k+s}}$ under some conditions.

Acknowledgements

This work was supported by National Basic Research Program of China (Grant No. 2007CB311002) and National Natural Science Foundation of China (Grant Nos. 90818020, 60873206).

References

1 Vapnik V. Statistical Learning Theory. New York: Wiley, 1998

2 Shawe-Taylor J, Bartlett P L, Williamson R C, et al. Structural risk minimization over data-dependent hierarchies. IEEE Trans Inf Theory, 1998, 44: 1926–1940

3 Cucker F, Smale S. On the mathematical foundations of learning. Bull Amer Math Soc, 2001, 39: 1–49

4 Cucker F, Smale S. Best choices for regularization parameters in learning theory: On the biasvariance problem. Found Comput Math, 2002, 1: 413–428

5 Wu Q, Ying Y M, Zhou D X. Learning rates of least-square regularized regression. Found Comput Math, 2006, 6: 171–192

6 Smale S, Zhou D X. Learning theory estimates via intergral operators and their approximation. Constr Approx, 2007, 26: 153–172

7 Caponnetto A, DeVito E. Optimal rates for the regularized least-squares algorithm. Found Comput Math, 2007, 7: 331–368

8 Temlyakov V N. Approximation in learning theory. IMI Preprints, 2005, 5: 1–42

9 Cucker F, Zhou D X. Learning Theory: An Approximation Theory Viewpoint. New York: Cambridge University Press, 2007

10 Tong H Z, Chen D R, Li Z P. Learning rates for regularized classifiers using multivariate polynomial kernels. J Complexity, 2008, 24: 619–631

11 Zhou D X, Jetter K. Approximation with polynomial kernels and SVM classifiers. Adv Comput Math, 2006, 25: 323–344

12 Temlyakov V N. Optimal estimators in learning theory. Inst Math Pol Acad Sci, 2006, 72: 341–366

13 Chen D R, Wu Q, Ying Y, et al. Support vector machine soft margin classifiers: error analysis. J Mach Learn Res, 2004, 5: 1143–1175

14 Guo Y, Bartlett P L, Shawe-Taylor J, et al. Covering numbers for support vector machines. IEEE Trans Inf Theory, 2002, 48: 239–250

15 Pontil M. A note different covering numbers in learning theory. J Complexity, 2003, 19: 665–671

16 Zhou D X. The covering number in learning theory. J Complexity, 2002, 18: 739–767

17 Zhou D X. Capacity of reproducing kernel spaces in learning theory. IEEE Trans Inf Theory, 2003, 49: 1734–1752

18 Smale S, Zhou D X. Estimating the approximation error in learning theory. Anal Appl, 2003, 1: 17–41

19 Wu Q, Ying Y M, Zhou D X. Learning theory: from regression to classification. In: Topics in Multivariate Approximation and Interpolation. Amsterdam: Elsevier B.V., 2004

20 Xie T F, Zhou S P. Real Function Approximation. Hangzhou: Hangzhou University Press, 1998

21 Xu Y. Fourier series and approximation on hexagonal and triangular domains. Constr Approx, 2009, 31: 115–138

22 Steinwart I, Hush D, Scovel C. Optimal rates for regularized least squares regression. Los Alamos National Laboratory Technical Report LA-UR-09-00901. 2009

Appendix

1 Approximation of Jackson operators

It is well known that the Jackson operator (see [20]) plays an important role in approximation theory. For the natural number d, r, let t $q = [d/r] + 1$, and define a kernel function:

$$K_{dr}(t) = L_{q,r}(t) = \frac{1}{\lambda_{qr}} \left(\frac{\sin \frac{qt}{2}}{\sin \frac{t}{2}} \right)^{2r}, \tag{A1}$$

where $\lambda_{qr} = \int_{-\pi}^{\pi} (\frac{\sin \frac{qt}{2}}{\sin \frac{t}{2}})^{2r} dt$. Jackson kernels have the following properties.

Lemma A1 [20]**.** Let $K_{dr}(t)$ be defined by (A1). Then it is a trigonometric polynomial with order of d, and

$$\int_{-\pi}^{\pi} K_{dr}(t)dt = 1, \quad \int_{-\pi}^{\pi} t^k K_{dr}(t)dt \leqslant C_k(d+1)^{-k}, \ k = 0, 1, \ldots, 2r-2.$$

Let $K_{dr}(\boldsymbol{t}) = K_{dr}(t_1) \cdots K_{dr}(t_s)$, and $|k| = k_1 + \cdots + k_s$. Then we get

$$\int_{[-\pi,\pi]^s} \boldsymbol{t}^k K_{dr}(\boldsymbol{t}) d\boldsymbol{t} \leqslant C'_k (d+1)^{-k_1} \cdots (d+1)^{-k_s} = C'_k (d+1)^{-|k|},$$

where $C'_k = C_{k_1} \cdots C_{k_s}$.

Let $\phi(\boldsymbol{t}) = f(\cos \boldsymbol{t})$. Then $\phi(\boldsymbol{t})$ is a periodic function. We define Jackson operator on $[-\pi, \pi]^s$ as

$$\begin{aligned} J_d(f, \boldsymbol{u}) &= -\int_{[-\pi,\pi]^s} K_{dr}(\boldsymbol{t}) \sum_{j=1}^{k} (-1)^j \binom{k}{j} f(\cos(\boldsymbol{u} + j\boldsymbol{t})) d\boldsymbol{t} \\ &= -\int_{[-\pi,\pi]^s} K_{dr}(\boldsymbol{t}) \sum_{j=1}^{k} (-1)^j \binom{k}{j} \phi(\boldsymbol{u} + j\boldsymbol{t}) d\boldsymbol{t}, \end{aligned} \tag{A2}$$

where r is the minimum positive integer satisfying $r \leqslant [(k+2)/2]$.

Denote by $L^2_{[-1,1]^s}$ the set of Lebesgue square-integrable functions on $[-1,1]^s$. Define $\|f\|_2 = (\int_{[-\pi,\pi]^s} |f(\boldsymbol{t})|^2 d\boldsymbol{t})^{1/2} < \infty$ for $f \in L^2_{[-1,1]^s}$, where $L^2_{[-1,1]^s}$ is a Banach space and denote the inner product on $L^2_{[-1,1]^s}$ by $\langle f,g\rangle = \int_{[-\pi,\pi]^s} f(\boldsymbol{t})\overline{g(\boldsymbol{t})}d\boldsymbol{t}$. Then $L^2_{[-1,1]^s}$ is a Hilbert space. Furthermore, we know that $\{e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\}_{\boldsymbol{k}\in\mathbb{Z}^s}$ is the orthogonal basis of Hilbert space $L^2_{[-1,1]^s}$ with respect to the above inner product. Therefore, for any $\psi \in L^2_{[-\pi,\pi]^s}$, we have

$$\psi(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^s} a_{\boldsymbol{k}}(\psi)e^{i\boldsymbol{k}\cdot\boldsymbol{x}},$$

where $a_{\boldsymbol{k}}(\psi) = \int_{[-\pi,\pi]^s} \psi(\boldsymbol{t})e^{i\boldsymbol{k}\cdot\boldsymbol{t}}d\boldsymbol{t}$, and $\mathbb{Z}^s$ denotes s-repetition exponent index set.

The vector $\boldsymbol{l}$ is divided exactly by integer j if and only if every component of the vector $\boldsymbol{l}$ can be divided by j exactly.

Lemma A2. Let $\phi \in L^2_{[-\pi,\pi]^s}$, $\boldsymbol{l} = (l_1,\ldots,l_s)$, and let l_j be the positive integer $(j = 1,2,\ldots)$. When the vector $\boldsymbol{l}$ is not divided exactly by integer j, we have

$$\int_{[-\pi,\pi]^s} \phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}d\boldsymbol{t} = 0.$$

Proof. Since the function $\phi(j\boldsymbol{t})$ has period $2\pi/j$ on each variable, we have

$$\begin{aligned}
\int_{[-\pi,\pi]^s} \phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}d\boldsymbol{t} &= \int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi} \phi(jt_1,\ldots,jt_s)e^{i(l_1t_1+\cdots+l_st_s)}dt_1\cdots dt_s \\
&= \int_{-\pi+2il_1\pi/j}^{\pi+2il_1\pi/j}\cdots\int_{-\pi}^{\pi} \phi(jt_1,\ldots,jt_s)e^{i(l_1t_1+\cdots+l_st_s)}dt_1\cdots dt_s \\
&= \int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi} \phi(jt_1,\ldots,jt_s)e^{i((l_1t_1+2il_1\pi/j)+\cdots+l_st_s)}dt_1\cdots dt_s \\
&= e^{2il_1\pi/j}\int_{[-\pi,\pi]}\cdots\int_{[-\pi,\pi]} \phi(jt_1,\ldots,jt_s)e^{i(l_1t_1+\cdots+l_st_s)}dt_1\cdots dt_s \\
&= e^{2il_1\pi/j}\int_{[-\pi,\pi]^s} \phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}d\boldsymbol{t}.
\end{aligned}$$

When l_1 cannot be divided by j exactly, there holds

$$\int_{[-\pi,\pi]^s} \phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}d\boldsymbol{t} = 0.$$

Similarly, we can prove that other component l_i of $\boldsymbol{l}$ cannot be divided by j exactly, $i = 2,3,\ldots,s$. The above integral is 0. Therefore, when $\boldsymbol{l}$ cannot be divided exactly by j, we have

$$\int_{[-\pi,\pi]^s} \phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}d\boldsymbol{t} = 0.$$

This finishes the proof of Lemma A2.

From Lemma A2 and the fact that

$$\begin{aligned}
\int_{[-\pi,\pi]^s} \phi(\boldsymbol{u}+j\boldsymbol{t})\cos\boldsymbol{l}\cdot\boldsymbol{t}d\boldsymbol{t} &= \int_{[-\pi,\pi]^s} \phi(j\boldsymbol{t})\cos\boldsymbol{l}\cdot(\boldsymbol{t}-\frac{\boldsymbol{u}}{j})d\boldsymbol{t} \\
&= \frac{1}{2}\int_{[-\pi,\pi]^s} \phi(j\boldsymbol{t})(e^{i\boldsymbol{l}\cdot(\boldsymbol{t}-\frac{\boldsymbol{u}}{j})} + e^{-i\boldsymbol{l}\cdot(\boldsymbol{t}-\frac{\boldsymbol{u}}{j})})d\boldsymbol{t} \\
&= \frac{1}{2}\int_{[-\pi,\pi]^s} \phi(j\boldsymbol{t})(e^{i\boldsymbol{l}\cdot\boldsymbol{t}}e^{-i\boldsymbol{l}\cdot\frac{\boldsymbol{u}}{j}} + e^{-i\boldsymbol{l}\cdot\boldsymbol{t}}e^{i\boldsymbol{l}\cdot\frac{\boldsymbol{u}}{j}})d\boldsymbol{t}, \qquad \text{(A3)}
\end{aligned}$$

where $\cos\boldsymbol{l}\cdot\boldsymbol{t} = \cos l_1t_1\cdots\cos l_st_s$, it follows that if $\boldsymbol{l}$ is not divided exactly by j, (A3) equals to 0. Otherwise, (5) is a trigonometric polynomial in $\mathcal{H}_d$ for $1 \leqslant j \leqslant k$, $0 \leqslant l_i \leqslant d(i = 1,2,\ldots,d)$. From the above discussion, we know K_{dr} is a trigonometric polynomial with an order of d. Therefore, $J_d(f,\boldsymbol{u})$ is linear combination of $\int_{-\pi}^{\pi} f(\boldsymbol{u}+j\boldsymbol{t})\cos\boldsymbol{l}\cdot\boldsymbol{t}d\boldsymbol{t}$, i.e., $J_d(f,\boldsymbol{u})$ is a trigonometric polynomial. Let $\boldsymbol{u} = \arccos\boldsymbol{x} = \arccos x_1\cdots\arccos x_s$. Then $Q_d(f,\boldsymbol{x}) = J_d(f,\arccos\boldsymbol{x})$ is an algebraic polynomial with order of $2d-2$.

Proposition A1. Let k be a natural number, $f \in C_{[-1,1]^s}$. For $d = 0,1,\ldots$, there holds

$$|f(\boldsymbol{x}) - Q_d(f,\boldsymbol{x})| \leqslant C_{ks}\omega_k\left(f,\frac{1}{d+1}\right), \qquad \forall\boldsymbol{x}\in[-1,1]^s.$$

Proof. From Lemma A1, the definition of $Q_d(f,\boldsymbol{x})$ and the fact $K_{dr}(\boldsymbol{t}) = K_{dr}(-\boldsymbol{t})$, we have

$$|f(\boldsymbol{x}) - Q_d(f,\boldsymbol{x})| = |f(\cos\boldsymbol{u}) - J_d(f,\boldsymbol{u})| = \left|\int_{[-\pi,\pi]^s} K_{dr}(\boldsymbol{t})\triangle^k_{\|\boldsymbol{t}\|_2}\phi(\boldsymbol{u})d\boldsymbol{t}\right|$$
$$\leqslant \int_{[-\pi,\pi]^s} K_{dr}(\boldsymbol{t})|\triangle^k_{\|\boldsymbol{t}\|_2}\phi(\boldsymbol{u})|d\boldsymbol{t} \leqslant 2^s\int_{[0,\pi]^s} K_{dr}(\boldsymbol{t})\omega_k(\phi,\|\boldsymbol{t}\|_2)d\boldsymbol{t},$$

where $\phi(\boldsymbol{u}) = f(\cos\boldsymbol{u})$.

For any $t,t' \in \mathcal{R}$, we have $|\cos y - \cos y'| \leqslant |y - y'|$. It follows that

$$\sup_{\|\boldsymbol{t}-\boldsymbol{t}'\|_2\leqslant h} |\phi(\boldsymbol{t}) - \phi(\boldsymbol{t}')| \leqslant \sup_{\|\cos\boldsymbol{t}-\cos\boldsymbol{t}'\|_2\leqslant h} |\phi(\boldsymbol{t}) - \phi(\boldsymbol{t}')| = \sup_{\|\boldsymbol{x}-\boldsymbol{x}'\|_2\leqslant h} |f(\boldsymbol{x}) - f(\boldsymbol{x}')|,$$

i.e., $\omega_k(\phi,\|\boldsymbol{t}\|_2) \leqslant \omega_k(f,\|\boldsymbol{t}\|_2)$.

From the definition of the modulus of smoothness of f, we know

$$\omega_k(f,\|\boldsymbol{t}\|_2) \leqslant (1+(d+1)\|\boldsymbol{t}\|_2)^k\omega_k\left(f,\frac{1}{d+1}\right).$$

For $n = 0,1,\ldots,k \leqslant 2r-2$ and any $\boldsymbol{x} \in [-1,1]^s$, applying Lemma A1, we get

$$|f(\boldsymbol{x}) - Q_d(f,\boldsymbol{x})| \leqslant 2^s\omega_k\left(f,\frac{1}{d+1}\right)\int_{[0,\pi]^s}(1+(d+1)\|\mathbf{t}\|_2)^kK_{dr}(\boldsymbol{t})d\boldsymbol{t}$$
$$\leqslant C_{ks}\omega_k\left(f,\frac{1}{d+1}\right),$$

where $C_{ks} = 2^sC'_k$.

The proof of Proposition A1 is completed.

2 Proof of Theorem 1

To estimate the error $\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m)$, we need to estimate

$$\begin{aligned}\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m) =;&\{(\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m)) - (\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}}) - \mathcal{E}_{\boldsymbol{z}}(m))\} + \{\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}}) - \mathcal{E}_{\boldsymbol{z}}(Q_d(m))\}\\ &+ \{(\mathcal{E}_{\boldsymbol{z}}(Q_d(m)) - \mathcal{E}_{\boldsymbol{z}}(m)) - (\mathcal{E}(Q_d(m)) - \mathcal{E}(m))\} + (\mathcal{E}(Q_d(m)) - \mathcal{E}(m))\\ \leqslant&\ \{(\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m)) - (\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}}) - \mathcal{E}_{\boldsymbol{z}}(m))\} + (\mathcal{E}(Q_d(m)) - \mathcal{E}(m))\\ &+ \{(\mathcal{E}_{\boldsymbol{z}}(Q_d(m)) - \mathcal{E}_{\boldsymbol{z}}(m)) - (\mathcal{E}(Q_d(m)) - \mathcal{E}(m))\}.\end{aligned} \tag{A4}$$

We first estimate the third term in (A4), we need the following lemma for the random variable $\xi = (Q_d(m,\boldsymbol{x}) - y)^2 - (m(\boldsymbol{x}) - y)^2$.

Lemma A3 [13]**.** Let ξ be a random variable on $Z = [-1,1]^s \times [-M,M]$ with mean μ and variance σ^2. Assume that $\mu \geqslant 0, |\xi - \mu| \leqslant B$ almost everywhere, and $E(\xi^2) \leqslant c_\xi E\xi$, then for every $\varepsilon > 0$, there holds

$$\mathrm{Prob}_{\boldsymbol{z}\in Z^n}\left\{\frac{\mu - \frac{1}{n}\sum_{i=1}^n\xi(z_i)}{\sqrt{\mu+\varepsilon}} \geqslant \sqrt{\varepsilon}\right\} \leqslant \exp\left\{-\frac{n\varepsilon}{2c_\xi + \frac{2}{3}B}\right\}.$$

In (A4), since $\xi = (f_{\boldsymbol{z}}(\mathrm{x}) - y)^2 - (m(\boldsymbol{x}) - y)^2$ is not a single random variable on $[-1,1]^s \times [-M,M]$, it depends on the sample $\boldsymbol{z}$. The variable ξ changing with the sample runs over the function set $\mathcal{F}_d$, and should not be considered as a fixed function. In the following, we shall bound the first part of (A4) by using the covering number of the unit ball.

Theorem A1. For any $0 < \delta \leqslant 1$, with confidence at least $1 - \frac{\delta}{2}$, there holds

$$(\mathcal{E}(Q_d(m)) - \mathcal{E}(m)) - (\mathcal{E}_{\boldsymbol{z}}(Q_d(m)) - \mathcal{E}_{\boldsymbol{z}}(m)) \leqslant \frac{70M^2}{n}\log\frac{2}{\delta} + \frac{1}{2}(\mathcal{E}(Q_d(m)) - \mathcal{E}(m)).$$

Proof. From (A2), we know that $Q_d(f,\boldsymbol{x}) \in \mathcal{H}_d$. It follows from Proposition A1 that

$$|Q_d(f,\boldsymbol{x})| \leqslant |f(\boldsymbol{x})| + \omega_k\left(f,\frac{1}{d}\right).$$

If the kth order of partial derivative of $f(\boldsymbol{x})$ belongs to $\mathrm{Lip}_C 1$, we have

$$\omega_k\left(f,\frac{1}{d}\right)\leqslant C_k'\frac{1}{d^k},$$

where C_k' is a constant depending on k.

When $d^k\geqslant\frac{C_k'}{M}$, we have

$$|Q_d(f,\boldsymbol{x})|\leqslant M+C_k^{'}\frac{1}{d^k}\leqslant 2M,\qquad\forall\boldsymbol{x}\in[-1,1]^s.$$

Hence, $Q_d(f,\boldsymbol{x})\in\mathcal{F}_d$, and $|m(x)|\leqslant M$. It follows that

$$|\xi|=|(Q_d(m,x)-m(x))(Q_d(x)+m(x)-2y)|\leqslant 15M^2,$$

Then we get

$$|\xi-\mu|\leqslant B=30M^2$$

and

$$\sigma^2\leqslant\boldsymbol{E}(\xi^2)\leqslant c_\xi\boldsymbol{E}(\xi)=25M^2\boldsymbol{E}(\xi).$$

Applying $(Q_d(m,\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2$ to Lemma A3, we have

$$\begin{aligned}(\mathcal{E}(Q_d(m))-\mathcal{E}(m))-(\mathcal{E}_{\boldsymbol{z}}(Q_d(m))-\mathcal{E}_{\boldsymbol{z}}(m))&\leqslant\sqrt{\varepsilon(\mathcal{E}(Q_d(m))-\mathcal{E}(m)+\varepsilon)}\\&\leqslant\varepsilon+\frac{1}{2}(\mathcal{E}(Q_d(m))-\mathcal{E}(m))\end{aligned}$$

with confidence at least $1-\exp\{-\frac{n\varepsilon}{70M^2}\}$.

Let $\exp\{-\frac{n\varepsilon}{70M^2}\}=\frac{\delta}{2}$. Then we have

$$\varepsilon=\frac{70M^2}{n}\log\frac{2}{\delta}.$$

Therefore, there holds

$$(\mathcal{E}(Q_d(m))-\mathcal{E}(m))-(\mathcal{E}_{\boldsymbol{z}}(Q_d(m))-\mathcal{E}_{\boldsymbol{z}}(m))\leqslant\frac{70M^2}{n}\log\frac{2}{\delta}+\frac{1}{2}(\mathcal{E}(Q_d(m))-\mathcal{E}(m))$$

with confidence at least $1-\frac{\delta}{2}$.

Definition A1 [16]**.** For a subset $\mathcal{F}$ of a metric space and $\varepsilon>0$, the covering number $\mathcal{N}(\mathcal{F},\varepsilon)$ is defined to be the minimal integer $l\in\mathbb{N}$ such that there exist l disks with radius ε covering $\mathcal{F}$.

The covering number has been extensively studied [13, 14][1). B_R is the closed ball with radius R in d-dimensional space, there holds (see [17])

$$\log\mathcal{N}(B_R,\varepsilon)\leqslant r\log\frac{4R}{\varepsilon}.\tag{A5}$$

In order to estimate the term $\{\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\}-\{\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}})-\mathcal{E}_{\boldsymbol{z}}(m)\}$ in (A4), we also need the following Lemma. For a function g on Z, denote $E(g)=\int_Z g(z)d\rho$.

Lemma A4 [13]**.** Let $\mathcal{G}$ be a set functions on Z such that for some $c_\rho\geqslant 0,|g-Eg|\leqslant B$ almost everywhere. If $E(g^2)\leqslant c_\rho E(g)$ for each $g\in\mathcal{G}$, for every $\varepsilon>0$, and $0<\alpha\leqslant 1$, we have

$$\mathrm{Prob}_{\boldsymbol{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}}\frac{E(g)-\frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{E(g)+\varepsilon}}\geqslant 4\alpha\sqrt{\varepsilon}\right\}\leqslant\mathcal{N}(\mathcal{G},\alpha\varepsilon)\exp\left\{-\frac{\alpha^2 m\varepsilon}{2c_\rho+\frac{2}{3}B}\right\}.$$

We apply Lemma A4 to a set of functions

$$\mathcal{G}=\{g:g(z)=(f(\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2,f\in\mathcal{F}_d\},$$

where $\mathcal{F}_d$ is defined in section 2.

[1) 1 Williamson R C, Smola A J, Schokopf B. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. IEEE Trans Inf Theory, 2001, 47: 2516–2532

2 Zhang T. Effective dimension and generalization of kernel learning. In: Proceedings of the 16th Annual Conference on Neural Information Processing Systems, Vancouver, Canada, 2002. 454–461

3 Zhang T. Leave-one-out bounds for kernel methods. Neural Comput, 2003, 13: 1397–1437

Theorem A2. For all $\varepsilon > 0$, we have

$$\begin{aligned}&\mathrm{Prob}_{\boldsymbol{z}\in Z^n}\left\{\sup_{f\in\mathcal{F}_d}\frac{\mathcal{E}(f)-\mathcal{E}(m)-(\mathcal{E}_{\boldsymbol{z}}(f)-\mathcal{E}_{\boldsymbol{z}}(m))}{\sqrt{\mathcal{E}(f)-\mathcal{E}(m)+\varepsilon}}\leqslant\sqrt{\varepsilon}\right\}\\&\geqslant 1-\exp\left\{(2d)^s\log\left(\frac{4M^2}{\varepsilon}\right)-\frac{n\varepsilon}{32M^2}\right\}.\end{aligned}$$

Proof. Consider the function set $\mathcal{G}$. Each function $g\in\mathcal{G}$ has the form $g(z)=(f(\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2$ with $f\in\mathcal{F}_d$, and satisfies $E(g)=\mathcal{E}(f)-\mathcal{E}(m)\geqslant 0$, where

$$g(z)=(f(\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2=(f(\boldsymbol{x})-m(\boldsymbol{x}))(f(\boldsymbol{x})+m(\boldsymbol{x})-2y).$$

Since $|f(\boldsymbol{x})|\leqslant M/4$ and $|m(\boldsymbol{x})|\leqslant M/4$ for any $\boldsymbol{x}\in[-1,1]^s$, we obtain

$$|g(z)|\leqslant M/2\times M=M^2/2.$$

So we have $|g(z)-E(g)|\leqslant M^2$ almost everywhere.

We take $c_0=M^2/2, B=M^2$. Applying Lemma A4 with $\alpha=\frac{1}{4}$ to the function $\mathcal{G}$, for every $\varepsilon>0$, with confidence at least

$$1-\mathcal{N}\left(\mathcal{G},\frac{\varepsilon}{4}\right)\exp\left\{-\frac{n\varepsilon}{32M^2}\right\},$$

there holds

$$\sup_{f\in\mathcal{G}}\frac{\mathcal{E}(f)-\mathcal{E}(m)-(\mathcal{E}_{\boldsymbol{z}}(f)-\mathcal{E}_{\boldsymbol{z}}(m))}{\sqrt{\mathcal{E}(f)-\mathcal{E}(m)+\varepsilon}}\leqslant\sqrt{\varepsilon}.$$

According to the definition of the function $g(\boldsymbol{z})$, we know

$$\begin{aligned}|g_1(z)-g_2(z)|&\leqslant|f_1(\boldsymbol{x})-f_2(\boldsymbol{x})||2y-f_1(\boldsymbol{x})-f_2(\boldsymbol{x})|\\&\leqslant 4M|f_1(\boldsymbol{x})-f_2(\boldsymbol{x})|.\end{aligned}$$

Therefore

$$\|g_1-g_2\|_\infty\leqslant M\|f_1-f_2\|_\infty,$$

which, (A5) implies

$$\log\mathcal{N}\left(\mathcal{G},\frac{\varepsilon}{4}\right)\leqslant\log\mathcal{N}\left(\mathcal{F}_d,\frac{\varepsilon}{M}\right)\leqslant(2d)^s\log\left(\frac{4M^2}{\varepsilon}\right).$$

The proof of Theorem A2 is completed.

Using Theorem A2, we can now start with the proof of Theorem 1.

Proof of Theorem 1. In the proof, we use the following error decomposition:

$$\begin{aligned}&\int_X|f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x})|^2\nu(d\boldsymbol{x})\\&\quad\leqslant\{(\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m))-(\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}})-\mathcal{E}_{\boldsymbol{z}}(m))\}+(\mathcal{E}(Q_d(m))-\mathcal{E}(m))\\&\quad+\{(\mathcal{E}_{\boldsymbol{z}}(Q_d(m))-\mathcal{E}_{\boldsymbol{z}}(m))-(\mathcal{E}(Q_d(m))-\mathcal{E}(m))\}\\&\quad=T_1+T_2+T_3.\end{aligned}$$

We begin with bounding T_1 in (A6). From Theorem 2, we know that for any $f\in\mathcal{F}_d$ there holds

$$\mathcal{E}(f)-\mathcal{E}(m)-(\mathcal{E}_{\boldsymbol{z}}(f)-\mathcal{E}_{\boldsymbol{z}}(m))\leqslant\sqrt{t}\sqrt{\mathcal{E}(f)-\mathcal{E}(m)+t}$$

with confidence at least $1-\exp\{(2d)^s\log(\frac{4M^2}{t})-\frac{mt}{32M^2}\}$.

Recall an elementary inequality :

$$ab\leqslant\frac{1}{2}(a^2+b^2),\quad\forall a,b\in\mathcal{R}.$$

We find that there holds

$$\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)-(\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}})-\mathcal{E}_{\boldsymbol{z}}(m))\leqslant t+\frac{1}{2}(\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m))$$

with confidence at least

$$1-\exp\left\{(2d)^s\log\left(\frac{4M^2}{t}\right)-\frac{nt}{32M^2}\right\}.$$

We need to bound the positive solution ε_0 to the equation

$$h(\varepsilon) = (2d)^s \log\left(\frac{4M^2}{\varepsilon}\right) - \frac{n\varepsilon}{32M^2} = \log\frac{\delta}{2}.$$

The function $h : \mathcal{R}_+ \to \mathcal{R}$ is strictly decreasing. Hence $\varepsilon_0 \leqslant \varepsilon^*$ if $h(\varepsilon^*) \leqslant \log\frac{\delta}{2}$.

When $\varepsilon \geqslant \frac{1}{n}$, we have

$$h(\varepsilon) \leqslant (2d)^s \log\left(4M^2 n\right) - \frac{n\varepsilon}{32M^2}.$$

Taking $\varepsilon^* \geqslant \frac{1}{n}$ satisfying

$$(2d)^s \log\left(4M^2 n\right) - \frac{n\varepsilon}{32M^2} \leqslant \log\frac{\delta}{2},$$

we have $h(\varepsilon^*) \leqslant \log\frac{\delta}{2}$.

For large enough $d > 0$, we have

$$\varepsilon^* \geqslant \frac{32M^2}{n}\log\frac{2}{\delta} + \frac{32M^2(2d)^s\log\left(4M^2n\right)}{n} \geqslant \frac{1}{n}.$$

Then we get

$$\varepsilon_0 \leqslant \frac{32M^2}{n}\log\frac{2}{\delta} + \frac{32M^2(2d)^s\log\left(4M^2n\right)}{n}.$$

From Theorem A2, there holds

$$\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m) - (\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}}) - \mathcal{E}_{\boldsymbol{z}}(m)) \leqslant \frac{32M^2}{n}\log\frac{2}{\delta} + \frac{32M^2(2d)^s\log\left(4M^2n\right)}{n} + \frac{1}{2}\left(\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m)\right)$$

with confidence at least $1 - \frac{\delta}{2}$.

Combining Theorem A1 and (A6), there holds

$$\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m) \leqslant \frac{204M^2}{n}\log\frac{2}{\delta} + \frac{64M^2(2d)^s\log\left(4M^2n\right)}{n} + 3\left(\mathcal{E}(Q_d(m)) - \mathcal{E}(m)\right)$$

with confidence at least $1 - \delta$.

The proof of Theorem 1 is completed.

In order to prove Corollary 1, we need to estimate

$$\inf_{f\in\mathcal{F}_d}\int_X (f(\boldsymbol{x}) - m(\boldsymbol{x}))^2\nu(d\boldsymbol{x}).$$

From (A2), we know that $Q_d(f,\boldsymbol{x}) \in \mathcal{H}_d$. And Proposition A1 tells us that

$$|Q_d(f,\boldsymbol{x})| \leqslant |f(\boldsymbol{x})| + \omega_k\left(f, \frac{1}{d}\right).$$

If the kth order of partial derivative of $f(\boldsymbol{x})$ belongs to $\mathrm{Lip}_C 1$, then we get

$$\omega_k\left(f, \frac{1}{d}\right) \leqslant C_k'\frac{1}{d^k},$$

where C_k' is a constant depending on k.

From Proposition A1 we have

$$\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m) \leqslant \frac{204M^2}{n}\log\frac{2}{\delta} + \frac{64M^2(2d)^s\log\left(4M^2n\right)}{n} + 3C_k'\frac{1}{d^{2k}},$$

and this is minimized for

$$d = \left[\left(\frac{3C_k'n}{2^s 64M^2\log(4M^2n)}\right)^{\frac{1}{2k+s}}\right],$$

where $[a]$ denotes the integer part of real number a.

For $n \geqslant 4M^2$, there holds

$$\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m) \leqslant \frac{204M^2}{n}\log\frac{2}{\delta} + C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}$$

with confidence at least $1-\delta$. $C_{K,s}=2(108M^2 2^s)^{\frac{2k}{2k+s}}+2(3C_K^{'})^{\frac{s}{2k+s}}$.

Let $t=\frac{204M^2}{n}\log\frac{2}{\delta}+C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}$. Then

$$\delta=2\exp\left\{-\frac{t-C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}}{\frac{204M^2}{n}}\right\}.$$

The above probability inequality can be rewritten as

$$\mathrm{Prob}_{\boldsymbol{z}\in Z^n}\{\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\geqslant t\}\leqslant 2\exp\left\{-\frac{t-C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}}{\frac{204M^2}{n}}\right\}.$$

For $\tau\geqslant\frac{1}{n}$, we have

$$\begin{aligned}\boldsymbol{E}\int_X(f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(d\boldsymbol{x})&=\int_0^\infty \mathrm{Prob}_{\boldsymbol{z}\in Z^n}\{\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\geqslant t\}dt\\&\leqslant\tau+\int_\tau^\infty 2\exp\left\{-\frac{t-C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}}{\frac{204M^2}{n}}\right\}dt,\end{aligned}$$

and this is minimized for

$$\tau=\frac{204M^2\log 2}{n}+C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}.$$

Hence

$$\boldsymbol{E}\int_X(f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(d\boldsymbol{x})\leqslant\frac{408M^2\log 2}{n}+2C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}.$$

The proof of Corollary 1 is finished.