应用概率统计 第二十七卷
第六期 2011年12月

Chinese Journal of Applied Probability
and Statistics Vol.27 No.6 Dec. 2011

Learning Rates of Empirical Risk Minimization Regression with Beta-Mixing Inputs *

Zou Bin[1,2] Xu Zongben[2] Zhang Hai[2,3]

([1]*Faculty of Mathematics and Computer Science, Hubei University, Wuhan, 430062*)

([2]*Institute for Information and System Science, Xi'an Jiaotong University, Xi'an, 710049*)

([3]*Department of Mathematics, Northwest University, Xi'an, 710069*)

Abstract

The study of empirical risk minimization (ERM) algorithm associated with least squared loss is one of very important issues in statistical learning theory. The main results describing the learning rates of ERM regression are almost based on independent and identically distributed (i.i.d.) inputs. However, independence is a very restrictive concept. In this paper we go far beyond this classical framework by establishing the bound on the learning rates of ERM regression with geometrically β-mixing inputs. We prove that the ERM regression with geometrically β-mixing inputs is consistent and the main results obtained in this paper are also suited to a large class of Markov chains samples and hidden Markov models.

Keywords: Learning rate, empirical risk minimization, β-mixing, least squared loss.

AMS Subject Classification: 68TXX.

§1. Introduction

The study of distribution free non-parametric estimation of regression is a very important issues in machine learning from random sampling. The previous results in this topic are almost based on the assumption of independent and identically distributed (i.i.d.) inputs. For example, Vapnik (1998), Cucker and Smale (2001), Smale and Zhou (2003, 2004), DeVore et al. (2006) established the theoretical justification in terms of both universal consistency and learning rates for the problem of regression estimation respectively.

However, independence is a very restrictive concept in several ways[6]. First, it is often an assumption, rather than a deduction on the basis of observations. Second, it is an all or nothing property, in the sense that two random variables are either independent or they are not — the definition does not permit an intermediate notion of being nearly independent. As a result, many of the proofs based on the assumption that the underlying stochastic

*This research is supported by National 973 project (2007CB311002), NSFC key project (70501030), NSFC project (61070225) and China Postdoctoral Science Foundation (20080440190, 200902592).

Received March 16, 2009.

sequence is i.i.d. are rather "fragile". The notion of mixing allows one to put the notion of "near independence" on a firm mathematical foundation, and moreover, permits one to derive a robust rather than a "fragile" theory. In addition, this i.i.d. assumption can not be strictly justified in real-world problems. For example, many machine learning applications such as market prediction, system diagnosis, and speech recognition are inherently temporal in nature, and consequently not i.i.d. processes[7]. Therefore, relaxations of such i.i.d. assumption have been considered for quite a while in both machine learning and statistics literatures. For example, Smale and Zhou (2009) studied online learning with Markov sampling. Yu (1994) established the rates of uniform convergence of the empirical means to their means based on stationary mixing sequences. Vidyasagar (2003) proved that most of the desirable properties (e.g. PAC property or UCEMUP property) of i.i.d. sequence are preserved when the underlying sequence is β-mixing sequence. Nobel and Dembo (1993) proved that, if a family of functions has the property that empirical means based on i.i.d. sequence uniform convergence to their values as the number of samples approaches infinity, then the family of functions continues to have the same property if the i.i.d. sequence is replaced by β-mixing sequence. Karandikar and Vidyasagar (2002) extended this result to the case where the underlying probability is itself not fixed, but varies over a family of measures. Steinwart et al. (2009) proved that the SVMs for both classification and regression are consistent if the data-generating process (e.g. β-mixing process, Markov process) satisfies a certain type of law of large numbers (e.g. WLLNE, SLLNE). Xu and Chen (2008) established the learning rates of regularized regression for exponentially strongly mixing sequence. Zou et al. (2009) established the bounds on the generalization performance of the ERM algorithm with strongly mixing observations.

There are many definitions of non-independent sequences in [6, 9], but we are only interested in β-mixing sequence in this paper, the reasons are as follows: First, Vidyasagar (2003) pointed out that in machine learning application, α-mixing is "too weak" an assumption and ϕ-mixing is "too strong" an assumption, β-mixing is "just right" and more meaningful in the context of PAC learning. Second, Markov chain samples appear so often and naturally in applications, especially in biological (DNA or protein) sequence analysis, speech recognition, character recognition, content-based web search and market prediction, and Vidyasagar (2003), Meyn and Tweedie (1993) proved that a very large class of Markov chains and hidden Markov models can produce β-mixing sequences. To extend previous results in [2, 5] on the study of regression estimation to the case where the i.i.d. inputs replaced by β-mixing inputs, and to improve the results in [6] based on β-mixing sequence, we first introduce the "effective number of observations" for geometrically β-mixing process by enlightening the idea from [15], and establish the bound on the rate of uniform convergence of ERM algorithm with geometrically β-mixing for regression

estimation. Then we obtain the bound on the learning rates of ERM regression with geometrically β-mixing and prove that the ERM algorithm with geometrically β-mixing for regression estimation is consistent.

The rest of this paper is organized as follows: In Section 2, we introduce the necessary notion and notations. In Section 3, we present the main results of this paper. In Section 4, we give the proof of our main results. We present some useful discussions and comparisons in Section 5. Finally, we conclude this paper with some useful remarks in Section 6.

§2. Preliminaries

We introduce some notations and do some preparations in this section.

let $\mathcal{Z} = \{z_i = (x_i, y_i)\}_{i=-\infty}^{\infty}$ be a stationary real-valued stochastic process defined on a probability space $(\Omega^\infty, \mathcal{F}^\infty, \mathsf{P})$. For $-\infty < i < \infty$, let $\mathcal{F}_{-\infty}^k$ denote the σ-algebra generated by the random variables z_i, $i \le k$, and similarly let $\mathcal{F}_k^\infty$ denote the σ-algebra generated by the random variables z_i, $i \ge k$. Let $\mathsf{P}_{-\infty}^k$ and P_k^∞ denote the corresponding marginal probability measures respectively. Let P_0 denote the marginal probability of each of the z_i. Let $\overline{\mathcal{F}}_1^{k-1}$ denote the σ-algebra generated by the random variables z_i, $i \le 0$ as well as z_j, $j \ge k$. Thus the bar over the $\mathcal{F}$ serves to remind us that the random variables between 1 and $k-1$ are missing from the list of variables that generated $\mathcal{F}$. With these notations, there are several definitions of mixing, but we shall be concerned with only one, namely, β-mixing in this literature[6, 9].

Definition 2.1 The sequence $\mathcal{Z}$ is called β-mixing, or completely regular[6], if

$$\sup_{C \in \overline{\mathcal{F}}_1^{k-1}} |\mathsf{P}(C) - (\mathsf{P}_{-\infty}^0 \times \mathsf{P}_1^\infty)(C)| = \beta(k) \to 0 \qquad \text{as} \qquad k \to \infty,$$

where $\beta(k)$ is called the β-mixing coefficient.

Assumption 2.1 A sequence $\mathcal{Z}$ is called geometrically β-mixing[6], if the β-mixing coefficient satisfies

$$\beta(k) \le \mu \lambda^k, \qquad k \ge 1$$

for some constants μ and $\lambda < 1$.

Remark 1 (i) In Definition 2.1, if the "future" events beyond time k were to be truly independent of the "past" events before time 0, then the probability measure P would exactly equal the "split" measure $\mathsf{P}_{-\infty}^0 \times \mathsf{P}_1^\infty$. The β-mixing coefficient thus measures how nearly the product measure approximates the actual measure P.

(ii) If the sequence $\mathcal{Z}$ consists of i.i.d. random variables, then P equals the measure $(\mathsf{P}_0)^\infty$, which denotes the measure on $(\Omega^\infty, \mathcal{F}^\infty)$. In such a case, $\beta(k)$ is zero for any integer k, that is, i.i.d. random variables satisfy the assumption with $\mu = 0$.

Denote by $\mathbf{z}$ the sample set of size m

$$\mathbf{z} = \{z_1, z_2, \ldots, z_m\}$$

drawn from the geometrically β-mixing sequence $\mathcal{Z}$. Let X be a compact domain or a manifold in Euclidean space and $Y = \mathbb{R}$. The goal of machine learning from sample set $\mathbf{z}$ is to find a function $f : X \to Y$ that assigns values to objects such that if new objects are given, the function f will forecast them correctly.

A main concept is the expected risk (or least squares error) of function f defined by

$$\mathcal{E}(f) = \mathsf{E}[\ell(f, z)] = \int_{\mathcal{Z}} (f(x) - y)^2 \mathrm{d}(\mathsf{P}_0),$$

where the function $\ell(f, z)$, which is integrable for any f, called loss function. Define $f_0 : X \to Y$ by

$$f_0 = \int_Y y \mathrm{d}\mathsf{P}(y|x).$$

The function f_0 is called the regression function of P_0, where $\mathsf{P}(y|x)$ is the conditional probability measure on Y with respect to x. It is clear that the regression function f_0 minimizes the expected risk $\mathcal{E}(f)$ (see [2]), that is,

$$\mathcal{E}(f_0) = \inf_{f \in L_2(X, \mathsf{P}_x)} \mathcal{E}(f),$$

where P_x is the marginal distribution of P_0 on X. For the sake of simplicity, we denote $\mathsf{E}[\zeta, \mathsf{P}_0^\infty]$ as the expected value of random variable ζ with respect to probability measure P_0^∞. If the probability measure is P_0, then we simply write $\mathsf{E}[\zeta]$ in the sequel.

A learning task is "learn" (i.e. to find a good approximation of) f_0 from random sample set $\mathbf{z}$. Thus we hope to find the function that approximates the regression function f_0 through minimizing the expected risk $\mathcal{E}(f)$. Since one knows only the set $\mathbf{z}$ of random samples instead of the distribution P_0, the minimizer of the expected risk $\mathcal{E}(f)$ can not be computed directly. According to the Empirical Risk Minimization (ERM) principle[1], we minimize, instead of the expected risk $\mathcal{E}(f)$, the so called empirical risk (or error)

$$\mathcal{E}_m(f) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2.$$

Given a function set $\mathcal{H}$, we define $f_{\mathcal{H}}$ to be a function minimizing the expected risk $\mathcal{E}(f)$ over the function set $\mathcal{H}$, i.e.,

$$f_{\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}(f) = \arg\min_{f \in \mathcal{H}} \int_{\mathcal{Z}} (f(x) - y)^2 \mathrm{d}(\mathsf{P}_0).$$

According to the principle of ERM, we shall consider the function $f_{\mathbf{z}}$

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}_m(f) = \arg\min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2,$$

as an approximation the function $f_{\mathcal{H}}$. Thus a central question of ERM algorithm for regression estimation is how well $f_{\mathbf{z}}$ really approximate $f_{\mathcal{H}}$. If it is well, the ERM algorithm for regression estimation is said to be of generalization ability. To characterize generalization capability of the ERM regression with geometrically β-mixing inputs requires in essence to decipher how close $f_{\mathbf{z}}$ is from $f_{\mathcal{H}}$. As mentioned earlier, we can primarily measure the difference between $f_{\mathbf{z}}$ and $f_{\mathcal{H}}$ in the $L_2(X,\mathsf{P}_x)$ norm. Notice that for any $f\in L_2(X,\mathsf{P}_x)$, we have

$$\begin{aligned}\mathcal{E}(f)-\mathcal{E}(f_{\mathcal{H}}) &= \int_{\mathcal{Z}}(f(x)-y)^2\mathrm{d}\mathsf{P}_0-\int_{\mathcal{Z}}(f_{\mathcal{H}}(x)-y)^2\mathrm{d}(\mathsf{P}_0)\\ &= \int_X\{f^2-2ff_{\mathcal{H}}+f_{\mathcal{H}}^2\}\mathrm{d}(\mathsf{P}_x)\\ &= \|f-f_{\mathcal{H}}\|^2. \end{aligned}\tag{2.1}$$

Thus the generalization capability of ERM algorithm with β-mixing inputs for regression estimation can be measured by $\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})$.

Note that $f_{\mathcal{H}}$ may not equal the regression function f_0, since $f_{\mathcal{H}}$ depends on the choice of the hypothesis space $\mathcal{H}$. Thus another aim of this paper is to estimate the deviation between $f_{\mathbf{z}}$ and f_0. For this purpose, we first give some basic assumptions on the hypothesis space $\mathcal{H}$ and the loss function $\ell(f,z)$:

(i) Assumption on the hypothesis space $\mathcal{H}$: we suppose that $\mathcal{H}$ is a compact subset of $\mathcal{C}(X)$, and $\mathcal{H}$ is contained in a finite ball in $\mathcal{C}(X)$. Thus for some positive constant C_0, the covering number of $\mathcal{H}$ satisfies (see [5])

$$\mathcal{N}(\mathcal{H},\varepsilon)\le\exp\{C_0\varepsilon^{-1/r}\}.\tag{2.2}$$

(ii) Assumption on the loss function: we assume that $|f(x)-y|\le M$ for any $z\in\mathcal{Z}$ and for any $f\in\mathcal{H}$, and M is finite, thus we have $\ell(f,z)\le M^2$.

§3. Main Results

To measure the generalization performance of learning machine, Vapnik (1998), Cucker and Smale (2001), DeVore et al. (2006) obtained the bound on the rate of the empirical risks uniform convergence to their expected risks on a given set $\mathcal{H}$ (or loss function set Q) based on i.i.d. sequences, that is, for any $\varepsilon>0$, they bounded the term

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|>\varepsilon\Big\}.\tag{3.1}$$

In order to prove that the PAC property is preserved if the i.i.d. input sequence is replaced by β-mixing process, Vidyasager (2003) also bound the term (3.1) for β-mixing sequence,

but his results (see Theorem 6.13 of [6]) consists of two terms, without an explicit convergence rate. The interested reader can consult [6] for the details.

In order to extend these results in [2, 5] to the case where the i.i.d. sequence is replaced by β-mixing sequence, and to improve the estimates in [6], we also bound the term (3.1) for ERM regression with geometrically β-mixing samples. Our main results can be stated as follows.

Theorem 3.1 Let $\mathcal{Z}$ be a stationary β-mixing sequence with the mixing coefficient satisfying Assumption 2.1. Assume that $|f(x)-y|<M$ for any $f\in\mathcal{H}$ and for all $z\in\mathcal{Z}$. Let

$$m^{(\beta)}=\Big\lfloor m\Big\lceil\Big\{\frac{8m}{\ln(1/\lambda)}\Big\}^{1/2}\Big\rceil^{-1}\Big\rfloor,$$

where m denotes the number of observations drawn from $\mathcal{Z}$ and $\lfloor u\rfloor$ ($\lceil u\rceil$) denotes the greatest (least) integer less (greater) than or equal to u. Then for any ε, $0<\varepsilon<3M^2/4$,

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}_m(f)-\mathcal{E}(f)|>\varepsilon\Big\}\le 2(1+\mu e^{-2})\mathcal{N}\Big(\mathcal{H},\frac{\varepsilon}{8M}\Big)\exp\Big\{\frac{-m^{(\beta)}\varepsilon^2}{2M^4}\Big\}. \tag{3.2}$$

In particular, if $\mathcal{Z}$ is an i.i.d. sequence, according to Remark 1, we take $\mu=0$ in Theorem 3.1 and ignore the multiplicative constant $1+\mu e^{-2}$, the following bound then immediately follows from Theorem 3.1.

Corollary 3.1 Let $\mathcal{Z}$ be an i.i.d. sequence, and assume that $|f(x)-y|<M$ for any $f\in\mathcal{H}$ and for all $z\in\mathcal{Z}$. Then for any ε, $0<\varepsilon\le 3M^2/4$,

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}_m(f)-\mathcal{E}(f)|>\varepsilon\Big\}\le 2\mathcal{N}\Big(\mathcal{H},\frac{\varepsilon}{8M}\Big)\exp\Big\{\frac{-m\varepsilon^2}{2M^4}\Big\}.$$

Remark 2 (i) $m^{(\beta)}$ is called the "effective number of observations" for geometrically β-mixing process. From Theorem 3.1 and Corollary 3.1, we can find that $m^{(\beta)}$ play the same role in our analysis as that played by the number of observations m in the i.i.d. case.

(ii) Since $m^{(\beta)}\to\infty$ as $m\to\infty$, by Theorem 3.1, we have that for any ε, $0<\varepsilon\le 3M^2/4$,

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}_m(f)-\mathcal{E}(f)|>\varepsilon\Big\}\to 0,\qquad \text{as } m\to\infty.$$

This shows that as long as the covering number of hypothesis space $\mathcal{H}$ is finite, the empirical risk $\mathcal{E}_m(f)$ can uniformly converge to the expected risk $\mathcal{E}(f)$, and the convergence speed may be exponential. This assertion is well known for the ERM regression with i.i.d. samples (see, e.g. [2, 5]). We have generalized this classical results in [2, 5] to geometrically β-mixing sequences.

Theorem 3.1 will be proven in the next section. Before going into the technical proofs, we first deduce the bound on the learning rates of ERM regression based on geometrically β-mixing inputs.

Proposition 3.1 Let $\mathcal{Z}$ be a stationary β-mixing sequence with the mixing coefficient satisfying Assumption 2.1. Assume that $|f(x)-y| \le M$ for all $z \in \mathcal{Z}$ and for all functions f in $\mathcal{H}$. Then for any $\eta \in (0,1]$, the inequality

$$\|f_{\mathbf{z}} - f_{\mathcal{H}}\|^2 \le \varepsilon(m,\eta) + M^2\sqrt{\frac{\ln(C/\eta)}{2m^{(\beta)}}} \tag{3.3}$$

holds true with probability at least $1-2\eta$ provided that

$$m^{(\beta)} \ge \max\Big\{\frac{64\ln(C/\eta)}{9}, \frac{2^{6+5/r}C_0}{3^{2+1/r}M^{1/r}}\Big\},$$

where

$$\varepsilon(m,\eta) \le \max\Big\{2M^2\Big[\frac{\ln(C/\eta)}{m^{(\beta)}}\Big]^{1/2}, \Big[\frac{4M^4C_0(8M)^{1/r}}{m^{(\beta)}}\Big]^{r/(2r+1)}\Big\}.$$

Remark 3 Since when $m \to \infty$, $m^{(\beta)} \to \infty$, we have $\varepsilon(m,\eta) \to 0$, as $m \to \infty$. By inequality (3.3), we then have $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \to 0$, as $m \to \infty$. This shows that the ERM regression based on geometrically β-mixing inputs over the hypothesis space $\mathcal{H}$ is consistent whenever the covering number of target function set $\mathcal{H}$ is finite.

By Proposition 3.1, we can find that as long as the covering number of hypothesis space $\mathcal{H}$ is finite, the function $f_{\mathbf{z}}$ minimizing the empirical risk $\mathcal{E}_m(f)$ will converge to the function $f_{\mathcal{H}}$ minimizing the expected risk $\mathcal{E}(f)$ over $\mathcal{H}$. But how good can we expect $f_{\mathbf{z}}$ to be as an approximation of the regression function f_0? Proposition 3.2 below gives an answer.

Proposition 3.2 With all notations as in Proposition 3.1, and assume that for some constant c, the entropy numbers of $\mathcal{H}$ satisfy

$$\epsilon_n(\mathcal{H}) \le cn^{-r}, \qquad n = 1,2,\ldots.$$

Then for any $\delta \in (0,1]$, the inequality $\|f_{\mathbf{z}} - f_0\| \le \varepsilon(m,\delta)$ holds true with probability at least $1-\delta$ provided that

$$m^{(\beta)} \ge \min\Big\{\frac{16^3\ln(2C/\delta)}{81M^4}, \frac{4^{4+4/r}C_0}{3^{4+2/r}M^{4+3/r}}\Big\},$$

where

$$\varepsilon(m,\delta) \le \max\Big\{2M\Big[\frac{\ln(2C/\delta)}{m^{(\beta)}}\Big]^{1/4}, \Big[\frac{16^{1+1/r}C_0M^{4+1/r}}{m^{(\beta)}}\Big]^{r/(4r+2)}\Big\}.$$

§4. Proof of Main Results

In this section, our aim is to prove the main results presented in the last section. To bound the learning rates of ERM regression with geometrically β-mixing inputs, we firstly establish the bound on the rate of uniform convergence of ERM regression with β-mixing inputs. Our approach is based on the following lemmas: the first one is the covariance inequality for β-mixing sequences, which is established by Yu in [9]. The second one is Hoeffding's inequality[16].

Lemma 4.1 Suppose $i_0 < i_1 < \cdots < i_l$ are integers[6], and define

$$k = \min_{0\le j\le l-1} i_{j+1} - i_j.$$

Suppose f is essentially bounded and depends only on $z_{i_0}, z_{i_1}, \ldots, z_{i_l}$. Then

$$|\mathsf{E}(f, \mathsf{P}) - \mathsf{E}(f, \mathsf{P}_0^\infty)| \le l\beta(k)\|f\|_\infty.$$

Lemma 4.2 Suppose that X is a zero-mean random variable assuming values in the interval $[a, b]$[16]. Then for any $s > 0$, we have

$$\mathsf{E}[\exp(sX)] \le \exp(s^2(b-a)^2/8).$$

To exploit the β-mixing property, we decompose the index set $I = \{1, 2, \ldots, m\}$ into different parts as follows: Given an integer m, choose any integer $k_m \le m$, and define $l_m = \lfloor m/k_m \rfloor$ to be the integer part of m/k_m. For the time being, k_m and l_m are denoted respectively by k and l, so as to reduce natational clutter. The dependence of k and l on m is restored near the end of the paper. Let $r = m - kl$, and define

$$I_i = \begin{cases} \{i, i+k, \ldots, i+lk\}, & i = 1, 2, \ldots, r, \\ \{i, i+k, \ldots, i+(l-1)k\}, & i = r+1, \ldots, k. \end{cases}$$

Note that $\bigcup_i I_i$ equals the index set I and that within each set I_i, the elements are pairwisely separated by at least k. Then we have the following theorem.

Theorem 4.1 With all notations as in Theorem 3.1, for any ε, $0 < \varepsilon < 3M^2/4$, and any $f \in \mathcal{H}$

$$\mathsf{P}\{|\mathcal{E}_m(f) - \mathcal{E}(f)| > \varepsilon\} \le 2(1 + \mu e^{-2}) \exp\left\{\frac{-2m^{(\beta)}\varepsilon^2}{M^4}\right\}. \tag{4.1}$$

Proof Let $p_i = |I_i|/m$ for $i = 1, 2, \ldots, k$, and define

$$T_i = \ell(f, z_i) - \mathsf{E}[\ell(f, z_i)], \qquad \pi_m(\mathbf{z}) = \frac{1}{m}\sum_{i=1}^m T_i, \qquad b_i(\mathbf{z}) = \frac{1}{|I_i|}\sum_{j\in I_i} T_i.$$

Then we have

$$\mathcal{E}_m(f) - \mathcal{E}(f) = \pi_m(\mathbf{z}) = \sum_{i=1}^{k} p_i b_i(\mathbf{z}).$$

Since $\exp(\cdot)$ is convex, we have that for any $\gamma > 0$,

$$\exp(\gamma\pi_m(\mathbf{z})) = \exp\Big[\sum_{i=1}^{k} \gamma p_i b_i(\mathbf{z})\Big] \le \sum_{i=1}^{k} p_i \exp(\gamma b_i(\mathbf{z})).$$

Now take the expectation of both sides with respect to P, we have

$$\mathsf{E}[\exp(\gamma\pi_m(\mathbf{z})), \mathsf{P}] \le \sum_{i=1}^{k} p_i \mathsf{E}[\exp(\gamma b_i(\mathbf{z})), \mathsf{P}].$$

Since

$$\exp(\gamma b_i(\mathbf{z})) = \exp\Big[\frac{\gamma}{|I_i|}\sum_{j\in I_i} T_j\Big] = \prod_{j\in I_i} \exp\Big(\frac{\gamma T_j}{|I_i|}\Big) \le \Big[\exp\Big(\frac{\gamma M^2}{|I_i|}\Big)\Big]^{|I_i|} \le e^{\gamma M^2},$$

where in the last step we use the fact that $T_i = \ell(f, z_i) - \mathsf{E}[\ell(f, z_i)] \le M^2$.

By Lemma 4.1, we get

$$\mathsf{E}[e^{\gamma b_i(\mathbf{z})}, \mathsf{P}] \le (|I_i| - 1)\beta(k)\|e^{\gamma b_i(\mathbf{z})}\|_\infty + \mathsf{E}[e^{\gamma b_i(\mathbf{z})}, \mathsf{P}_0^\infty]. \tag{4.2}$$

Since under the measure P_0^∞, the various z_i are independent, we have

$$\mathsf{E}[e^{\gamma b_i(\mathbf{z})}, \mathsf{P}_0^\infty] = \mathsf{E}\Big[\prod_{j\in I_i} \exp(\gamma T_j/|I_i|), \mathsf{P}_0^\infty\Big] = \{\mathsf{E}[\exp(\gamma T_j/|I_i|), \mathsf{P}_0]\}^{|I_i|}.$$

Apply Lemma 4.2 to the function T_j, since T_j has zero mean and values in an interval of width M^2. It follows from Lemma 4.2 that $\mathsf{E}[\exp(\gamma T_j/|I_i|)] \le \exp(\gamma^2 M^4/8|I_i|^2)$. So

$$\mathsf{E}[e^{\gamma b_i(\mathbf{z})}, \mathsf{P}] \le \exp\Big(\frac{\gamma^2 M^4}{8|I_i|}\Big) + (|I_i| - 1)\beta(k)e^{\gamma M^2}.$$

It follows that

$$\mathsf{E}[e^{\gamma\pi_m(\mathbf{z})}, \mathsf{P}] \le \sum_{i=1}^{k} p_i\Big[\exp\Big(\frac{\gamma^2 M^4}{8|I_i|}\Big) + (|I_i| - 1)\beta(k)e^{\gamma M^2}\Big]. \tag{4.3}$$

We now bound the second term on the right-hand side of inequality (4.3) which is denoted henceforth by ϕ. We suppose $\gamma \le 3|I_i|/M^2$, then we have that

$$\begin{aligned}
\phi &= \exp\Big(\frac{\gamma^2 M^4}{8|I_i|}\Big) + (|I_i| - 1)\beta(k)e^{\gamma M^2} \\
&\le \exp\Big(\frac{\gamma^2 M^4}{8|I_i|}\Big) + e^{|I_i|}e^{-2}\mu\lambda^k \cdot e^{\gamma M^2} \\
&\le \exp\Big(\frac{\gamma^2 M^4}{8|I_i|}\Big) + \mu e^{-2}\exp\{k\ln(\lambda) + 4|I_i|\}.
\end{aligned}$$

The second inequality follows from Assumption 2.1 and the fact that $|I_i| - 1 \le e^{|I_i|-2}$ for any $|I_i| \ge 2$.

We require $\exp\{k\ln(\lambda) + 4|I_i|\} \le 1$, which holds if $k\ln(\lambda) + 4|I_i| < 0$. But $|I_i| \le (m/k+1)$, thus the bound holds if $4(m/k+1) \le k\ln(1/\lambda)$. Since $m + k \le 2m$, then the bound holds if $8m \le k\ln(1/\lambda)$ or $\{8m/\ln(1/\lambda)\}^{1/2} \le k$. Let $k = \lceil\{8m/\ln(1/\lambda)\}^{1/2}\rceil$. Then we have

$$\phi \le \exp\Big(\frac{\gamma^2 M^4}{8|I_i|}\Big) + \mu e^{-2}. \tag{4.4}$$

Since inequality (4.4) is true for all γ, $0 < \gamma < 3|I_i|/M^2$. To make the constraint uniform over all i, we then require γ satisfy $0 < \gamma < 3l/M^2 < 3|I_i|/M^2$. Since $\gamma^2 M^4/(8l) > 0$, we have

$$\phi \le (1 + \mu e^{-2}) \exp\Big(\frac{\gamma^2 M^4}{8l}\Big).$$

Returning to inequality (4.3) we have

$$\mathsf{E}[e^{\gamma\pi_m(\mathbf{z})}, \mathsf{P}] \le (1 + \mu e^{-2}) \exp\Big(\frac{\gamma^2 M^4}{8l}\Big).$$

By Markov's inequality, we have that for any $\gamma > 0$

$$\begin{aligned}\mathsf{P}\{\pi_m(\mathbf{z}) > \varepsilon\} &= \mathsf{P}\{e^{\gamma\pi_m(\mathbf{z})} > e^{\gamma\varepsilon}\} \le \frac{\mathsf{E}[\exp\{\gamma\pi_m(\mathbf{z})\}, \mathsf{P}]}{\exp\{\gamma\varepsilon\}} \\ &\le (1 + \mu e^{-2}) \exp\Big\{\frac{\gamma^2 M^4}{8l} - \gamma\varepsilon\Big\}.\end{aligned}$$

Now by substituting $\gamma = 4l\varepsilon/M^4$ and noting that if $\varepsilon \le 3M^2/4$, then γ satisfies $\gamma \le 3l/M^2$. We then obtain that for any ε, $0 < \varepsilon \le 3M^2/4$, the inequality

$$\mathsf{P}\{\pi_m(\mathbf{z}) > \varepsilon\} \le (1 + \mu e^{-2}) \exp\Big\{\frac{-2l\varepsilon^2}{M^4}\Big\}$$

is valid. Since $l = \lfloor m/k \rfloor$, replacing l by $m^{(\beta)}$ then implies that for any ε, $0 < \varepsilon \le 3M^2/4$,

$$\mathsf{P}\{\pi_m(\mathbf{z}) > \varepsilon\} \le (1 + \mu e^{-2}) \exp\Big\{\frac{-2m^{(\beta)}\varepsilon^2}{M^4}\Big\}.$$

By symmetry, we also have

$$\mathsf{P}\{\pi_m(\mathbf{z}) < -\varepsilon\} \le (1 + \mu e^{-2}) \exp\Big\{\frac{-2m^{(\beta)}\varepsilon^2}{M^4}\Big\}.$$

Combining these two bounds leads to the desired inequality (4.1). Then we finish the proof of Theorem 4.1. □

By Theorem 4.1, we now can prove our main theorem on the rate of empirical risks uniform converging to their expected risks for ERM regression with geometrically β-mixing sequence $\mathcal{Z}$.

Proof of Theorem 3.1 Let $\mathcal{H}=\mathcal{H}_1\cup\mathcal{H}_2\cup\cdots\cup\mathcal{H}_{n_1}$, $n_1\in\mathbb{N}$, $L_{\mathbf{z}}(f)=\mathcal{E}(f)-\mathcal{E}_m(f)$ then for any $\varepsilon>0$, whenever $\sup\limits_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon$, there exists k, $1\leq k\leq n_1$, such that $\sup\limits_{f\in\mathcal{H}_k}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon$. This implies the equivalence

$$\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\Longleftrightarrow\exists\, k,\ 1\leq k\leq n_1,\ \text{s.t.}\ \sup_{f\in\mathcal{H}_k}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon. \tag{4.5}$$

By the equivalence (4.5), and by the fact that the probability of a union of events is bounded by the sum of the probabilities of these events, we have

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\Big\}\leq\sum_{k=1}^{n_1}\mathsf{P}\Big\{\sup_{f\in\mathcal{H}_k}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\Big\}. \tag{4.6}$$

Now we estimate the term on the right-hand side of inequality (4.6). Let the balls D_k, $k\in\{1,2,\ldots,n_1\}$ be a cover of $\mathcal{H}$ with center at f_k and radius $\varepsilon/(4M)$. Then, for all $\mathbf{z}\in\mathcal{Z}^m$ and all $f\in D_k$,

$$\begin{aligned}
|\mathcal{E}(f)-\mathcal{E}(f_k)| &= |\mathsf{E}\ell(f,z)]-\mathsf{E}[\ell(f_k,z)]|\\
&\leq \|f-f_k\|_\infty\int_{\mathcal{Z}}|(f(x)-y)+(f_k(x)-y)|\mathrm{d}(\mathsf{P}_0)\\
&\leq 2M\cdot\|f-f_k\|_\infty,\\
|\mathcal{E}_m(f)-\mathcal{E}_m(f_k)| &= \Big|\frac{1}{m}\sum_{i=1}^m\ell(f,z_i)-\frac{1}{m}\sum_{i=1}^m\ell(f_k,z_i)\Big|\\
&\leq \|f-f_k\|_\infty\frac{1}{m}\sum_{i=1}^m|(f(x_i)-y_i)+(f_k(x_i)-y_i)|\\
&\leq 2M\cdot\|f-f_k\|_\infty.
\end{aligned}$$

Therefore we have $|L_{\mathbf{z}}(f)-L_{\mathbf{z}}(f_k)|\leq 4M\cdot\|f-f_k\|_\infty\leq 4M\cdot\varepsilon/(4M)=\varepsilon$. It follows that for any $\mathbf{z}\in\mathcal{Z}^m$ and all $f\in D_k$,

$$\sup_{f\in D_k}|L_{\mathbf{z}}(f)|\geq 2\varepsilon\Longrightarrow|L_{\mathbf{z}}(f_k)|\geq\varepsilon.$$

We thus conclude that for any $k\in\{1,2,\ldots,n\}$,

$$\mathsf{P}\Big\{\sup_{f\in D_k}|L_{\mathbf{z}}(f)|\geq 2\varepsilon\Big\}\leq\mathsf{P}\{|L_{\mathbf{z}}(f_k)|\geq\varepsilon\}.$$

By Theorem 4.1, we can get

$$\mathsf{P}\{|L_{\mathbf{z}}(f_k)|\geq\varepsilon\}\leq 2(1+\mu e^{-2})\exp\Big\{\frac{-2m^{(\beta)}\varepsilon^2}{M^4}\Big\}.$$

Then

$$\mathsf{P}\Big\{\sup_{f\in D_k}|L_{\mathbf{z}}(f)|\geq 2\varepsilon\Big\}\leq 2(1+\mu e^{-2})\exp\Big\{\frac{-2m^{(\beta)}\varepsilon^2}{M^4}\Big\}. \tag{4.7}$$

By inequalities (4.6) and (4.7), we have

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\Big\}\leq 2(1+\mu e^{-2})\mathcal{N}\Big(\mathcal{H},\frac{\varepsilon}{4M}\Big)\exp\Big\{\frac{-2m^{(\beta)}\varepsilon^2}{M^4}\Big\}. \tag{4.8}$$

Theorem 3.1 thus follows from inequality (4.8) by replacing ε by $\varepsilon/2$. □

To prove the bound (Proposition 3.1) on the learning rates of ERM regression based on geometrically β-mixing samples, our main tool is the following lemma established by Cucker and Smale in [17].

Lemma 4.3 Let $c_1, c_2 > 0$, and $s > q > 0$[17]. Then the equation $x^s - c_1x^q - c_2 = 0$ has a unique positive zero x^*. In addition $x^* \leq \max\{(2c_1)^{1/(s-q)}, (2c_2)^{(1/s)}\}$.

Proof of Proposition 3.1 By inequality (2.2), we have

$$\mathcal{N}\Big(\mathcal{H},\frac{\varepsilon}{8M}\Big)\leq\exp\Big\{C_0\Big(\frac{\varepsilon}{8M}\Big)^{-1/r}\Big\}.$$

By inequality (3.2) of Theorem 3.1, we have that for any ε, $0<\varepsilon<3M^2/4$,

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}_m(f)-\mathcal{E}(f)|>\varepsilon\Big\}\leq 2(1+\mu e^{-2})\exp\Big\{C_0\Big(\frac{\varepsilon}{8M}\Big)^{-1/r}-\frac{m^{(\beta)}\varepsilon^2}{2M^4}\Big\}. \tag{4.9}$$

Let us rewrite inequality (4.9) in the equivalent form. We equate the right-hand side of inequality (4.9) to a positive value η $(0<\eta\leq 1)$

$$(1+\mu e^{-2})\exp\Big\{C_0\Big(\frac{\varepsilon}{8M}\Big)^{-1/r}-\frac{m^{(\beta)}\varepsilon^2}{2M^4}\Big\}=\eta.$$

It follows that

$$\varepsilon^{2+1/r}-\frac{2M^4\ln(C/\eta)}{m^{(\beta)}}\varepsilon^{1/r}-\frac{2M^4C_0(8M)^{1/r}}{m^{(\beta)}}=0,$$

where $C=1+\mu e^{-2}$. By Lemma 4.3, we can solve this equation with respect to ε. The solution is then given by

$$\varepsilon\doteq\varepsilon(m,\eta)\leq\max\Big\{2M^2\Big[\frac{\ln(C/\eta)}{m^{(\beta)}}\Big]^{1/2},\Big[\frac{4M^4C_0(8M)^{1/r}}{m^{(\beta)}}\Big]^{r/(2r+1)}\Big\}.$$

It is used further to solve inequality

$$\sup_{f\in\mathcal{H}}\mathcal{E}(f)-\mathcal{E}_m(f)\leq\varepsilon(m,\eta).$$

Then we deduce that with probability at least $1-\eta$ simultaneously for all functions in the function set $\mathcal{H}$, the inequality $\mathcal{E}(f)\leq\mathcal{E}_m(f)+\varepsilon(m,\eta)$ is valid. Since with probability at least $1-\eta$, this inequality holds for all functions in the function set $\mathcal{H}$, it holds in

particular for the function $f_{\mathbf{z}}$ that minimizes the empirical risk $\mathcal{E}_m(f)$ over $\mathcal{H}$. For this function with probability at least $1-\eta$,

$$\mathcal{E}(f_{\mathbf{z}}) \le \mathcal{E}_m(f_{\mathbf{z}}) + \varepsilon(m,\eta). \tag{4.10}$$

By Theorem 4.1, we conclude that for the same η as above, and for the function $f_{\mathcal{H}}$ that minimizes the expected risk $\mathcal{E}(f)$ over $\mathcal{H}$, the inequality

$$\mathcal{E}(f_{\mathcal{H}}) > \mathcal{E}_m(f_{\mathcal{H}}) - \varepsilon'(m,\eta) \tag{4.11}$$

holds with probability $1-\eta$, where

$$\varepsilon'(m,\eta) = M^2\sqrt{\frac{\ln(C/\eta)}{2m^{(\beta)}}}.$$

Note that

$$\mathcal{E}_m(f_{\mathcal{H}}) \ge \mathcal{E}_m(f_{\mathbf{z}}). \tag{4.12}$$

From inequalities (4.10), (4.11) and (4.12), we deduce that with probability at least $1-2\eta$, the inequality

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \le \varepsilon(m,\eta) + M^2\sqrt{\frac{\ln(C/\eta)}{2m^{(\beta)}}}$$

is valid. In addition, if

$$m^{(\beta)} \ge \max\Big\{\frac{64\ln(C/\eta)}{9}, \frac{2^{6+5/r}C_0}{3^{2+1/r}M^{1/r}}\Big\},$$

then we have $\varepsilon \le 3M^2/4$. This leads to Proposition 3.1. □

Proof of Proposition 3.2 Define $\mathrm{dist}(f_0,\mathcal{H}) = \mathrm{dist}(f_0,\mathcal{H})_{L_2(X,\mathsf{P}_x)} = \|f_0 - f_{\mathcal{H}}\|$. By inequality (2.1), we have

$$\|f_{\mathbf{z}} - f_{\mathcal{H}}\|^2 = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \le \{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_m(f_{\mathbf{z}})\} + \{\mathcal{E}_m(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}})\}.$$

By Theorems 3.1 and 4.1, we have

$$\begin{aligned}
\mathsf{P}\{\|f_{\mathbf{z}} - f_{\mathcal{H}}\|^2 \ge 2\varepsilon\} &= \mathsf{P}\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \ge 2\varepsilon\} \\
&\le \mathsf{P}\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_m(f_{\mathbf{z}}) \ge \varepsilon\} + \mathsf{P}\{\mathcal{E}_m(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \ge \varepsilon\} \\
&\le 2(1+\mu e^{-2})\mathcal{N}\Big(\mathcal{H}, \frac{\varepsilon}{8M}\Big)\exp\Big\{\frac{-m^{(\beta)}\varepsilon^2}{2M^4}\Big\}.
\end{aligned}$$

It follows that

$$\mathsf{P}\{\|f_{\mathbf{z}} - f_{\mathcal{H}}\| \ge \varepsilon\} \le 2(1+\mu e^{-2})\mathcal{N}\Big(\mathcal{H}, \frac{\varepsilon^2}{16M}\Big)\exp\Big\{\frac{-m^{(\beta)}\varepsilon^4}{8M^4}\Big\}. \tag{4.13}$$

But

$$\begin{aligned}\|f_{\mathbf{z}}-f_0\| &= \Big(\int_X (f_{\mathbf{z}}-f_0)^2 \mathrm{d}\mathsf{P}_x\Big)^{1/2} = \Big(\int_X (f_{\mathbf{z}}-f_{\mathcal{H}}+f_{\mathcal{H}}-f_0)^2 \mathrm{d}\mathsf{P}_x\Big)^{1/2}\\ &\le \|f_{\mathbf{z}}-f_{\mathcal{H}}\|+\|f_{\mathcal{H}}-f_0\|.\end{aligned}$$

Thus for any $\varepsilon>0$, if $\|f_{\mathbf{z}}-f_{\mathcal{H}}\|\le\varepsilon$, then we have that

$$\|f_{\mathbf{z}}-f_0\|\le \mathrm{dist}(f_0,\mathcal{H})+\varepsilon \tag{4.14}$$

for a set $\nu_m(\varepsilon)$ which satisfies

$$\mathsf{P}\{\mathbf{z}\in\nu_m(\varepsilon)\}\ge 1-2(1+\mu e^{-2})\mathcal{N}\Big(\mathcal{H},\frac{\varepsilon^2}{16M}\Big)\exp\Big\{\frac{-m^{(\beta)}\varepsilon^4}{8M^4}\Big\}.$$

Since in the last section, we have supposed that $\mathcal{H}$ is contained in a finite ball in $\mathcal{C}(X)$, and $\mathcal{H}$ is compact in $\mathcal{C}(X)$, then its entropy numbers $\epsilon_n(\mathcal{H})$ tend to 0 with $n\to\infty$. Thus we have that if these entropy numbers behave like (see [5]) $\epsilon_n(\mathcal{H})\le cn^{-r}$, $n=1,2,\ldots$, the covering number $\mathcal{N}(\mathcal{H},\varepsilon)\le\exp\{C_0\varepsilon^{-1/r}\}$, and $\mathrm{dist}(f_0,\mathcal{H})=0$.

By inequality (4.14), we then have $\|f_{\mathbf{z}}-f_0\|\le\varepsilon$, $\mathbf{z}\in\nu_m(\varepsilon)$. In other words, for any ε, $0<\varepsilon\le 3M^2/4$, we have

$$\mathsf{P}\{\|f_{\mathbf{z}}-f_0\|\ge\varepsilon\}\le 2(1+\mu e^{-2})\exp\Big\{C_0\Big(\frac{\varepsilon^2}{16M}\Big)^{-1/r}-\frac{m^{(\beta)}\varepsilon^4}{8M^4}\Big\}. \tag{4.15}$$

By the same argument with inequality (4.9), we can rewrite inequality (4.15) in an equivalent form. Thus we conclude that for any $\delta\in(0,1]$, with probability at least $1-\delta$, the inequality $\|f_{\mathbf{z}}-f_0\|\le\varepsilon(m,\delta)$ holds, where

$$\varepsilon(m,\delta)\le\max\Big\{2M\Big[\frac{\ln(2C/\delta)}{m^{(\beta)}}\Big]^{1/4},\Big[\frac{16^{1+1/r}C_0M^{4+1/r}}{m^{(\beta)}}\Big]^{r/(4r+2)}\Big\}.$$

In addition, if

$$m^{(\beta)}\ge\min\Big\{\frac{16^3\ln(2C/\delta)}{81M^4},\frac{4^{4+4/r}C_0}{3^{4+2/r}M^{4+3/r}}\Big\},$$

we have $\varepsilon\le 3M^2/4$. This arrives at Proposition 3.2. □

§5. Comparison and Discussion

In this section, we compare our main results with previously known results and present some useful discussions.

First, Cucker and Smale (2001), DeVore et al. (2006) established the bound on the learning rates of ERM algorithm with i.i.d. inputs for regression estimation by using Bernstein's inequality. In the last section, we obtained the bound on the learning rates

of ERM algorithm with geometrically β-mixing inputs for regression estimation by using Hoeffding's inequality. Comparing these results in [2, 5] with Theorem 3.1, we can find that if the input samples are i.i.d., the bound in Theorem 3.1 has the same convergence rate as that in [2] and [5]. Thus we extended these results of i.i.d. inputs to the case of geometrically β-mixing inputs. Since geometrically β-mixing samples usually contain less information than i.i.d. samples, it therefore might lead to worse learning rates. This property of dependent samples is just what we can expect as reflected as in our results.

Second, comparing our main results (Theorem 3.1) with the results (Theorem 6.13) obtained by Vidyasager in [6], we can find that although our proof techniques have many steps similar to that of Theorem 6.13 (or Theorem 3.4) in [6], the difference between Theorem 3.1 and the bound in Theorem 6.13 is obvious: the bound in Theorem 6.13 consists of two terms, which has not an explicit convergence rate. In this paper we introduce the sign $m^{(\beta)}$, the "effective number of observations" for geometrically β-mixing process to establish a new bound on the uniform convergence of ERM algorithm with geometrically β-mixing inputs, which consists of only one exponential term, and has an explicit convergence rate. Thus our main results improve the corresponding results in [6].

In addition,in some sense, β-mixing is a very "natural" assumption on non-i.i.d. sequences. For example, Vidyasager (2003), Meyn and Tweedie (1993) proved that if a Markov chain $\{z_i\}$ is V-geometrically ergodic, then the sequence $\{z_i\}$ is geometrically β-mixing, i.e., there exist constants B and $\lambda < 1$ such that the β-mixing coefficient $\beta(k)$ satisfies

$$\beta(k) \leq B\lambda^k \tag{5.1}$$

for all k. Moreover, the β-mixing coefficient is given by

$$\beta(k) \leq \mathsf{E}\{\rho[\mathsf{P}^k(x, A), \pi], \pi\} \leq \int_X \rho[\mathsf{P}^k(x, A), \pi]\pi(\mathrm{d}x),$$

where $\mathsf{P}^k(x, A)$ is the transition probability that the state x will belong to the set A after k time steps. π is the stationary distribution of the Markov chain $\{z_i\}$. ρ is the total variation metric between two probability measures. Especially, if a Markov chain can be described by the recursion relation $x_{t+1} = f(x_t) + e_t$ where $x_t \in \mathbb{R}^k$ for some integer k, subject to three suitable assumptions (see Theorem 3.11 in [6]), then we can define a V-function such that the Markov chain is geometrically β-mixing[6]. Moreover, Meyn and Tweedie (1994) have presented a method to compute the parameters B and λ in inequality (5.1). Thus we can obtain the parameters B and λ of geometrically β-mixing coefficient in inequality (5.1) for the Markov chain described by the above recursion relation, but other mixing sequences (e.g., α-mixing, and ϕ-mixing) have not this property of β-mixing sequence. The interested readers can consult [6] for the details. Moreover, Vidyasagar (2003) proved

that in hidden Markov models, if the underlying Markov chain has β-mixing property, then so does the corresponding hidden Markov model. Therefore, the results established in this paper are suited to geometrically β-mixing inputs are also suited to geometrically ergodic Markov chain inputs and hidden Markov models.

§6. Conclusions

In this paper, we have studied the learning rates of ERM regression with geometrically β-mixing inputs. We first established a new bound on the rate of uniform convergence of ERM algorithm with geometrically β-mixing input samples for regression estimation. Then we derived the bounds on the learning rates of ERM regression based on geometrically β-mixing inputs and proved that the ERM algorithm with geometrically β-mixing inputs for regression estimation is consistent. To our knowledge, these results here are the first explicit bounds on the rate of convergence in this topic. In order to better understand the significance and value of the established results in this paper, we compared our main results with previously known results, and concluded that the established results in this paper not only improve previously known results in [6], but also extend the results for i.i.d. samples in [2, 5] to the case of β-mixing sequence. Since a very large class of Markov chains and hidden Markov models can produce β-mixing sequences, we also shown that the results on learning rates of the ERM algorithm with geometrically β-mixing inputs for regression estimation are suited to geometrically ergodic Markov chain inputs and hidden Markov models.

Further directions of research include the question of how to control the learning rates of ERM regression with geometrically β-mixing inputs? What is the essential difference of generalization ability of ERM algorithm for regression estimation with i.i.d. samples and geometrically β-mixing samples? All these problems are under our current investigation.

References

[1] Vapnik, V., *Statistical Learning Theory*, John Wiley, New York, 1998.

[2] Cucker, F. and Smale, S., On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, **39**(2001), 1–49.

[3] Smale, S. and Zhou, D.X., Estimating the approximation error in learning theory, *Anal. Appl.*, **1**(2003), 17–41.

[4] Smale, S. and Zhou, D.X., Shannon sampling and function reconstruction from point values, *Bull. Amer. Math. Soc.*, **41**(2004), 279–305.

[5] DeVore, R., Kerkyacharian, G., Picard, D. and Temlyakov, V., Approximation methods for supervised learning, *Foundations of Computational Mathematics*, **6**(2006), 3–58.

[6] Vidyasagar, M., *Learning and Generalization with Applications to Neural Networks*, Springer, London, 2003.

[7] Steinwart, I., Hush, D. and Scovel, C., Learning from dependent observations, *Multivariate Anal.*, **100**(2009), 175–194.

[8] Smale, S. and Zhou, D.X., Online learning with Markov sampling, *Anal. Appl.*, **7**(2009), 87–113.

[9] Yu, B., Rates of convergence for empirical processes of stationary mixing sequences, *Ann. Probab.*, **22**(1994), 94–114.

[10] Nobel, A. and Dembo, A., A note on uniform laws of averages for dependent processes, *Statist. Probab. Lett.*, **17**(1993), 169–172.

[11] Karandikar, R.L. and Vidyasagar, M., Rates of uniform convergence of empirical means with mixing processes, *Statist. Probab. Lett.*, **58**(2002), 297–307.

[12] Xu, Y.L. and Chen, D.R., Learning rates of regularized regression for exponentially strongly mixing sequence, *Journal Statist. Plann.*, **138**(2008), 2180–2189.

[13] Zou, B., Li, L.Q. and Xu, Z.B., The generalization performance of ERM algorithm with strongly mixing observations, *Machine Learning*, **75**(2009), 275–295.

[14] Meyn, S.P. and Tweedie, R.L., *Markov chains and Stochastic Stability*, Springer-Verlag, 1993.

[15] Modha, S. and Masry, E., Minimum complexity regression estimation with weakly dependent observations, *IEEE Trans. Inform. Theory*, **42**(1996), 2133–2145.

[16] Hoeffding, W., Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.*, **58**(1963), 13–30.

[17] Cucker, F. and Smale, S., Best choices for regularization parameters in learning theory: on the bias-variance problem, *Found. Comput. Math.*, **2**(2002), 413–428.

[18] Meyn, S.P. and Tweedie, R.L., Computable bounds for geometric convergence rates of Markov chains, *The Annals of Applied Probability*, **4**(1994), 981–1011.

基于β-混合输入的经验风险最小化回归的学习速率

邹 斌[1,2] 徐宗本[2] 张 海[2,3]

([1]湖北大学数学与计算机科学学院, 武汉, 430062; [2]西安交通大学信息与系统科学研究所, 西安, 710049)

([3]西北大学数学系, 西安, 710069)

研究最小平方损失下的经验风险最小化算法是统计学习理论中非常重要研究内容之一. 而以往研究经验风险最小化回归学习速率的几乎所有工作都是基于独立同分布输入假设的. 然而, 独立的输入样本是一个非常强的条件. 因此, 在本文, 我们超出了独立输入样本这个经典框架来研究了基于β混合输入样本的经验风险最小化回归算法学习速率的界. 我们证明了基于β混合输入样本的经验风险最小化回归算法是一致的, 指出了本文所建立的结果同样适合输入样本是马氏链、隐马氏链的情形.

关键词: 学习速率, 经验风险最小化, β混合, 最小平方损失.

学科分类号: O234.