应用概率统计　第三十卷
第六期　2014年12月

Chinese Journal of Applied Probability
and Statistics Vol.30 No.6 Dec. 2014

New Bernstein's Inequalities for Dependent Observations and Applications to Learning Theory *

Zou Bin[1]　Tang Yuanyan[2]　Li Luoqing[1]　Xu Jie[3⋆]

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

([2]*Faculty of Science and Technology, University of Macau, Macau*)

([3]*School of Computer Science and Information Engineering, Hubei University, Wuhan, 430062*)

Abstract

The classical concentration inequalities deal with the deviations of functions of independent and identically distributed (i.i.d.) random variables from their expectation and these inequalities have numerous important applications in statistics and machine learning theory. In this paper we go far beyond this classical framework by establish two new Bernstein type concentration inequalities for β-mixing sequence and uniformly ergodic Markov chains. As the applications of the Bernstein's inequalities, we also obtain the bounds on the rate of uniform deviations of empirical risk minimization (ERM) algorithms based on β-mixing observations.

Keywords: Concentration inequality, β-mixing, Markov chains, uniform deviation, empirical risk minimization.

AMS Subject Classification: 68T01.

§1. Introduction

The laws of large numbers of classical probability theory state that sums of independent random variables are close to their expectation with a large probability, which can be applied to statistical learning problems under very mild assumptions. Such sums are the most basic examples of random variables concentrated around their expectation. In recent years some new tools and methods have been introduced making it possible to establish simple and powerful concentration inequalities such as martingale methods (see Milman and Schechtman, 1986), information-theoretic methods (see Marton, 1986), Talagrand's

*The research was supported in part by National Natural Science Foundation of China (11371007, 61370002, 61403132), Multi-Year Research of University of Macau under Grants No. MYRG205(Y1-L4)-FST11-TYY and No. MYRG187(Y1-L3))-FST11-TYY, Start-up Research of University of Macau under Grant No. SRG010-FST11-TYY and Natural Science Foundation of Hubei Province (2011CDA003).

⋆Corresponding author, E-mail: frangipani@hubu.edu.cn.

Received December 10, 2012. Revised January 27, 2013.
doi: 10.3969/j.issn.1001-4268.2014.06.002

induction method (see Talagrand, 1996), the decoupling method (see de la Peña and Giné, 1999) and various problem-specific methods (see Janson et al., 2000). These inequalities are at the heart of the mathematical analysis of various problems in machine learning and made it possible to derive new efficient algorithms. Most of these inequalities are mainly focus on the case of i.i.d. process. However, independence is a very restrictive concept (see Vidyasagar, 2003). Therefore, relaxations of such i.i.d. assumption have been considered for quite a while in both machine learning and statistics literatures. For example, Yu (1994) established the rates of convergence for empirical processes of stationary mixing sequences. Modha and Masry (1996) established the minimum complexity regression estimation with m-dependent observations and strongly mixing observations respectively. Vidyasagar (2003) considered the notions of mixing and proved that most of the desirable properties (e.g., PAC, UCEMUP) of i.i.d. sequence are preserved when the underlying sequence is mixing sequence. Kontorovich and Ramanan (2008) established the concentration inequalities for dependent random variables via the martingale method.

There are many definitions of non-independent sequences in Vidyasagar (2003), but we are only interested in β-mixing sequence and Markov chains in this paper, the reasons are as follows: First, Vidyasagar (2003) pointed out that in machine learning applications, 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 marking prediction, and Vidyasagar (2003) proved that a very large class of Markov chains and hidden Markov models (HMM) can produce β-mixing sequences. In addition, in statistical learning theory, we can use the variance of the functions to establish the better rates of convergence of learning algorithms such as support vector machine classification (SVMC) and regularization algorithms (see Chen et al., 2004). For these purposes, in this paper we first establish a new Bernstein's concentration inequality for β-mixing sequences. As the applications of the Bernstein's inequality for β-mixing sequences, we not only establish a new Bernstein's concentration inequality for Markov chains, but also apply this Bernstein's inequality for β-mixing sequence to establish the bound on the rate of uniform convergence of ERM based on β-mixing observations.

This paper is organized as follows: In Section 2 we introduce some notions and notations used in this paper. In Section 3 we present the main result obtained in this paper. As an application of the obtained main results, we obtain a new Bernstein's concentration inequality for uniformly ergodic Markov chains. In Section 4, we apply these Bernstein's

inequalities to establish the bounds on the rate of uniform convergence of ERM algorithm based on β-mixing observations.

§2. Preliminaries

In this section we introduce the definitions and notations used throughout the paper.

2.1 β-Mixing Sequence

let $Z = \{z_i\}_{i=-\infty}^{\infty}$ be a stationary real-valued stochastic process defined on the probability space $(\mathcal{Z}^\infty, \mathcal{F}^\infty, \mathsf{P})$. 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 $P_{-\infty}^k$ and P_k^∞ denote the corresponding marginal probability measures respectively. Let P_0 denote the marginal probability of each of the z_i and $\overline{\mathcal{F}}_1^{k-1}$ denote the σ-algebra generated by the random variables z_i, $i \le 0$ as well as z_j, $j \ge k$. With these notations, we can present the definition of β-mixing as follows (see Vidyasagar, 2003):

Definition 2.1 The sequence Z is called β-mixing, or completely regular, if

$$\sup_{C \in \overline{\mathcal{F}}_1^{k-1}} |P(C) - (P_{-\infty}^0 \times P_1^\infty)(C)| = \beta(k) \to 0 \qquad \text{as } k \to \infty,$$

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

Assumption 2.1 The sequence Z is called geometrically β-mixing (see Vidyasagar, 2003), if for some constants γ and $\rho < 1$, the β-mixing coefficient $\beta(k)$ satisfies $\beta(k) \le \gamma\rho^k$, $k \ge 1$.

Remark 1 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 $P_{-\infty}^0 \times P_1^\infty$. The β-mixing coefficient thus measures how nearly the product measure approximates the actual measure P. If the sequence Z consists of i.i.d. random variables, then P equals the measure $(P_0)^\infty$, which denotes the measure on $(\mathcal{Z}^\infty, \mathcal{F}^\infty)$. In such a case, the mixing coefficient $\beta(k)$ is zero for any integer k, that is, i.i.d. random variables satisfy Assumption 2.1.

2.2 Uniformly Ergodic Markov Chains

Suppose $(\mathcal{Z}, \mathcal{S})$ is a measurable space, a Markov chain is a sequence of random variables $\{Z_t\}_{t\ge1}$ together with a set of transition probability measures $P^n(A|z_i)$, $A \in \mathcal{S}$,

$z_i \in \mathcal{Z}$. It is assumed that $P^n(A|z_i) := \mathsf{P}\{Z_{n+i} \in A|Z_j, j < i, Z_i = z_i\}$. Thus $P^n(A|z_i)$ denotes the probability that the state z_{n+i} will belong to the set A after n time steps, starting from the initial state z_i at time i. It is common to denote the one-step transition probability by $P^1(A|z_i) := \mathsf{P}\{Z_{i+1} \in A|Z_j, j < i, Z_i = z_i\}$. The fact that the transition probability does not depend on the values of Z_j prior to time i is the Markov property, that is $\mathsf{P}\{Z_{n+i} \in A|Z_j, j < i, Z_i = z_i\} = \mathsf{P}\{Z_{n+i} \in A|Z_i = z_i\}$. This is commonly expressed in words as "given the present state, the future and past states are independent". Given two probabilities ν_1, ν_2 on the measure space $(\mathcal{Z}, \mathcal{S})$, we define the total variation distance between the two measures ν_1, ν_2 as follows: $\|\nu_1 - \nu_2\|_{\mathrm{TV}} := \sup_{A\in\mathcal{S}} |\nu_1(A) - \nu_2(A)|$. Thus we have the following definition of uniformly ergodic Markov chain (see Vidyasagar, 2003).

Definition 2.2　A Markov chain $\{Z_t\}_{t\geq 1}$ is said to be uniformly ergodic if

$$\|P^n(\cdot|z) - \pi(\cdot)\|_{\mathrm{TV}} \leq \gamma_1 \rho_1^n, \qquad \forall\, n \geq 1$$

for some $\gamma_1 < \infty$ and $\rho_1 < 1$, where $\pi(\cdot)$ is the stationary distribution of $\{Z_t\}_{t\geq 1}$.

A weaker condition than uniformly ergodic is V-geometrically ergodic (see Definition 3.5.1 of Vidyasagar, 2003). The difference between V-geometrically ergodic and uniformly ergodic is that here the constant γ_1 in Definition 2.2 is not depend on the initial state z.

§3. Two New Bernstein's Inequalities

In this section, we establish two new Bernstein's inequalities for β-mixing sequence and uniformly ergodic Markov chains. Our main tools are the following three useful lemmas.

Lemma 3.1 (Vidyasagar, 2003)　Suppose $i_0 < i_1 < \cdots < i_l$ are integers, and define $k = \min\limits_{0\leq j\leq l-1} i_{j+1} - i_j$. Suppose g is essentially bounded and depends only on $z_{i_0}, z_{i_1}, \ldots, z_{i_l}$. Then $|\mathsf{E}(g, P) - \mathsf{E}(g, P_0^\infty)| \leq l\beta(k)\|g\|_\infty$, where $\mathsf{E}(g, P)$ and $\mathsf{E}(g, P_0^\infty)$ are the expectation values of g with respect to P and P_0^∞ respectively.

Lemma 3.2 (Craig, 1933)　Let W be a random variable such that $\mathsf{E}(W) = 0$, and W satisfies the Bernstein moment condition, that is, for some $K_1 > 0$,

$$\mathsf{E}|W|^k \leq \frac{\mathsf{Var}\,(W)}{2} k! K_1^{k-2} \tag{3.1}$$

for all $k \geq 2$. Then, for all $0 < \zeta < 1/K_1$, $\mathsf{E}[\exp(\zeta W)] \leq \exp[\zeta^2 \mathsf{E}|W|^2/(2(1 - \zeta K_1))]$.

Remark 2　If $|W| \leq 3K_1$ almost everywhere, then the Bernstein moment condition (3.1) holds true (see Modha and Masry, 1996).

Lemma 3.3 (Vidyasagar, 2003) Let $\{Z_t\}_{t\geq 1}$ be a Markov chain V-geometrically ergodic. Then the sequence $\{Z_t\}_{t\geq 1}$ is geometrically β-mixing, and the β-mixing coefficient $\beta(k)$ is given by

$$\beta(k) = \mathsf{E}\{\|P^k(\cdot|\xi) - \pi(\cdot)\|_{\mathrm{TV}}, \pi\} = \int \|P^k(\cdot|\xi) - \pi(\cdot)\|_{\mathrm{TV}}\pi(\mathrm{d}\xi).$$

By these lemmas, we first establish the following new Bernstein's concentration inequality for β-mixing sequence.

Theorem 3.1 Let $\mathcal{Z} = \{z_i\}_{i=-\infty}^{\infty}$ be a stationary β-mixing sequence with the mixing coefficient satisfying Assumption 2.1. Suppose that ξ is a random variable on the probability space $(\mathcal{Z}^\infty, \mathcal{F}^\infty, \mathsf{P})$ with mean $\mathsf{E}(\xi) = \mu$ and variance $\sigma^2(\xi) = \sigma^2$. Set $m^{(\beta)} = \lfloor m\lceil\{8m/\ln(1/\rho)\}^{1/2}\rceil^{-1}\rfloor$, where m denotes the number of observations and $\lfloor u\rfloor(\lceil u\rceil)$ denotes the greatest (least) integer less (greater) than or equal to u. If $|\xi(z) - \mu| \leq B$ for almost all $z \in \mathcal{Z}$, then for any $\varepsilon > 0$,

$$\mathsf{P}\Big\{\Big|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\Big| \geq \varepsilon\Big\} \leq 2(1+\gamma \mathrm{e}^{-2})\exp\Big\{\frac{-\varepsilon^2 m^{(\beta)}}{2(\sigma^2 + \varepsilon B/3)}\Big\}. \tag{3.2}$$

Proof We decompose the proof into three steps.

Step 1: To exploit the β-mixing property, we decompose the index set $I = \{1, 2, \ldots, m\}$ into different parts by following the idea from Vidyasagar (2003), that is, given an integer m, choose any integer $k_m \leq m$, 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. Let $r = m - kl$, and define

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

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

$$T_i = \xi(z_i) - \mu, \qquad a_m(z) = \frac{1}{m}\sum_{i=1}^{m} T_i, \qquad b_i(z) = \frac{1}{|I_i|}\sum_{j\in I_i} T_j.$$

Then we have

$$\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu = a_m(z) = \sum_{i=1}^{k} p_i b_i(z).$$

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

$$\exp[s a_m(z)] = \exp\Big[\sum_{i=1}^{k} p_i s b_i(z)\Big] \leq \sum_{i=1}^{k} p_i \exp[s b_i(z)].$$

We have

$$\mathsf{E}(\mathrm{e}^{sa_m(z)}, \widetilde{P}) \leq \sum_{i=1}^{k} p_i \mathsf{E}(\mathrm{e}^{sb_i(z)}, \widetilde{P}). \tag{3.3}$$

Since

$$\exp[sb_i(z)] = \exp\Big[\frac{s}{|I_i|}\sum_{j\in I_i} T_j\Big] = \prod_{j\in I_i} \exp\Big(\frac{sT_j}{|I_i|}\Big) \leq \Big[\exp\Big(\frac{sB}{|I_i|}\Big)\Big]^{|I_i|} \leq \mathrm{e}^{sB},$$

where in the last step we use the assumption $|T_j| = |\xi(z_j) - \mu| \leq B$. By Lemma 3.1, we have that for any $s > 0$,

$$\begin{aligned}\mathsf{E}(\mathrm{e}^{sb_i(z)}, \widetilde{P}) &\leq (|I_i| - 1)\beta(k)\|\mathrm{e}^{sb_i(z)}\|_\infty + \mathsf{E}(\mathrm{e}^{sb_i(z)}, \widetilde{P}_0^\infty) \\ &\leq (|I_i| - 1)\beta(k)\|\mathrm{e}^{sB} + \mathsf{E}(\mathrm{e}^{sb_i(z)}, \widetilde{P}_0^\infty). \end{aligned}\tag{3.4}$$

Since under the measure $\widetilde{P}_0^\infty$, the various z_i are independent, using Lemma 3.2, we have that for any $0 < s < 3|I_i|/B$,

$$\begin{aligned}\mathsf{E}(\mathrm{e}^{sb_i(z)}, \widetilde{P}_0^\infty) &= \mathsf{E}\Big[\prod_{j\in I_i} \exp(sT_j/|I_i|), \widetilde{P}_0^\infty\Big] = \prod_{j\in I_i} \mathsf{E}[\exp(sT_j/|I_i|), \widetilde{P}_0^\infty] \\ &= \Big\{\mathsf{E}\Big[\exp\Big(\frac{sT_1}{|I_i|}\Big), \widetilde{P}_0\Big]\Big\}^{|I_i|} \leq \exp\Big[\frac{s^2\mathsf{E}|T_1|^2}{2|I_i|(1 - sB/3|I_i|)}\Big].\end{aligned}$$

By inequality (3.4), we have that for any $3|I_i|/B > s > 0$,

$$\mathsf{E}(\mathrm{e}^{sb_i(z)}, \widetilde{P}) \leq \exp\Big[\frac{s^2\mathsf{E}|T_1|^2}{2|I_i|(1 - sB/3|I_i|)}\Big] + (|I_i| - 1)\beta(k)\mathrm{e}^{sB}.$$

Thus by inequality (3.3) and the inequality above, we have that for any $3|I_i|/B > s > 0$,

$$\mathsf{E}(\mathrm{e}^{sa_m(z)}, \widetilde{P}) \leq \sum_{i=1}^{k} p_i\Big\{\exp\Big[\frac{s^2\mathsf{E}|T_1|^2}{2|I_i|(1 - sB/3|I_i|)}\Big] + (|I_i| - 1)\beta(k)\mathrm{e}^{sB}\Big\}. \tag{3.5}$$

Step 2: We now bound the second term on the right-hand side of inequality (3.5) which is denoted henceforth by ϕ. By Assumption 2.1, we have that for any $0 < s \leq 3|I_i|/B$,

$$\begin{aligned}\phi = &\exp\Big[\frac{s^2\mathsf{E}|T_1|^2}{2|I_i|(1 - sB/3|I_i|)}\Big] + (|I_i| - 1)\beta(k)\mathrm{e}^{sB} \\ \leq &\exp\Big[\frac{s^2\mathsf{E}|T_1|^2}{2|I_i|(1 - sB/3|I_i|)}\Big] + \mathrm{e}^{|I_i|}\mathrm{e}^{-2}\gamma\rho^k \cdot \mathrm{e}^{sB} \\ \leq &\exp\Big[\frac{s^2\mathsf{E}|T_1|^2}{2|I_i|(1 - sB/3|I_i|)}\Big] + \gamma\mathrm{e}^{-2}\exp\{k\ln(\rho) + 4|I_i|\}.\end{aligned}$$

The above inequality follows from the fact that $|I_i - 1| \leq \mathrm{e}^{|I_i|-2}$ for $|I_i| \geq 2$. We require $\exp\{k\ln(\rho) + 4|I_i|\} \leq 1$. But $|I_i| \leq (m/k + 1)$, thus the bound holds if $4(m/k +$

1) $\le k\ln(1/\rho)$ or $4(m+k) \le k^2\ln(1/\rho)$. Since $m+k \le 2m$, then the bound holds if $\{8m/\ln(1/\rho)\}^{1/2} \le k$. Let $k = \lceil\{8m/\ln(1/\rho)\}^{1/2}\rceil$. Since for all $i = 1, 2, \ldots, k$, $|I_i| \ge l$, and $l = \lfloor m/k\rfloor$, we have

$$\phi \le \exp\Big[\frac{s^2\mathsf{E}|T_1|^2}{2l(1-sB/3l)}\Big] + \gamma\mathrm{e}^{-2}. \tag{3.6}$$

Since inequality (3.6) is true for all s, $0 < s \le 3|I_i|/B$. To make the constraint uniform over all i, we then require s satisfy $0 < s < 3l/B \le 3|I_i|/B$. Since $s^2\mathsf{E}|T_1|^2/[2l(1-sB/3l)] > 0$, we have that for any $0 < s < 3l/B$,

$$\phi \le (1+\gamma\mathrm{e}^{-2})\exp\Big[\frac{s^2\mathsf{E}|T_1|^2}{2l(1-sB/3l)}\Big].$$

Returning to inequality (3.5), we have that for any $0 < s < 3l/B$,

$$\mathsf{E}(\mathrm{e}^{sa_m(z)}, \widetilde{P}) \le (1+\gamma\mathrm{e}^{-2})\exp\Big[\frac{s^2\mathsf{E}|T_1|^2}{2l(1-sB/3l)}\Big]. \tag{3.7}$$

Step 3: By Markov's inequality and inequality (3.7), we have that for any $0 < s \le 3l/B$,

$$\begin{aligned}\mathsf{P}\Big\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu \ge \varepsilon\Big\} &= \mathsf{P}\Big\{\mathrm{e}^{s\left[m^{-1}\sum\limits_{i=1}^m \xi(z_i)-\mu\right]} \ge \mathrm{e}^{s\varepsilon}\Big\}\\ &\le \frac{\mathsf{E}\Big\{\mathrm{e}^{s\left[m^{-1}\sum\limits_{i=1}^m \xi(z_i)-\mu\right]}\Big\}}{\mathrm{e}^{s\varepsilon}}\\ &\le (1+\gamma\mathrm{e}^{-2})\exp\Big\{-s\varepsilon + \frac{s^2\mathsf{E}|T_1|^2}{2l(1-sB/3l)}\Big\}.\end{aligned}$$

Substituting $s = l\varepsilon/(\mathsf{E}|T_1|^2 + \varepsilon B/3)$ and noting that the selected value for s satisfies $s \le 3l/B$, then we have that for any $\varepsilon > 0$,

$$\mathsf{P}\Big\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu \ge \varepsilon\Big\} \le (1+\gamma\mathrm{e}^{-2})\exp\Big\{\frac{-l\varepsilon^2}{2(\mathsf{E}|T_1|^2 + \varepsilon B/3)}\Big\}.$$

By symmetry, we also have that for any $\varepsilon > 0$,

$$\mathsf{P}\Big\{\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i) \ge \varepsilon\Big\} \le (1+\gamma\mathrm{e}^{-2})\exp\Big\{\frac{-l\varepsilon^2}{2(\mathsf{E}|T_1|^2 + \varepsilon B/3)}\Big\}.$$

By these inequalities above and replacing l and $\mathsf{E}|T_1|^2$ by $m^{(\beta)}$ and σ^2 respectively, we can complete the Proof of Theorem 3.1. □

Remark 3 Since inequality (3.2) in Theorem 3.1 contains the information of variance of random variables, inequality (3.2) is a Bernstein type concentration inequality for β-mixing sequence. To our knowledge, this inequality in Theorem 3.1 is the first Bernstein type inequality for β-mixing sequence in this topic. $m^{(\beta)}$ in Theorem 3.1 is called the "effective number of observations" for the β-mixing processes. From Theorem 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 (see Cucker and Smale, 2002a). In particular, if $\mathcal{Z}$ is i.i.d., according to Remark 1, we take $\gamma = 0$ in Theorem 3.1 and ignore the multiplicative constant $1+\gamma \mathrm{e}^{-2}$, thus by Theorem 3.1, we can recover the classical Bernstein's inequality (see Cucker and Smale, 2002a) for sums of independent random variables.

As an application of the Bernstein's inequality for β-mixing sequence, we establish a new Bernstein's inequality for uniformly ergodic Markov chains.

Theorem 3.2 Let $\{z_i\}_{i=1}^m$ be a uniformly ergodic Markov chain. Suppose that ξ is a random variable on a probability space with mean $\mathsf{E}(\xi) = \mu$ and variance $\sigma^2(\xi) = \sigma^2$. Set $\widetilde{m}^{(\beta)} = \lfloor m \lceil \{8m/\ln(1/\rho_1)\}^{1/2} \rceil^{-1} \rfloor$. If $|\xi(z) - \mu| \le B$ for any $z \in \mathcal{Z}$, then for any $\varepsilon > 0$,

$$\mathsf{P}\Big\{\Big|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\Big| \ge \varepsilon\Big\} \le 2(1+\gamma_1 \mathrm{e}^{-2}) \exp\Big\{\frac{-\varepsilon^2 \widetilde{m}^{(\beta)}}{2(\sigma^2 + \varepsilon B/3)}\Big\}. \tag{3.8}$$

Proof By Definition 2.2 and Lemma 3.3, we have $\beta(k) = \mathsf{E}\{\|P^k(\cdot|\xi) - \pi(\cdot)\|_{\mathrm{TV}}, \pi\} \le \gamma_1 \rho_1^k$. Then replacing γ and ρ by γ_1 and ρ_1 in Theorem 3.1, respectively, we can finish the Proof of Theorem 3.2. □

Remark 4 In Theorem 3.2, we established a new Bernstein's inequality for uniformly ergodic Markov chains. To our knowledge, this inequality here is the first Bernstein type inequality for uniformly ergodic Markov chains in this topic.

§4. Applications to Learning Theory

In this section, we apply the Bernstein's inequality to study the bounds on the rate of uniform convergence of ERM algorithm and the generalization bounds of ERM algorithm. Denote by $\boldsymbol{z} = \{z_1, z_2, \ldots, z_m\}$ the sample set of size m observations drawn from the β-mixing sequence $\mathcal{Z}$. The goal of machine learning from random sampling is to find a function f that assigns values to objects such that if new objects are given, the function f will forecast them correctly. Let $\mathcal{E}(f) = \mathsf{E}[\ell(f,z)] = \int \ell(f,z)\mathrm{d}P$ be the expected risk of function f, where the function $\ell(f,z)$, which is integrable for any f and depends on f and z, called loss function. A learning task is to find the minimizer of the expected risk

$\mathcal{E}(f)$ over a given hypothesis space $\mathcal{H}$. Since one knows only the set $\boldsymbol{z}$ of samples instead of the distribution P, the minimizer of the expected risk $\mathcal{E}(f)$ can not be computed directly. According to the principle of ERM (see Vapnik, 1998), we minimize, instead of the expected risk $\mathcal{E}(f)$, the so called empirical risk $\mathcal{E}_m(f) = m^{-1}\sum_{i=1}^{m}\ell(f, z_i)$. Let $f_{\mathcal{H}}$ be a function minimizing the expected risk $\mathcal{E}(f)$ over $f \in \mathcal{H}$, i.e.,

$$f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}}\mathcal{E}(f) = \arg\min_{f\in\mathcal{H}}\int \ell(f, z)\mathrm{d}P.$$

We define the empirical target function $f_{\boldsymbol{z}}$ to be a function minimizing the empirical risk $\mathcal{E}_m(f)$ over $f \in \mathcal{H}$, i.e., $f_{\boldsymbol{z}} = \arg\min_{f\in\mathcal{H}}\mathcal{E}_m(f) = \arg\min_{f\in\mathcal{H}} m^{-1}\sum_{i=1}^{m}\ell(f, z_i)$. According to the principle of ERM, we shall consider the function $f_{\boldsymbol{z}}$ as an approximation of the target function $f_{\mathcal{H}}$. Thus a central question of ERM learning algorithm is how well $f_{\boldsymbol{z}}$ really approximate $f_{\mathcal{H}}$. If it is well, the ERM algorithm is said to be of generalization ability. To characterize generalization capability of a learning algorithm requires in essence to decipher how close $f_{\boldsymbol{z}}$ is from $f_{\mathcal{H}}$, which is a very difficult issue in general (see Vapnik, 1998). In framework of statistical learning, however, this is then relaxed to considering how close the expected risk $\mathcal{E}(f_{\boldsymbol{z}})$ is from $\mathcal{E}(f_{\mathcal{H}})$, or equivalently, how small the excess risk $\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(f_{\mathcal{H}})$. In order to bound the excess risk, we should first estimate the bound on the rate of the empirical risks uniform convergence to their expected risk on a given set $\mathcal{H}$, that is, for any $\varepsilon > 0$, we should to bound the uniform convergence bound $\mathsf{P}\big\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f) - \mathcal{E}_m(f)| > \varepsilon\big\}$. Since the uniform convergence bound is valid for all function of set $\mathcal{H}$, we have to regulate the capacity of the function set $\mathcal{H}$. Here the capacity is measured by the covering number (see Cucker and Smale, 2002a). We present the definition of covering number, some assumptions and lemma as follows:

Definition 4.1 For a subset $\mathcal{F}$ of a metric space and $\epsilon > 0$, the covering number $\mathcal{N}(\mathcal{F}, \epsilon)$ of the function set $\mathcal{F}$ is the minimal $n \in \mathbb{N}$ such that there exist n disks in $\mathcal{F}$ with radius ϵ covering $\mathcal{F}$.

We give some basic assumptions on the space $\mathcal{H}$ and the loss function $\ell(f, z)$:

(i) Assumption on the hypothesis space: We suppose that $\mathcal{H}$ is contained in a ball of a Hölder space $\mathcal{C}^p$ on a compact subset of a Euclidean space $\mathbb{R}^d$ for some $p > 0$. Then there exists constant $C_0 > 0$ such that

$$\mathcal{N}(\mathcal{H}, \varepsilon) \le \exp\{C_0\varepsilon^{-2d/p}\}. \tag{4.1}$$

(ii) Assumption on the loss function: We define

$$M = \sup_{f\in\mathcal{H}}\max_{z\in\mathcal{Z}}|\ell(f, z)|, \qquad L = \sup_{g_1,g_2\in\mathcal{H}}\max_{z\in\mathcal{Z}}\frac{|\ell(g_1, z) - \ell(g_2, z)|}{|g_1 - g_2|}.$$

We assume that M and L are finite. Note that $f_{\boldsymbol{z}}$ is dependent on the sample $\boldsymbol{z}$, and its existence follows from the compactness of $\mathcal{H}$.

Lemma 4.1 (Cucker and Smale, 2002b)　Let $c_1, c_2 > 0$, and $s_1 > q_1 > 0$. Then the equation $x^{s_1} - c_1 x^{q_1} - c_2 = 0$ has a unique positive zero x^*. In addition $x^* \leq \max\{(2c_1)^{1/(s_1-q_1)}, (2c_2)^{(1/s_1)}\}$.

Theorem 4.1　Let $\mathcal{Z}$ be a stationary β-mixing sequence with the mixing coefficient satisfying Assumption 2.1. Then for any $\varepsilon > 0$,

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq\varepsilon\Big\}\leq 2(1+\gamma\mathrm{e}^{-2})\mathcal{N}\Big(\mathcal{H},\frac{\varepsilon}{4L}\Big)\exp\Big\{\frac{-\varepsilon^2 m^{(\beta)}}{8(\sigma^2+\varepsilon M/6)}\Big\}. \tag{4.2}$$

Proof　Let $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots \cup \mathcal{H}_{n_1}$, $n_1 \in \mathbb{N}$, $L_{\boldsymbol{z}}(f) = \mathcal{E}(f) - \mathcal{E}_m(f)$ 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

$$\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon \iff \exists\, k,\ 1\leq k\leq n_1,\ \text{s.t.}\ \sup_{f\in\mathcal{H}_k}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon. \tag{4.3}$$

By the equivalence (4.3), 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

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\Big\}\leq\sum_{k=1}^{n_1}\mathsf{P}\Big\{\sup_{f\in\mathcal{H}_k}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\Big\}. \tag{4.4}$$

Now we estimate the term on the right-hand side of inequality (4.4). 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 $\varepsilon/(2L)$. Then, for all $\boldsymbol{z} \in \mathcal{Z}^m$ and all $f \in D_k$,

$$\begin{aligned}|L_{\boldsymbol{z}}(f)-L_{\boldsymbol{z}}(f_k)| &\leq |\mathcal{E}(f)-\mathcal{E}(f_k)|+|\mathcal{E}_m(f)-\mathcal{E}_m(f_k)|\\ &\leq 2L\cdot\|f-f_k\|_\infty\leq 2L\cdot\varepsilon/2L=\varepsilon.\end{aligned}$$

It follows that for any $\boldsymbol{z} \in \mathcal{Z}^m$ and all $f \in D_k$, $\sup\limits_{f\in D_k}|L_{\boldsymbol{z}}(f)| \geq 2\varepsilon \Longrightarrow |L_{\boldsymbol{z}}(f_k)| \geq \varepsilon$. We thus conclude that for any $k \in \{1, 2, \ldots, n_1\}$,

$$\mathsf{P}\Big\{\sup_{f\in D_k}|L_{\boldsymbol{z}}(f)|\geq 2\varepsilon\Big\}\leq\mathsf{P}\{|L_{\boldsymbol{z}}(f_k)|\geq\varepsilon\}. \tag{4.5}$$

By Theorem 3.1, we can get

$$\mathsf{P}\{|L_{\boldsymbol{z}}(f_k)|\geq\varepsilon\}\leq 2(1+\gamma\mathrm{e}^{-2})\exp\Big\{\frac{-\varepsilon^2 m^{(\beta)}}{2(\sigma^2+\varepsilon M/3)}\Big\}.$$

By inequalities (4.4) and (4.5), we have that for any $\varepsilon > 0$,

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\Big\}\leq 2(1+\gamma \mathrm{e}^{-2})\mathcal{N}\Big(\mathcal{H},\frac{\varepsilon}{2L}\Big)\exp\Big\{\frac{-\varepsilon^2 m^{(\beta)}}{2(\sigma^2+\varepsilon M/3)}\Big\}. \tag{4.6}$$

Theorem 4.1 thus follows from inequality (4.6) by replacing ε by $\varepsilon/2$. Then we finish the proof of Theorem 4.1. □

Remark 5 Since $m^{(\beta)}\to\infty$ as $m\to\infty$, by Theorem 4.1, we have that as long as the covering number of the space $\mathcal{H}$ is finite, the empirical risks $\mathcal{E}_m(f)$ can uniformly converge to their expected risks $\mathcal{E}(f)$, and the convergence speed may be exponential. This assertion is well known for the ERM algorithm with i.i.d. samples (see Cucker and Smale, 2002a; Vapnik, 1998). In particular, if $\mathcal{Z}$ is i.i.d., according to Remark 3, we take $\gamma = 0$ in Theorem 4.1 and ignore the multiplicative constant $1+\gamma \mathrm{e}^{-2}$, we can recover the classical results for i.i.d. samples in Cucker and Smale (2002a).

Now we can apply Theorem 4.1 to obtain the bounds on the generalization ability of ERM algorithm based on β-mixing observations: By assumption (4.1), and using Theorem 4.1 we have that for any $0<\varepsilon<M$,

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq \varepsilon\Big\}\leq 2(1+\gamma \mathrm{e}^{-2})\exp\Big\{C_0\Big(\frac{\varepsilon}{4L}\Big)^{-2d/p}-\frac{\varepsilon^2 m^{(\beta)}}{2(\sigma^2+M^2/3)}\Big\}.$$

Let us rewrite the above inequality in the equivalent form. We equate the right-hand side of the above inequality to a positive value δ $(0<\delta<1)$

$$(1+\gamma \mathrm{e}^{-2})\exp\Big\{C_0\Big(\frac{\varepsilon}{4L}\Big)^{-2d/p}-\frac{\varepsilon^2 m^{(\beta)}}{2(\sigma^2+M^2/3)}\Big\}=\delta.$$

It follows that

$$\varepsilon^{2+2d/p}-\frac{2(\sigma^2+M^2/3)\ln[(1+\gamma \mathrm{e}^{-2})/\eta]}{m^{(\beta)}}\varepsilon^{2d/p}-\frac{2C_0(4L)^{2d/p}(\sigma^2+M^2/3)}{m^{(\beta)}}=0.$$

By Lemma 4.1, we can solve this equation with respect to ε. The solution is then given by

$$\begin{aligned}&\varepsilon\doteq\varepsilon(m,\delta)\\&\leq 2\max\Big\{\Big[\frac{(\sigma^2+M^2/3)\ln[(1+\gamma \mathrm{e}^{-2})/\delta]}{m^{(\beta)}}\Big]^{1/2},\Big[\frac{C_0(4L)^{2d/p}(\sigma^2+M^2/3)}{m^{(\beta)}}\Big]^{p/(2p+2d)}\Big\}.\end{aligned}$$

Then we deduce that for the function $f_{\boldsymbol{z}}$ that minimizes the empirical risk $\mathcal{E}_m(f)$ over $\mathcal{H}$, with probability at least $1-\delta$, the following inequality is holds true

$$\mathcal{E}(f_{\boldsymbol{z}})\leq\mathcal{E}_m(f_{\boldsymbol{z}})+\varepsilon(m,\delta). \tag{4.7}$$

In addition, by Theorem 3.1, we have that for any ε, $M > \varepsilon > 0$

$$\mathsf{P}\{|\mathcal{E}(f) - \mathcal{E}_m(f)| \geq \varepsilon\} \leq 2(1 + \gamma \mathrm{e}^{-2}) \exp\left\{\frac{-\varepsilon^2 m^{(\beta)}}{2(\sigma^2 + M^2/3)}\right\}.$$

Then we conclude that for the same δ as above, and for the function $f_{\mathcal{H}}$ that minimizes the expected risk $\mathcal{E}(f)$ over $\mathcal{H}$, the following inequality holds with probability $1 - \delta$,

$$\mathcal{E}(f_{\mathcal{H}}) > \mathcal{E}_m(f_{\mathcal{H}}) + \sqrt{\frac{2(\sigma^2 + M^2/3)\ln((1 + \gamma \mathrm{e}^{-2})/\delta)}{m^{(\beta)}}}. \tag{4.8}$$

Note that $\mathcal{E}_m(f_{\mathcal{H}}) \geq \mathcal{E}_m(f_{\boldsymbol{z}})$, and by inequalities (4.7) and (4.8), we deduce that with probability at least $1 - 2\delta$, the inequality

$$\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq \varepsilon(m, \delta) + \sqrt{\frac{2(\sigma^2 + M^2/3)\ln((1 + \gamma \mathrm{e}^{-2})/\delta)}{m^{(\beta)}}} \tag{4.9}$$

is valid provided that $m^{(\beta)}$ satisfies

$$m^{(\beta)} \geq \max\left\{\frac{4(\sigma^2 + M^2/3)\ln[(1 + \gamma \mathrm{e}^{-2})/\delta]}{M^2}, \frac{2^{(2p+2d)/p} C_0 (4L)^{2d/p}(\sigma^2 + M^2/3)}{M^{2+2d/p}}\right\}.$$

Remark 6　Bounds (4.7) and (4.9) describe the generalization performance of the ERM algorithm based on β-mixing observations on the given function set $\mathcal{H}$. Different from that results in Zou et al. (2011) for β-mixing observations, the generalization bounds (4.7) and (4.9) are based on the Bernstein's inequality for β-mixing sequence, in other words, these generalization bounds (4.7) and (4.9) contain the information of variance of β-mixing observations.

However, inequality (4.2) in Theorem 4.1 fails to capture the phenomenon that for those functions $f \in \mathcal{H}$ for which the expected risk $\mathcal{E}(f)$ is small, the deviation $\mathcal{E}(f) - \mathcal{E}_m(f)$ is also small with large probability (see Vapnik, 1998). As an application of Theorem 3.1, we also establish the bound on the relative uniform convergence bound for β-mixing mixing sequence. Our result can be stated as follows:

Theorem 4.2　With all notations as in Theorem 4.1. Then for $\varepsilon > 0$ and $0 < \alpha \leq 1$,

$$\mathsf{P}\left\{\sup_{f \in \mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_m(f)}{\sqrt{\mathcal{E}(f) + \varepsilon}} > 4\alpha\sqrt{\varepsilon}\right\} \leq \mathcal{N}(\mathcal{H}, \alpha\varepsilon)(1 + \gamma \mathrm{e}^{-2}) \exp\left\{\frac{-3\alpha^2 \varepsilon m^{(\beta)}}{8M}\right\}.$$

Proof　By Theorem 3.1, we have that for $\varepsilon > 0$ and $0 < \alpha \leq 1$,

$$\mathsf{P}\left\{\frac{\mathcal{E}(f) - \mathcal{E}_m(f)}{\sqrt{\mathcal{E}(f) + \varepsilon}} > \alpha\sqrt{\varepsilon}\right\} \leq (1 + \gamma \mathrm{e}^{-2}) \exp\left\{\frac{-\alpha^2 \varepsilon(\mathcal{E}(f) + \varepsilon) m^{(\beta)}}{2(\sigma^2 + M\alpha\sqrt{\varepsilon}\sqrt{\mathcal{E}(f) + \varepsilon}/3)}\right\}. \tag{4.10}$$

Here $\sigma^2 \le \mathsf{E}[(\ell(f,z))^2] \le M\mathcal{E}(f)$, since $0 \le \mathcal{E}(f) \le M$. Then we have that

$$\sigma^2 + M\alpha\sqrt{\varepsilon}\sqrt{\mathcal{E}(f)+\varepsilon}/3 \le M\mathcal{E}(f) + M(\mathcal{E}(f)+\varepsilon)/3 \le 4M(\mathcal{E}(f)+\varepsilon)/3.$$

By inequality (4.10), we have

$$\mathsf{P}\Big\{\frac{\mathcal{E}(f)-\mathcal{E}_m(f)}{\sqrt{\mathcal{E}(f)+\varepsilon}} > \alpha\sqrt{\varepsilon}\Big\} \le (1+\gamma \mathrm{e}^{-2})\exp\Big\{\frac{-3\alpha^2\varepsilon m^{(\beta)}}{8M}\Big\}. \tag{4.11}$$

We denote the covering number $(\mathcal{H}, \alpha\varepsilon)$ by n_2, then there exist n_2 disks $\{D_j\}_{j=1}^{n_2}$ covering $\mathcal{H}$, for which $D_j = \{f \in \mathcal{H} : \|f - f_j\|_\infty \le \alpha\varepsilon\}$. We have that for any $f \in D_j$,

$$\begin{aligned}
\mathsf{P}\Big\{\sup_{f\in D_j}\frac{\mathcal{E}(f)-\mathcal{E}_m(f)}{\sqrt{\mathcal{E}(f)+\varepsilon}} > 4\alpha\sqrt{\varepsilon}\Big\} &= \mathsf{P}\Big\{\sup_{f\in D_j}\frac{\mathcal{E}(f)-\mathcal{E}_m(f)}{\sqrt{\mathcal{E}(f)+\varepsilon}} + S_1 + S_2 > 4\alpha\sqrt{\varepsilon}\Big\} \\
&\le \mathsf{P}\Big\{\frac{\mathcal{E}(f_j)-\mathcal{E}_m(f_j)}{\sqrt{\mathcal{E}(f_j)+\varepsilon}} > \alpha\sqrt{\varepsilon}\Big\} \\
&\le (1+\gamma \mathrm{e}^{-2})\exp\Big\{\frac{-3\alpha^2\varepsilon m^{(\beta)}}{8M}\Big\},
\end{aligned}$$

where

$$S_1 = \frac{\mathcal{E}(f_j)-\mathcal{E}_m(f_j)}{\sqrt{\mathcal{E}(f)+\varepsilon}}, \qquad S_2 = \frac{\mathcal{E}_m(f_j)-\mathcal{E}_m(f)}{\sqrt{\mathcal{E}(f)+\varepsilon}}.$$

We finish the proof of Theorem 4.2 by summing up the inequalities and noting the fact

$$\mathsf{P}\Big\{\sup_{f\in\mathcal{H}}\frac{\mathcal{E}(f)-\mathcal{E}_m(f)}{\sqrt{\mathcal{E}(f)+\varepsilon}} > 4\alpha\sqrt{\varepsilon}\Big\} \le \sum_{j=1}^{n_2}\mathsf{P}\Big\{\sup_{f\in D_j}\frac{\mathcal{E}(f)-\mathcal{E}_m(f)}{\sqrt{\mathcal{E}(f)+\varepsilon}} > 4\alpha\sqrt{\varepsilon}\Big\}. \qquad \square$$

Now we apply Theorem 4.2 to obtain the bounds on the generalization ability of ERM algorithm based on β-mixing observations: Taking $\alpha = 1/4$ and the fact that $\sqrt{\varepsilon}\sqrt{\mathcal{E}(f)+\varepsilon} \le \mathcal{E}(f)/2 + \varepsilon$, by Theorem 4.2, we have that for any $\varepsilon > 0$,

$$\mathsf{P}\Big\{\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}_m(f_{\boldsymbol{z}}) > \frac{1}{2}\mathcal{E}(f_{\boldsymbol{z}}) + \varepsilon\Big\} \le \mathcal{N}(\mathcal{H}, \varepsilon/4)(1+\gamma \mathrm{e}^{-2})\exp\Big\{\frac{-3\varepsilon m^{(\beta)}}{128M}\Big\}.$$

By assumption (4.1), we have

$$\mathsf{P}\{\mathcal{E}(f_{\boldsymbol{z}}) - 2\mathcal{E}_m(f_{\boldsymbol{z}}) > 2\varepsilon\} \le (1+\gamma \mathrm{e}^{-2})\exp\Big\{C_0\Big(\frac{\varepsilon}{4}\Big)^{-2d/p} - \frac{3\varepsilon m^{(\beta)}}{128M}\Big\}.$$

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

$$(1+\gamma \mathrm{e}^{-2})\exp\Big\{C_0\Big(\frac{\varepsilon}{4}\Big)^{-2d/p} - \frac{3\varepsilon m^{(\beta)}}{128M}\Big\} = \eta.$$

It follows that

$$\varepsilon^{1+2d/p} - \frac{128M\ln[(1+\gamma e^{-2})/\eta]}{m^{(\beta)}}\varepsilon^{2d/p} - \frac{128MC_0 4^{2d/p}}{m^{(\beta)}} = 0.$$

By Lemma 4.1, we can solve this equation with respect to ε. The solution is then given by

$$\varepsilon \doteq \varepsilon(m,\eta) \le 4\max\left\{\left[\frac{64M\ln[(1+\gamma e^{-2})/\delta]}{m^{(\beta)}}\right], \left[\frac{64C_0M}{m^{(\beta)}}\right]^{p/(p+2d)}\right\}.$$

Then we conclude that with probability at least $1-\eta$, the following inequality is valid.

$$\mathcal{E}(f_z) \le 2\mathcal{E}_m(f_z) + 2\varepsilon(m,\eta). \tag{4.12}$$

In addition, by inequality (4.11), we have that with probability $1-\eta$, the inequality

$$\mathcal{E}(f_{\mathcal{H}}) > 2\mathcal{E}_m(f_{\mathcal{H}}) + \frac{16M\ln[(1+\gamma e^{-2})/\eta]}{3m^{(\beta)}}$$

holds true. Then we conclude that with probability at least $1-2\eta$,

$$\mathcal{E}(f_z) - \mathcal{E}(f_{\mathcal{H}}) \le 2\varepsilon(m,\eta) + \frac{16M\ln[(1+\gamma e^{-2})/\eta]}{3m^{(\beta)}}. \tag{4.13}$$

References

[1] Milman, V.D. and Schechtman, G., *Asymptotic Theory of Finite Dimensional Normed Spaces*, Springer-Verlag, New York, 1986.

[2] Marton, K., A simple proof of the blowing-up lemma, *IEEE Transactions on Information Theory*, **32(3)**(1986), 445–446.

[3] Talagrand, M., New concentration inequalities in product spaces, *Inventiones mathematicae*, **126(3)** (1996), 505–563.

[4] de la Peña, V. and Giné, E., *Decoupling: From Dependence to Independence*, Springer, New York, 1999.

[5] Janson, S., Luczak, T. and Ruciński, A., *Random Graphs*, John Wiley, New York, 2000.

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

[7] Yu, B., Rates of convergence for empirical processes of stationary mixing sequences, *The Annals of Probability*, **22(1)**(1994), 94–116.

[8] Modha, D.S. and Masry, E., Minimum complexity regression estimation with weakly dependent observations, *IEEE Transactions on Information Theory*, **42(6)**(1996), 2133–2145.

[9] Kontorovich, L. and Ramanan, K., Concentration inequalities for dependent random variables via the martingale method, *The Annals of Probability*, **36(6)**(2008), 2126–2158.

[10] Chen, D.R., Wu, Q., Ying, Y.M. and Zhou, D.X., Support vector machine soft margin classifiers: error analysis, *Journal of Machine Learning Research*, **5**(2004), 1143–1175.

[11] Craig, C.C., On the tchebychef inequality of Bernstein, *The Annals of Mathematical Statistics*, **4(2)** (1933), 94–102.

[12] Cucker, F. and Smale, S., On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, **39(1)**(2002a), 1–49.

[13] Zou, B., Xu, Z.B. and Zhang, H., Learning rates of empirical risk minimization regression with beta-mixing inputs, *Chinese Journal of Applied Probability and Statistics*, **27(6)**(2011), 597–613.

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

[15] Cucker, F. and Smale, S., Best choices for regularization parameters in learning theory: on the bias-variance problem, *Foundations of Computational Mathematics*, **2(4)**(2002b), 413–428.

相依观察值下新的伯恩斯坦不等式及其在学习理论中的应用

邹 斌[1] 唐远炎[2] 李落清[1] 徐 婕[3]

([1]湖北大学数学与统计学学院, 武汉, 430062; [2]澳门大学科技学院, 澳门)

([3]湖北大学计算机与信息工程学院, 武汉, 430062)

经典的集中不等式描述了基于独立同分布随机变量的函数与其数学期望的偏离程度, 并且这些不等式在统计学和机器学习理论中都有许多重要的应用. 在本文, 我们超出了独立同分布随机变量这个经典框架来建立了基于β-混合序列、一致遍历马氏链的两个新的伯恩斯坦不等式. 作为这些不等式的应用, 我们又建立了基于β-混合序列的经验风险最小化算法的一致偏差速率的界.

关键词: 集中不等式, β-混合, 马氏链, 一致偏差, 经验风险最小化.

学科分类号: O211.1.