Journal of Statistical Planning and Inference 141 (2011) 1077–1087

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Learning performance of Tikhonov regularization algorithm with geometrically beta-mixing observations ☆

Bin Zou [a,b], Rong Chen [b], Zongben Xu [b,*]

[a] Faculty of Mathematics and Computer Science, Hubei University, Wuhan 430062, China
[b] Institute for Information and System Science, Faculty of Science, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 11 November 2008
Received in revised form
16 July 2010
Accepted 21 July 2010
Available online 29 July 2010

Keywords:
Generalization performance
Tikhonov regularization
Beta-mixing
Regularization error
Sample error

ABSTRACT

Estimating the generalization performance of learning algorithms is one of the main purposes of machine learning theoretical research. The previous results describing the generalization ability of Tikhonov regularization algorithm are almost all based on independent and identically distributed (i.i.d.) samples. In this paper we go far beyond this classical framework by establishing the bound on the generalization ability of Tikhonov regularization algorithm with geometrically beta-mixing observations. We first establish two refined probability inequalities for geometrically beta-mixing sequences, and then we obtain the generalization bounds of Tikhonov regularization algorithm with geometrically beta-mixing observations and show that Tikhonov regularization algorithm with geometrically beta-mixing observations is consistent. These obtained bounds on the learning performance of Tikhonov regularization algorithm with geometrically beta-mixing observations are proved to be suitable to geometrically ergodic Markov chain samples and hidden Markov models.

1. Introduction

Recently there has been a large increase of the interest for theoretical issues in the machine learning community. It is mainly due to the fact that statistical learning theory has demonstrated its usefulness by providing the ground for developing successful and well-founded learning algorithms such as support vector machines (SVMs) (Vapnik, 1998). Besides their good performance in practical applications they also enjoy a good theoretical justification in terms of both universal consistency and learning rates (see Steinwart and Christmann, 2008; Chen et al., 2004) if the training samples come from an i.i.d. process. This renewed interest for theory naturally boosted the development of performance bounds (see Chen et al., 2004; Cucker and Smale, 2001; Cucker and Zhou, 2007; Smale and Zhou, 2003). However, independence is a very restrictive concept in several ways (Steinwart et al., 2009; Vidyasagar, 2003). 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 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

☆ Supported by National 973 project (2007CB311002), NSFC key project (70501030), NSFC project (61070225) and Hubei Key Laboratory of Applied Mathematics (Hubei University) and China Postdoctoral Science Foundation (20080440190, 200902592).
* Corresponding author. Tel.: +86 29 82668005.
E-mail address: zbxu@mail.xjtu.edu.cn (Z. Xu).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.07.011

cannot be strictly justified in real-world problems. Therefore, relaxations of the independence assumption have been considered for quite a while in both machine learning and statistical literature. For example, Yu (1994) established the rates of convergence for empirical processes of stationary mixing sequences. White (1989) considered cross-validated regression estimators for strongly mixing processes. 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 property or UCEMUP property) of i.i.d. sequence are preserved when the underlying sequence is mixing sequence. Nobel and Dembo (1993) proved that if a family of functions has the property that the empirical means based on i.i.d. sequences converge uniformly 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 $\beta-$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 algorithm for both classification and regression are consistent if the samples of processes satisfying the law of large numbers. Xu and Chen (2008) established the learning rates of regularized regression for exponentially strongly mixing sequences. Smale and Zhou (2009) studied online learning algorithm with Markov sampling. Zou and Li (2007) established the performance bounds of ERM learning algorithms with exponentially strongly mixing sequences. Sun and Wu (2010) considered the regularized least square regression with dependent samples.

There are many definitions of non-independent sequences in Vidyasagar (2003) and Steinwart et al. (2009), but we are only interested in $\beta-$mixing sequence in this paper, the reasons are as follows: First, Vidyasagar (2003) pointed out that in machine learning applications, $\alpha-$mixing is "too weak" an assumption and $\phi-$mixing is "too strong" an assumption, $\beta-$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) and Meyn and Tweedie (1993) proved that a very large class of Markov chains and hidden Markov models (HMM) can produce $\beta-$mixing sequences. To study the generalization performance of Tikhonov regularization algorithm with geometrically beta-mixing observations, in this paper we first establish two refined concentration inequalities for geometrically beta-mixing sequences. We then obtain the bound on the learning rates of Tikhonov regularization algorithm with geometrically beta-mixing observations, and prove that Tikhonov regularization algorithm with geometrically beta-mixing observations is consistent.

The rest of this paper is organized as follows: In Section 2, we introduce the definitions of beta-mixing sequence and Tikhonov regularization algorithm. In Section 3 we establish two refined concentration inequalities for geometrically beta-mixing sequences. We obtain the bound on the learning rates of Tikhonov regularization algorithm with geometrically beta-mixing observations in Section 4. Finally, we give some significant conclusions in Section 5.

2. Preliminaries

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

Let $Z=\{z_i=(x_i,y_i)\}_{i=-\infty}^{\infty}$ be a stationary real-valued stochastic process defined on a probability space $(\mathcal{Z}^{\infty},\mathcal{F}^{\infty},P)$. For $-\infty<i<\infty$, let $\mathcal{F}_{-\infty}^{k}$ denote the $\sigma-$algebra generated by the random variables $z_i, i\le k$, and similarly let $\mathcal{F}_{k}^{\infty}$ denote the $\sigma-$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. Let $\overline{\mathcal{F}}_{1}^{k-1}$ denote the $\sigma-$algebra generated by the random variables $z_i, i\le 0$ as well as $z_j, j\ge k$. With these notations, there are several definitions of mixing, but we shall be concerned with only one, namely, $\beta-$mixing in this literature (see Steinwart et al., 2009; Vidyasagar, 2003; Yu, 1994).

Definition 1 (*Vidyasagar, 2003*). The sequence $Z=\{z_i=(x_i,y_i)\}_{i=-\infty}^{\infty}$ is called $\beta-$mixing, or completely regular, if

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

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

Assumption 1 (*Vidyasagar, 2003*). The sequence Z is called geometrically $\beta-$mixing, if for some constants μ and $\alpha<1$, the $\beta-$mixing coefficient satisfies

$$\beta(k)\le\mu\alpha^{k},\quad k\ge 1.$$

Remark 1. (i) In Definition 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 $\beta-$mixing coefficient thus measures how nearly the product measure approximates the actual measure P.

(ii) 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 1.

Denote by S the training sample set of size m

$$S=\{z_1=(x_1,y_1),z_2=(x_2,y_2),\ldots,z_m=(x_m,y_m)\}$$

drawn from the geometrically $\beta-$mixing sequence Z. Given a function set $\mathcal{H}$, the goal of machine learning from the sample set S is to find a function $f:\mathcal{X}\to\mathcal{Y}$ so that it has small expected risk (or error)

$$\mathcal{E}(f)=E[\ell(f,z)]=\int_{\mathcal{Z}}\ell(f,z)\,d(P_0),$$

where $\mathcal{X}$ is a compact space, and the function $\ell(f,z)$ is a non-negative loss function. Since our aim is to discuss general learning problems, we will consider the loss function of general form $\ell(f,z)$ in the sequel.

Let $\Omega:\mathcal{H}\to\mathbb{R}_+$ be a penalty functional over the hypothesis space $\mathcal{H}$. The ERM with Tikhonov (1963) regularization solves the problem

$$f_{S,\lambda}=\operatorname*{argmin}_{f\in\mathcal{H}}\{\mathcal{E}_m(f)+\lambda\Omega(f)\} \tag{1}$$

with $\lambda>0$ a constant, where $\mathcal{E}_m(f)$ is defined as

$$\mathcal{E}_m(f)=\frac{1}{m}\sum_{i=1}^{m}\ell(f,z_i).$$

The functional $\Omega(f)$ is called the regularizer and the constant λ is called the regularization parameter, it often depends on the sample size m: $\lambda=\lambda(m)$ and satisfies $\lambda\to0$ as $m\to\infty$.

Thus our purpose of this paper is to estimate the difference

$$\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f^*)$$

between the value of achieved risk $\mathcal{E}(f_{S,\lambda})$ and the value of minimal possible risk $\mathcal{E}(f^*)$ over all measure functions. According to the definition of the output function $f_{S,\lambda}$, for any $f_\lambda\in\mathcal{H}$, there holds

$$\mathcal{E}_m(f_{S,\lambda})+\lambda\Omega(f_{S,\lambda})\le\mathcal{E}_m(f_\lambda)+\lambda\Omega(f_\lambda).$$

Hence we have

$$\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f^*)\le\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f^*)+\lambda\Omega(f_{S,\lambda})\le\{\mathcal{E}(f_{S,\lambda})-\mathcal{E}_m(f_{S,\lambda})+\mathcal{E}_m(f_\lambda)-\mathcal{E}(f_\lambda)\}+\{\mathcal{E}(f_\lambda)-\mathcal{E}(f^*)+\lambda\Omega(f_\lambda)\}. \tag{2}$$

The second term in inequality (2) depends on the choice of $\mathcal{H}$, but is independent of sampling, we will call it the regularization error (see Cucker and Smale, 2001; Steinwart and Scovel, 2005; Wu et al., 2006). The first term is called the sample error.

Definition 2 (*Wu et al., 2006*). The regularization error for a function $f_\lambda\in\mathcal{H}$ is defined as

$$D(\lambda)=\mathcal{E}(f_\lambda)-\mathcal{E}(f^*)+\lambda\Omega(f_\lambda).$$

The function f_λ is called the regularizing function.

Since the minimization (1) is taken over the discrete quantity $\mathcal{E}_m(f)$, to estimate the difference $\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f^*)$, we need to estimate the capacity of the function set that contains $f_{S,\lambda}$. The capacity is measured by the covering number of $\mathcal{H}$ in this paper.

Definition 3. For a subset $\mathcal{M}$ of a metric space and any $\varepsilon>0$, the covering number $\mathcal{N}(\mathcal{M},\varepsilon)$ of the set $\mathcal{M}$ is the minimal $n\in\mathbb{N}$ such that there exist n disks in $\mathcal{M}$ with radius ε covering $\mathcal{M}$.

Define the ball of radius $R>0$ in the hypothesis space $\mathcal{H}$ as

$$B_\Omega(R)=\{f\in\mathcal{H}:\Omega(f)\le R^\theta\},\quad\theta\ge1.$$

We close this section by presenting some basic assumptions on the hypothesis space $\mathcal{H}$ and the loss function $\ell(f,z)$:

(i) We suppose that $\mathcal{H}$ is contained in a ball of a Hölder space $\mathcal{C}^p$ on a compact subset of an Euclidean space $\mathbb{R}^d$ for some $p>0$ (Zhou, 2003). Then we can assume that for any $\varepsilon>0$, the covering number of the unit ball satisfies

$$\mathcal{N}(B_\Omega(1),\varepsilon)\le\exp\{C_0\varepsilon^{-2d/p}\}$$

for some constant $C_0>0$. By dilation, we thus have that for any $\varepsilon>0$,

$$\mathcal{N}(B_\Omega(R),\varepsilon)\le\exp\{C_0R^{2d/p}\varepsilon^{-2d/p}\}. \tag{3}$$

(ii) Let $\mathcal{H}'=\mathcal{H}\cup\{f^*\}$, we define

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

and we assume that M and L are finite in this paper.

Remark 2. Note that reproducing kernel Hilbert spaces (RKHS) plays an essential role in the analysis of learning theory (see e.g. Chen et al., 2004; Cucker and Smale, 2001, 2002). But Zhou (2003) proved that if a Mercer kernel is $\mathcal{C}^p(\mathcal{X})$ ($p>0$), then the RKHS associated with this kernel can be embedded into $\mathcal{C}^{p/2}(\mathcal{X})$. This is the reason why we consider the function space $\mathcal{C}^p(\mathcal{X})$ in this paper.

3. Refined probability inequalities

In this section, we establish two refined concentration inequalities for $\beta-$mixing sequences. Our approach is based on the following three lemmas:

Lemma 1 (*Vidyasagar, 2003*). *Suppose $i_0<i_1<\cdots<i_l$ are integers, and define*

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

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

$$|E(g,P)-E(g,P_0^\infty)|\le l\beta(k)\|g\|_\infty,$$

where $E(g,P)$ and $E(g,P_0^\infty)$ are the expectation values of g with respect to P and P_0^∞ respectively.

Lemma 2 (*Hoeffding, 1963*). *Suppose that ξ is a zero-mean random variable assuming values in the interval $[a,b]$. Let $E[g]$ denote the expectation value of g. Then for any $s>0$,*

$$E[\exp(s\xi)]\le\exp(s^2(b-a)^2/8).$$

Lemma 3 (*Cucker and Smale, 2002*). *Let $c_1,c_2>0$, and $s>q>0$. Then the equation*

$$x^s-c_1x^q-c_2=0$$

has a unique positive zero x^*. In addition

$$x^*\le\max\{(2c_1)^{1/(s-q)},(2c_2)^{(1/s)}\}.$$

To exploit the $\beta-$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 notational clutter. The dependence of k and l on m is restored near the end of the paper. Let $r=m-kl$, and define

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

Note that $\bigcup_i I_i$ equals the index set $\{1,2,\ldots,m\}$ and that within each set I_i, the elements are pairwisely separated by at least k. Then we first establish the following theorem.

Theorem 1. *Let Z be a stationary β mixing sequence with the mixing coefficient satisfying Assumption* 1. *Let*

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

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. Then for any ε, $0<\varepsilon<3M$,

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

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

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

Then we have

$$\mathcal{E}_m(f)-\mathcal{E}(f)=\pi_m(S)=\sum_{i=1}^k p_ib_i(S).$$

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

$$\exp(\gamma\pi_m(S))=\exp\left[\sum_{i=1}^k\gamma p_ib_i(S)\right]\le\sum_{i=1}^k p_i\exp(\gamma b_i(S)).$$

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

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

Since

$$\exp(\gamma b_i(S)) = \exp\left[\frac{\gamma}{|I_i|}\sum_{j\in I_i} T_j\right] = \prod_{j\in I_i}\exp\left(\frac{\gamma T_j}{|I_i|}\right) \le \left[\exp\left(\frac{\gamma M}{|I_i|}\right)\right]^{|I_i|} \le e^{\gamma M},$$

where in the last step we use the fact that $T_i = \ell(f,z_i) - E[\ell(f,z_i)] \le M$ for any $i=1,2,\ldots,k$.

By Lemma 1, we get

$$E[e^{\gamma b_i(S)},P] \le (|I_i|-1)\beta(k)\|e^{\gamma b_i(S)}\|_\infty + E[e^{\gamma b_i(S)},P_0^\infty].$$

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

$$E[e^{\gamma b_i(S)},P_0^\infty] = E\left[\prod_{j\in I_i}\exp(\gamma T_j/|I_i|),P_0^\infty\right] = \{E[\exp(\gamma T_j/|I_i|),P_0]\}^{|I_i|}.$$

Apply Lemma 2 to the function T_j, since T_j has zero mean and values in an interval of width $2M$. It follows from Lemma 2 that

$$E[\exp(\gamma T_j/|I_i|)] \le \exp(\gamma^2 M^2/2|I_i|^2).$$

Thus

$$E[e^{\gamma b_i(S)},P] \le \exp\left(\frac{\gamma^2 M^2}{2|I_i|}\right) + (|I_i|-1)\beta(k)e^{\gamma M}.$$

It follows that

$$E[e^{\gamma\pi_m(S)},P] \le \sum_{i=1}^{k} p_i\left[\exp\left(\frac{\gamma^2 M^2}{2|I_i|}\right) + (|I_i|-1)\beta(k)e^{\gamma M}\right]. \tag{4}$$

We now bound the second term on the right-hand side of inequality (4) which is denoted henceforth by ϕ. We suppose $\gamma \le 3|I_i|/M$, then we have that

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

The second inequality follows from Assumption 1 and the fact that $|I_i|-1 \le e^{|I_i|-2}$ for any $|I_i| \ge 2$. We require $\exp\{k\ln(\alpha)+4|I_i|\} \le 1$, which holds if $k\ln(\alpha)+4|I_i| \le 0$. But $|I_i| \le (m/k+1)$, thus the bound holds if $4(m/k+1) \le k\ln(1/\alpha)$. Since $m+k \le 2m$, then the bound holds if $8m \le k^2\ln(1/\alpha)$ or $\{8m/\ln(1/\alpha)\}^{1/2} \le k$. Let

$$k = \left\lceil \left\{\frac{8m}{\ln(1/\alpha)}\right\}^{1/2}\right\rceil.$$

Then we have

$$\phi \le \exp\left(\frac{\gamma^2 M^2}{2|I_i|}\right) + \mu e^{-2}. \tag{5}$$

Since inequality (5) is true for all $\gamma, 0 < \gamma < 3|I_i|/M$. To make the constraint uniform over all i, we then require γ satisfies

$$0 < \gamma < \frac{3l}{M} < \frac{3|I_i|}{M}.$$

Since $\gamma^2 M^2/2l > 0$, we have

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

Returning to inequality (4), we have

$$E[e^{\gamma\pi_m(S)},P] \le (1+\mu e^{-2})\exp\left(\frac{\gamma^2 M^2}{2l}\right).$$

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

$$\begin{aligned}\text{Prob}\{\pi_m(S) > \varepsilon\} &= \text{Prob}\{e^{\gamma\pi_m(S)} > e^{\gamma\varepsilon}\}\\ &\le \frac{E[\exp\{\gamma\pi_m(S)\},P]}{\exp\{\gamma\varepsilon\}}\\ &\le (1+\mu e^{-2})\exp\left\{\frac{\gamma^2 M^2}{2l} - \gamma\varepsilon\right\}.\end{aligned}$$

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

$$\text{Prob}\{\pi_m(S) > \varepsilon\} \le (1+\mu e^{-2})\exp\left\{\frac{-l\varepsilon^2}{2M^2}\right\}$$

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

$$\text{Prob}\{\pi_m(S) > \varepsilon\} \le (1+\mu e^{-2})\exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2M^2}\right\}.$$

By symmetry, we also have

$$P\{\pi_m(S) < -\varepsilon\} \le (1+\mu e^{-2})\exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2M^2}\right\}.$$

Combining these two bounds leads to the desired inequality in Theorem 1. Then we finish the proof of Theorem 1. □

From Theorem 1, the following corollary is then immediate.

Corollary 1. *With all notations as in Theorem* 1, *then for any* $\delta \in (0,1)$, *inequality*

$$\mathcal{E}(f) - \mathcal{E}_m(f) \le M\sqrt{\frac{2\ln(C/\delta)}{m^{(\beta)}}}$$

holds true with probability at least $1-\delta$ *provided that* $m^{(\beta)} \ge 18\ln(C/\delta)$, *where* $C = 1+\mu e^{-2}$. *The same bound holds true for* $\mathcal{E}_m(f) - \mathcal{E}(f)$.

Proof. For any $\delta \in (0,1)$, the positive solution to the equation with the variable ε

$$(1+\mu e^{-2})\exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2M^2}\right\} = \delta$$

is given by

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

In addition, if $m^{(\beta)} \ge 18\ln(C/\delta)$, we have $\varepsilon < 3M$. Then by Theorem 1 we can complete the proof of Corollary 1. □

By Theorem 1, we obtain the following theorem on the rate of empirical risks uniformly converging to their expected risk over the hypothesis space $\mathcal{H}$ with the same method that used in Cucker and Smale (2001). For completeness, we give a proof.

Theorem 2. *With all notations as in Theorem* 1, *then for any* ε, $0 < \varepsilon < 3M$,

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}_m(f) - \mathcal{E}(f)| > \varepsilon\right\} \le 2(1+\mu e^{-2})\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{4L}\right)\exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{8M^2}\right\}. \tag{6}$$

Proof. Let

$$\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots \cup \mathcal{H}_n, \quad L_S(f) = \mathcal{E}(f) - \mathcal{E}_m(f),$$

then for any $\varepsilon > 0$, whenever $\sup_{f\in\mathcal{H}}|\mathcal{E}(f) - \mathcal{E}_m(f)| \ge 2\varepsilon$, there exists $k, 1 \le k \le n$ such that $\sup_{f\in\mathcal{H}_k}|\mathcal{E}(f) - \mathcal{E}_m(f)| \ge 2\varepsilon$. This implies the equivalence

$$\sup_{f\in\mathcal{H}}|\mathcal{E}(f) - \mathcal{E}_m(f)| \ge 2\varepsilon \iff \exists k,\ 1 \le k \le n, \quad \text{s.t.} \sup_{f\in\mathcal{H}_k}|\mathcal{E}(f) - \mathcal{E}_m(f)| \ge 2\varepsilon. \tag{7}$$

By the equivalence (7), 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

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\right\}\leq\sum_{k=1}^{n}\text{Prob}\left\{\sup_{f\in\mathcal{H}_k}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\right\}. \tag{8}$$

Now we estimate the term on the right-hand side of inequality (8). Let the balls $D_k, 1\leq k\leq n$ be a cover of $\mathcal{H}$ with center at f_k and radius $\varepsilon/2L$. Then, for all $S\in\mathcal{Z}^m$ and all $f\in D_k$,

$$\begin{aligned}|L_S(f)-L_S(f_k)|&\leq|\mathcal{E}(f)-\mathcal{E}(f_k)|+|\mathcal{E}_m(f)-\mathcal{E}_m(f_k)|\\&\leq E|[\ell(f,z)]-[\ell(f_k,z)]|+\frac{1}{m}\sum_{i=1}^{m}|\ell(f,z_i)-\ell(f_k,z_i)|\\&\leq 2L\cdot\|f-f_k\|_\infty\leq 2L\cdot\frac{\varepsilon}{2L}=\varepsilon.\end{aligned}$$

It follows that for any $S\in\mathcal{Z}^m$ and all $f\in D_k$

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

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

$$\text{Prob}\left\{\sup_{f\in D_k}|L_S(f)|\geq 2\varepsilon\right\}\leq\text{Prob}\{|L_S(f_k)|\geq\varepsilon\}.$$

By Theorem 1, we can get

$$\text{Prob}\{|L_S(f_k)|\geq\varepsilon\}\leq 2(1+\mu e^{-2})\exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2M^2}\right\}.$$

Then

$$\text{Prob}\left\{\sup_{f\in D_k}|L_S(f)|\geq 2\varepsilon\right\}\leq 2(1+\mu e^{-2})\exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2M^2}\right\}. \tag{9}$$

By inequalities (8) and (9), we obtain

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\right\}\leq 2(1+\mu e^{-2})\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{2L}\right)\exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2M^2}\right\}. \tag{10}$$

Theorem 2 thus follows from inequality (10) by replacing ε by $\varepsilon/2$. □

Remark 3. (i) $m^{(\beta)}$ in Theorems 1 and 2 is called the "effective number of observations" for the beta-mixing processes. From Theorems 1 and 2, we can find that $m^{(\beta)}$ plays 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, 2001; Wu et al., 2006).

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

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq\varepsilon\right\}\to 0\quad\text{as } m\to\infty.$$

This shows that as long as the covering number of the hypothesis space $\mathcal{H}$ is finite, the empirical risk $\mathcal{E}_m(f)$ will uniformly converge to the expected risk $\mathcal{E}(f)$, and the convergence speed may be exponential. This assertion is well known for the ERM algorithm with i.i.d. samples (see, e.g. Vapnik, 1998; Cucker and Smale, 2001). Then we have generalized this classical results in Vapnik (1998) and Cucker and Smale (2001) to the geometrically beta-mixing sequences.

By Theorem 2, we also get the following corollary.

Corollary 2. *With all notations as in Theorem* 1. *If for any* $\varepsilon>0$, *the covering number of function set* $\mathcal{H}$ *satisfies*

$$\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{4L}\right)\leq\exp\left\{C_0\left(\frac{\varepsilon}{4L}\right)^{-2d/p}\right\}$$

for some constant $C_0>0$. *Then for any* $\delta\in(0,1)$, *and for all functions in* $\mathcal{H}$, *inequality*

$$\mathcal{E}(f)-\mathcal{E}_m(f)\leq\varepsilon(m,\delta)$$

holds true with probability at least $1-\delta$ *provided that* $m^{(\beta)} \geq 18\ln(C/\delta)$, *where*

$$\varepsilon(m,\delta) = \max\left\{4M\left[\frac{\ln(C/\delta)}{m^{(\beta)}}\right]^{1/2}, 4\left[\frac{C_0 L^{2d/p}M^2}{m^{(\beta)}}\right]^{p/(2p+2d)}\right\}.$$

The same bound holds true for $\mathcal{E}_m(f)-\mathcal{E}(f)$.

Proof. By Theorem 2, we have that for any ε, $0<\varepsilon<3M$,

$$P\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|>\varepsilon\right\} \leq 2(1+\mu e^{-2})\exp\left\{C_0\left(\frac{\varepsilon}{4L}\right)^{-2d/p}-\frac{m^{(\beta)}\varepsilon^2}{8M^2}\right\}.$$

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 $(0<\delta<1)$

$$(1+\mu e^{-2})\exp\left\{C_0\left(\frac{\varepsilon}{4L}\right)^{-2d/p}-\frac{m^{(\beta)}\varepsilon^2}{8M^2}\right\}=\delta.$$

It follows that

$$\varepsilon^{2+2d/p}-\frac{8\ln(C/\delta)M^2}{m^{(\beta)}}\cdot\varepsilon^{2d/p}-\frac{8C_0(4L)^{2d/p}M^2}{m^{(\beta)}}=0.$$

By Lemma 3, this above equation with respect to ε has a unique positive zero ε^*, and

$$\varepsilon^* \leq \varepsilon(m,\delta) := \max\left\{4M\left[\frac{\ln(C/\delta)}{m^{(\beta)}}\right]^{1/2}, 4\left[\frac{C_0 L^{2d/p}M^2}{m^{(\beta)}}\right]^{p/(2p+2d)}\right\}.$$

Then we deduce that inequality

$$\mathcal{E}(f)-\mathcal{E}_m(f)\leq\varepsilon(m,\delta)$$

is valid with probability at least $1-\delta$ simultaneously for all functions in $\mathcal{H}$. In addition, if $m^{(\beta)}\geq 18\ln(C/\delta)$, we have $\varepsilon<3M$. Then we complete the proof of Corollary 2. □

4. Estimates error bounds

By the two refined probability inequalities (Corollaries 1 and 2) obtained in the last section, we can establish the error bound of Tikhonov regularization algorithm with geometrically $\beta-$mixing observations as follows:

Theorem 3. *Let* Z *be a stationary* β*-mixing sequence with the mixing coefficient satisfying Assumption* 1, *that is, the* β*-mixing coefficient of sequence* Z *satisfies*

$$\beta(k)\leq\mu\alpha^k,\quad k\geq 1$$

for some constants μ *and* $\alpha<1$. *Then for any* $\tau\geq 1$, *inequality*

$$\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f^*)\leq M\sqrt{\frac{2(\ln(2C)+\tau)}{m^{(\beta)}}}+\varepsilon''(m,\tau)+D(\lambda) \tag{11}$$

holds true with probability at least $1-e^{-\tau}$ *provided that* $m^{(\beta)}\geq 18(\ln(2C)+\tau)$, *where*

$$\varepsilon''(m,\tau)=\max\left\{4M\left[\frac{\ln(2C)+\tau}{m^{(\beta)}}\right]^{1/2}, 4\left[\frac{C_0[L\cdot(M/\lambda)^{1/\theta}]^{2d/p}M^2}{m^{(\beta)}}\right]^{p/(2p+2d)}\right\}.$$

Proof. By Corollary 1, we have that there exists a subset V_1 of $\mathcal{Z}^m$ with probability at least $1-e^{-\tau}$ such that for any $S\in V_1$

$$\mathcal{E}_m(f_\lambda)-\mathcal{E}(f_\lambda)\leq M\sqrt{\frac{2(\ln C+\tau)}{m^{(\beta)}}}. \tag{12}$$

Applying Corollary 2 to $B_\Omega(R)$, we have that for all $f\in B_\Omega(R)$, there exists a subset $V(R)$ of $\mathcal{Z}^m$ with probability at least $1-e^{-\tau}$,

$$\mathcal{E}(f)-\mathcal{E}_m(f)\leq\varepsilon(m,\tau), \tag{13}$$

where

$$\varepsilon(m,\tau)=\max\left\{4M\left[\frac{(\ln C+\tau)}{m^{(\beta)}}\right]^{1/2}, 4\left[\frac{C_0(LR)^{2d/p}M^2}{m^{(\beta)}}\right]^{p/(2p+2d)}\right\}.$$

Let

$$W(R)=\{S\in V_1 : f_{S,\lambda}\in B_\Omega(R)\}.$$

Combine inequalities (12) and (13) with inequality (2), we deduce that for any $S\in V(R)\cap W(R)$, with probability at least $1-e^{-\tau}$,

$$\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f^*)+\lambda\Omega(f_{S,\lambda})\le M\sqrt{\frac{2(\ln(2C)+\tau)}{m^{(\beta)}}}+\varepsilon'(m,\tau)+D(\lambda), \tag{14}$$

where

$$\varepsilon'(m,\tau)=\max\left\{4M\left[\frac{\ln(2C)+\tau}{m^{(\beta)}}\right]^{1/2}, 4\left[\frac{C_0(LR)^{2d/p}M^2}{m^{(\beta)}}\right]^{p/(2p+2d)}\right\}.$$

In addition, since for all $\lambda>0$, and almost all $S\in\mathcal{Z}^m$, we have

$$\mathcal{E}_m(f_{S,\lambda})+\lambda\Omega(f_{S,\lambda})\le\mathcal{E}_m(0)+0\le M.$$

It follows that $\Omega(f_{S,\lambda})\le M/\lambda$ for almost all $S\in\mathcal{Z}^m$. Take $R:=(M/\lambda)^{1/\theta}$ and use inequality (14), we complete the proof of Theorem 3. □

Remark 4. Since $m^{(\beta)}\to\infty$ and $\lambda:=\lambda(m)\to 0$ as $m\to\infty$, we can find that

$$M\sqrt{\frac{2(\ln(2C)+\tau)}{m^{(\beta)}}}\to 0,\quad \varepsilon''(m,\tau)\to 0,\quad D(\lambda)\to 0.$$

Then by Theorem 3, we conclude that Tikhonov regularization algorithm with geometrically beta-mixing observations is consistent. Thus we have generalized this classical results on Tikhonov regularization algorithm with i.i.d. samples in Wu (2005) to geometrically $\beta-$mixing sequences.

By Theorem 3, we can easily obtain the following learning rates in weak forms.

Corollary 3. *With all notations as in Theorem* 3, *and let* $D(\lambda)\le C_1(1/m^{(\beta)})^{p/(2p+2d)}$ *for some constant* $C_1>0$. *Then for any* $\tau\ge 1$, *there exists a constant* C_2 *such that inequality*

$$\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f^*)\le M\sqrt{\frac{2(\ln(2C)+\tau)}{m^{(\beta)}}}+C_2\left(\frac{1}{m^{(\beta)}}\right)^{p/(2p+2d)}$$

holds true with probability at least $1-e^{-\tau}$ *provided that*

$$m^{(\beta)}\ge\max\left\{18(\ln(2C)+\tau), \left[\frac{M^{[\theta(d-p)-2d]/\theta p}(\ln(2C)+\tau)^{(p+d)/p}\lambda^{2d/\theta p}}{C_0L^{2d/p}}\right]^{p/d}\right\}.$$

To improve the error estimates presented in Theorem 3, we also use iteration technique to find a small ball $B_\Omega(R)$ that contains $f_{S,\lambda}$, this technique was first used in Steinwart and Scovel (2005) and later developed in Wu et al. (2006).

Proposition 1. *Take* $0<\lambda<1/M^{\theta-1}$ *and* $R\ge M$, *then for any* $\delta\in(0,1)$, *and any* $\varepsilon>0$,

$$\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f^*)\le D(\lambda)[2+(R_\varepsilon)^{d/(p+d)}]$$

holds true with probability at least $1-\delta$ *provided that* $m^{(\beta)}\ge\max\{m_1,m_2\}$, *where*

$$m_1=\max\left\{18\ln\left(\frac{2C}{\delta}\right), \frac{[\ln(2C/\delta)]^{(p+d)/d}}{(C_0)^{p/d}L^2}\right\},\quad m_2=\max\left\{\frac{2\ln(2C/\delta)M^2}{(D(\lambda))^2}, \frac{16C_0(4L)^{2d/p}}{(D(\lambda))\frac{2p+2d}{p}}\right\}.$$

Proof. For any $\tau\ge 1$, when $0<\lambda<1/M^{\theta-1}$, we have $(M/\lambda)^{1/\theta}>M$. Take $R\ge M$, and notice that if $m^{(\beta)}>m_1$, we have

$$\varepsilon'(m,\tau)=4R^{d/(p+d)}\cdot\left[\frac{C_0L^{2d/p}}{m^{(\beta)}}\right]^{p/(2p+2d)}.$$

In addition, if $m^{(\beta)}>m_2$, we also have

$$M\sqrt{\frac{2(\ln(2C)+\tau)}{m^{(\beta)}}}\le D(\lambda),\quad 4\left[\frac{C_0(L)^{2d/p}}{m^{(\beta)}}\right]^{p/(2p+2d)}\le D(\lambda).$$

Then from inequality (14), we have that for any $S\in V(R)\cap W(R)$, inequality

$$\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f^*)+\lambda\Omega(f_{S,\lambda})\le R^{d/(p+d)}D(\lambda)+2D(\lambda)=D(\lambda)(2+R^{d/(p+d)}) \tag{15}$$

holds with probability at least $1-e^{-\tau}$. This implies $f_{S,\lambda} \in B_\Omega(g(R))$, i.e. $\Omega(f_{S,\lambda}) \le (g(R))^\theta$, where $g : \mathbb{R}_+ \to \mathbb{R}_+$ is a univariate function defined by

$$g(R) := \left(\frac{D(\lambda)}{\lambda}\right)^{1/\theta} (2+R^{d/(p+d)})^{1/\theta}.$$

It follows that

$$W(R) \cap V(R) \subseteq W(g(R)). \tag{16}$$

Denote $R_j = g(R_{j-1})$ for $j \in \mathbb{N}$, and let $R_0 = (M/\lambda)^{1/\theta}$. According to (16), we have

$$W(R_0) \cap \left(\bigcap_{i=0}^{j-1} V(R_j)\right) \subseteq W(R_j).$$

Define $r_j = (2+(r_{j-1})^{d/(p+d)})^{1/\theta}$. By Lemma 5.17 in Wu (2005), we have $r_j = [(M/\lambda)^{(1/\theta)\cdot(d/(p+d)\theta)^j} + b]$, where $b = ((3(p+d)\theta - d)/((p+d)\theta - d))^{1/\theta}$. Thus, for $\varepsilon > 0$, choose $J \in \mathbb{N}$ such that

$$J = \left\lfloor \frac{\ln(\varepsilon\theta)}{\ln\left(\frac{d}{(p+d)\theta}\right)} \right\rfloor + 1,$$

where $\lfloor v \rfloor$ denotes the integer part of $v \in \mathbb{R}_+$. It follows that $R_J \le [(M/\lambda)^\varepsilon + b]$. Set

$$R_\varepsilon := (1+b)\left(\frac{D(\lambda)}{\lambda}\right)^{1/\theta}\left(\frac{M}{\lambda}\right)^{\varepsilon}.$$

Then $W(R_J) \subset W(R_\varepsilon)$ and hence $W(R_\varepsilon)$ has measure at least $1-(J+2)e^{-\tau}$.

Applying (15) to $R = R_\varepsilon$, we have

$$\mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f^*) \le D(\lambda)[2+(R_\varepsilon)^{p/(p+d)}]$$

holds for any $S \in W(R_\varepsilon) \cap V(R_\varepsilon)$. Taking $\tau = \ln((J+3)/\delta)$, the measure of the set $W(R_\varepsilon) \cap V(R_\varepsilon)$ is at least $1-(J+3)e^{-\tau} = 1-\delta$. Then we complete the proof of Proposition 1. □

Remark 5. In the proof of Proposition 1, we use two technical conditions, that is, $\lambda < M$ and $0 < \lambda < 1/M^{\theta-1}$. It is natural because $\lambda \to 0$ as $m \to \infty$.

Remark 6. In order to better understand the significance and value of the established results for Tikhonov regularization algorithm with geometrically $\beta-$mixing samples, we give some useful discussions as follows: First, in some sense, $\beta-$mixing is a very "natural" assumption on non-i.i.d. sequences. For example, Vidyasagar (2003) and Meyn and Tweedie (1993) proved that if a Markov chain $\{z_i\}$ is V-geometrically ergodic, then the sequence $\{z_i\}$ is geometrically $\beta-$mixing. Namely, there exist constants μ and $\alpha < 1$ such that the $\beta-$mixing coefficient $\beta(k)$ satisfies

$$\beta(k) \le \mu\alpha^k \tag{17}$$

for all $k \in \mathbb{N}$. Moreover, the $\beta-$mixing coefficient is given by

$$\beta(k) \le E\{\rho[P^k(z,A),\pi],\pi\} \le \int_{\mathcal{Z}} \rho[P^k(z,A),\pi]\pi(dz),$$

where $P^k(z,A)$ is the transition probability that the state z 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

$$z_{t+1} = f(z_t) + e_t,$$

where e_t is noise sequence, $z_t \in \mathbb{R}^k$ for some integer k, subject to three suitable assumptions (see Theorem 3.11 in Vidyasagar, 2003 for details), then we can define a Lyapunov function V such that the Markov chain is geometrically $\beta-$mixing. Moreover, Meyn and Tweedie (1994) have presented a method to compute the parameters μ and α in inequality (17). Thus we can obtain the parameters μ and α of geometrically $\beta-$mixing coefficient in inequality (17) for the Markov chain described by the above recursion relation. However, other mixing sequences (i.e. $\alpha-$mixing and $\phi-$mixing) do not have this property of $\beta-$mixing sequences. The interested readers can consult Vidyasagar (2003) for the details. This implies that these results on the learning performance of Tikhonov regularization algorithm with geometrically $\beta-$mixing observations are suited to geometrically ergodic Markov chain samples.

Second, Vidyasagar (2003) proved that in hidden Markov models, if the underlying Markov chain has $\beta-$mixing property (or geometrically $\beta-$mixing), then so does the corresponding hidden Markov model. Therefore, the established results in this paper are also suited to hidden Markov models.

5. Conclusions

In this paper, we studied the learning performance of Tikhonov regularization algorithm with geometrically $\beta-$mixing observations. We first established two new refined probability inequalities for geometrically $\beta-$mixing sequences. We then derived the bounds on the learning performance of Tikhonov regularization algorithm with geometrically $\beta-$mixing samples, and proved that Tikhonov regularization algorithm with geometrically $\beta-$mixing observations is consistent. To our knowledge, these results for geometrically $\beta-$mixing 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 also give some useful discussions in the last section. By these discussions, we concluded that these established results on the learning performance of Tikhonov regularization algorithm for geometrically $\beta-$mixing observations are not only suitable to geometrically ergodic Markov chain samples, but also suitable to hidden Markov models. In addition, the obtained results extended the well-known statistical learning theory for Tikhonov regularization algorithm justified previously for i.i.d. observations in Wu et al. (2006).

Further directions of research include establishing the bounds on the better learning rates of Tikhonov regularization algorithm with geometrically $\beta-$mixing samples, and the essential difference between the generalization ability of Tikhonov regularization algorithm with i.i.d. samples and that for geometrically $\beta-$mixing samples. All these problems are under our current investigation.

Acknowledgements

The authors are grateful to the reviewers for their valuable comments and suggestions that helped to improve the original version of this paper.

References

Chen, D.R., Wu, Q., Ying, Y.M., Zhou, D.X., 2004. Support vector machine soft margin classifiers: error analysis. J. Mach. Learn. Res. 5, 1143–1175.
Cucker, F., Smale, S., 2001. On the mathematical foundations of learning. Bull. Amer. Math. Soc. 39, 1–49.
Cucker, F., Smale, S., 2002. Best choices for regularization parameters in learning theory: on the bias-variance problem. Found. Comput. Math. 2, 413–428.
Cucker, F., Zhou, D.X., 2007. Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge.
Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58, 13–30.
Karandikar, R.L., Vidyasagar, M., 2002. Rates of uniform convergence of empirical means with mixing processes. Statist. Probab. Lett. 58, 297–307.
Meyn, S.P., Tweedie, R.L., 1993. Markov Chains and Stochastic Stability. Springer-Verlag.
Meyn, S.P., Tweedie, R.L., 1994. Computable bounds for geometric convergence rates of Markov chains. Ann. Appl. Probab. 4, 981–1011.
Modha, S., Masry, E., 1996. Minimum complexity regression estimation with weakly dependent observations. IEEE Trans. Inform. Theory 42, 2133–2145.
Nobel, A., Dembo, A., 1993. A note on uniform laws of averages for dependent processes. Statist. Probab. Lett. 17, 169–172.
Smale, S., Zhou, D.X., 2003. Estimating the approximation error in learning theory. Anal. Appl. 1, 17–41.
Smale, S., Zhou, D.X., 2009. Online learning with Markov sampling. Anal. Appl. 7, 87–113.
Steinwart, I., Scovel, C., 2005. Fast rates for support vector machines. In: 18th Annual Conference on Learning Theory (COLT 2005), vol. 6, Bertinoro, Italy, pp. 279–294.
Steinwart, I., Christmann, A., 2008. Support Vector Machines. Springer, New York.
Steinwart, I., Hush, D., Scovel, C., 2009. Learning from dependent observations. Multivariate Anal. 100, 175–194.
Sun, H.W., Wu, Q., 2010. Regularized least square regression with dependent samples. Adv. Comput. Math. 32, 175–189.
Tikhonov, A.N., 1963. On solving ill-posed problem and method of regularization. Dokl. Akad. Nauk. USSR 501–504.
Vapnik, V., 1998. Statistical Learning Theory. John Wiley, New York.
Vidyasagar, M., 2003. Learning and Generalization with Applications to Neural Networks, second ed. Springer, London.
White, H., 1989. Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. Neural Networks 3, 535–549.
Wu, Q., 2005. Classification and regularization in learning theory. Thesis of Doctor of Philosophy, City University of Hong Kong, Hong Kong.
Wu, Q., Ying, Y.M., Zhou, D.X., 2006. Learning rates of least-square regularized regression. Found. Comput. Math. 6, 171–192.
Xu, Y.L., Chen, D.R., 2008. Learning rates of regularized regression for exponentially strongly mixing sequence. J. Statist. Plann. Inference 138, 2180–2189.
Yu, B., 1994. Rates of convergence for empirical processes of stationary mixing sequences. Ann. Probab. 22, 94–114.
Zhou, D.X., 2003. Capacity of reproducing kernel spaces in learning theory. IEEE Trans. Inform. Theory 49, 1743–1752.
Zou, B., Li, L.Q., 2007. The performance bounds of learning machines based on exponentially strongly mixing sequence. Comput. Math. Appl. 53 (7), 1050–1058.