Adv Comput Math (2012) 36:99–114
DOI 10.1007/s10444-011-9182-7

Generalization bounds of ERM algorithm with V-geometrically Ergodic Markov chains

Bin Zou · Zongben Xu · Xiangyu Chang

Received: 10 May 2009 / Accepted: 5 June 2010 /
Published online: 17 September 2011

Abstract The previous results describing the generalization ability of Empirical Risk Minimization (ERM) algorithm are usually based on the assumption of independent and identically distributed (i.i.d.) samples. In this paper we go far beyond this classical framework by establishing the first exponential bound on the rate of uniform convergence of the ERM algorithm with V-geometrically ergodic Markov chain samples, as the application of the bound on the rate of uniform convergence, we also obtain the generalization bounds of the ERM algorithm with V-geometrically ergodic Markov chain samples and prove that the ERM algorithm with V-geometrically ergodic Markov chain samples is consistent. The main results obtained in this paper extend the previously known results of i.i.d. observations to the case of V-geometrically ergodic Markov chain samples.

Communicated by Ding Xuan Zhou.

This work is supported in part by National 973 project (2007CB311002), NSFC key project (70501030), NSFC project (61070225), FHEC (Q20091003) and China Postdoctoral Science Foundation (20080440190, 200902592).

B. Zou (✉)
Faculty of Mathematics and Computer Science, Hubei University,
Wuhan, 430062, China
e-mail: zoubin0502@hubu.edu.cn

B. Zou · Z. Xu · X. Chang
Institute for Information and System Science, Faculty of Science,
Xi'an Jiaotong University, Xi'an, 710049, China

Z. Xu
e-mail: zbxu@mail.xjtu.edu.cn

X. Chang
e-mail: xiangyuchang@gmail.com

Keywords Generalization bounds · ERM algorithm · Geometrically ergodic · Markov chains · Learning theory

Mathematics Subject Classifications (2010) 68W40 · 68T10

1 Introduction

Quantifying the generalization performance of learning algorithms is one of the central problems in machine learning theory. The previously known results on the learning performance of ERM algorithm are usually based on the assumption that the training samples are independent and identically distributed (i.i.d.) (see e.g. [1–6, 13, 20, 21]). However, independence is a very restrictive concept in several ways (see [16–19]). 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". In addition, this i.i.d. assumption can not be strictly justified in real-world problems and 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 (see [16]). 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 [23] established the rates of convergence for empirical processes of stationary mixing sequences. Modha and Masry [10] established the minimum complexity regression estimation with m-dependent observations and strongly mixing observations respectively. Vidyasagar [19] considered the notions of mixing and proved that most of the desirable properties (e.g. property of Probably Approximately Correct (PAC) or property of Uniform Convergence of Empirical Means Uniformly in Probability (UCEMUP)) of i.i.d. sequence are preserved when the underlying sequence is mixing sequence. Smale and Zhou [13] considered least-square regularized regression without the independent assumption for Shannon sampling when the inputs samples were deterministic. Gamarnik [8] extended the PAC learning from i.i.d. samples to the case of Markov chain with finite and countably infinite state space by establishing the bounds on the sample sizes which would guarantee the PAC learning for Markov chain samples. More recently, Steinwart et al. [16] proved that the SVMs for both classification and regression are consistent only if the data-generating process satisfies a certain type of law of large numbers (e.g. Weak Law of Large Numbers for Events (WLLNE), Strong Law of Large Numbers for Events (SLLNE)). Smale and Zhou [14] considered online learning algorithm based on Markov sampling. Zou et al. [26] established the bounds on the generalization performance of the ERM algorithm with strongly mixing observations. Xu and Chen [22] considered the learning rates of regularized regression algorithm with strongly mixing sequences. Sun and Wu [17] studied the regularized least square regression with dependent samples. Steinwart and Christmann [15] considered the fast learning rates of regularized empirical risk minimizing algorithms for α-mixing process.

In real-world problems, Markov chain samples appear so often and naturally in applications, such as biological (DNA or protein) sequence analysis, times series prediction content-based web search and marking prediction and so on (see [16]). We can present an example Markov chain as follows:

Example 1 We usually have the following quantitative example in the models of random walk and predicting the weather, that is, suppose that $\{x_i\}$ is a Markov chain consisting of five states $1, 2, 3, 4, 5$ and having transition probability matrix

$$P = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.1 & 0.3 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.1 & 0.3 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.1 & 0.3 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.1 & 0.3 \\ & & & & \end{bmatrix}.$$

By the matrix P, we can create a sequence with Markov property, for example, $x_1 = 1, x_2 = 1, x_3 = 5, x_4 = 3, \cdots$. Through target function $y = f(x) = x^2 + 10x + 3$, we also can produce the corresponding values of x_i, that is, $y_1 = 14$, $y_2 = 14$, $y_3 = 78$, $y_4 = 42, \cdots$. Then a problem is posed: how can we learn the target function $f(x) = x^2 + 10x + 3$ from the Markov chain input samples

$$\{z_1 = (1, 14), z_2 = (1, 14), z_3 = (5, 78), z_4 = (3, 42), \cdots\}.$$

More importantly, many empirical evidences show that learning algorithms very often perform well with Markov chain samples (e.g. biological sequence analysis, speech recognition). Why it is so, however, has been unknown (particularly, it is unknown how well it performs in terms of consistency and generalization). For these purposes, in this paper we consider the general definition of Markov chains, V-geometrically ergodic Markov chains. We establish the first generalization bounds of the ERM algorithm with V-geometrically ergodic Markov chain samples, and show that the ERM algorithm with V-geometrically ergodic Markov chain samples is consistent.

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 results of this paper. In Section 4 we present the proofs of the obtained main results. Finally, we conclude the paper with some useful remarks in Section 5.

2 Preliminaries

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

Suppose $(\mathcal{X}, \mathcal{S})$ is a measurable space, where $\mathcal{X}$ is a compact subset of $\mathbf{R}^N (N \geq 1)$. A Markov chain is a sequence of random variables $\{X_t\}$ together with a set of probability measures $P^n(x_{n+i}|x_i)$, $x_{n+i}, x_i \in \mathcal{X}$. It is assumed that

$$P^n(x_{n+i}|x_i) := \mathrm{Prob}\{X_{n+i} = x_{n+i}|X_j,\ j < i,\ X_i = x_i\}.$$

Thus $P^n(x_{n+i}|x_i)$ denotes the probability that the state x_{n+i} after n time steps, starting from the initial state x_i at time i. It is common to denote the one-step transition probability by

$$P^1(x_{i+1}|x_i) := \mathrm{Prob}\{X_{i+1} = x_{i+1}|X_j,\ j < i,\ X_i = x_i\},$$

so that $P^1(x_{i+1}|x_i) = P(x_{i+1}|x_i)$. The fact that the transition probability does not depend on the values of X_j prior to time i is the Markov property, that is

$$\mathrm{Prob}\{X_{n+i} = x_{n+i}|X_j,\ j < i,\ X_i = x_i\} = \mathrm{Prob}\{X_{n+i} = x_{n+i}|X_i = x_i\}.$$

This is commonly expressed in words as "given the present state, the future and past states are independent". The fact that the transition probability does not depend on the initial time i means that the Markov chain is stationary (see [11, 19]), that is, if a Markov chain is started off with the initial state distributed according to a stationary distribution π, then at all subsequent times the state continues to be distributed according to the stationary distribution π.

Given two probabilities ν_1, ν_2 on the measure space $(\mathcal{X}, \mathcal{S})$, we define the total variation distance between the two measures ν_1, ν_2 as follows

$$||\nu_1 - \nu_2||_{TV} = \sup_{A \in \mathcal{S}} \left|\nu_1(A) - \nu_2(A)\right|.$$

To extend the PAC learning from i.i.d. samples to the case of Markov chain with countably infinite state space, Gamarnik [8] considered the Markov chain with countably infinite state space under two assumptions (see Assumptions A and B in [8]). Different from these definitions of [8], in this paper we consider V-geometrically ergodic Markov chains.

Definition 1 ([19]) A Markov chain $\{X_t\}_{t \geq 1}$ is said to be V-geometrically ergodic with respect to the measurable function $V : \mathcal{X} \to [1, \infty)$ if there exist constants $\gamma < \infty$ and $\rho < 1$ such that

$$||P^n(x_j|x_i) - \pi(x_j)||_{TV} \leq \gamma \rho^n V(x_i), \quad x_j, x_i \in \mathcal{X}, \ \forall n \geq 1,$$

and in addition

$$\mathrm{E}(V, \pi) = \int_{\mathcal{X}} V(x)\pi(dx) < B < \infty,$$

where π is the stationary distribution of Markov chain $\{X_t\}_{t\geq 1}$ and $\mathrm{E}(V, \pi)$ is the expectation of V with respect to the stationary distribution π.

Remark 1

(i) V-geometrically ergodic is a weaker condition than uniformly ergodic. The difference between V-geometrically ergodic and uniformly ergodic is that here the total variation distance between the n-step transition probability $P^n(\cdot|x_i)$ and the invariant measure $\pi(\cdot)$ approaches zero at a geometric rate multiplied by $V(x_i)$. Thus the rate of geometric convergence is independent of x_i, but the multiplicative constant is allowed to depend on x_i. Especially, if the state space of a Markov chain is finite, then all irreducible and aperiodic Markov chains are V-geometrically (in fact, uniformly) ergodic. And a Markov chain is V-geometrically ergodic if the condition that $V(\cdot)$ has finite expectation with respect to the invariant measure π holds. This is the reason why we consider V-geometrically ergodic Markov chains in this paper.

(ii) In [14], Smale and Zhou researched online learning algorithm based on Markov sampling. Compared Definition 1 with Definition 1 in [14], we can find that Definition 1 is a weaker definition than Definition 1 in [14] since in Definition 1, the distance between the n-step transition probability $P^n(\cdot|x_i)$ and the invariant measure $\pi(\cdot)$ approaches zero at a geometric rate multiplied by $V(x_i)$, which is allowed to depend on x_i, while in [14], the geometric convergence of the distance between probability measures $\rho_X^{(t)}$ and ρ_X is independent of X.

Denote by $\mathbf{z}$ the V-geometrically ergodic Markov chain sample set of size m

$$\mathbf{z} = \{z_1 = (x_1, y_1), z_2 = (x_2, y_2), \cdots, z_m = (x_m, y_m)\}$$

drawn from $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ according to an unknown distribution $\tilde{P}_0$. The goal of learning from the sample set $\mathbf{z}$ is to choose a function $f : \mathcal{X} \to \mathcal{Y}$ from a given function space $\mathcal{H}$ such that it has small expected risk

$$\mathcal{E}(f) = \mathrm{E}[\ell(f, z)] = \mathrm{E}[\ell(f(x), y)],$$

where $\ell(f, z)$ is a nonnegative loss function. In this paper, we would like to establish a general framework which includes classification and regression problems, so we consider the loss function of general form $\ell(f, z)$. The important feature of the regression estimation problem is that the loss function $\ell(f, z)$ can take arbitrary non-negative values whereas in pattern recognition problem it can take only two values $\{0, 1\}$ (see [4, 20]).

Since one knows only the sample set $\mathbf{z}$, the minimizer of the expected risk $\mathcal{E}(f)$ can not be computed directly. By the principle of Empirical Risk Minimization (ERM) (see [20]), we minimize, instead of the expected risk

$\mathcal{E}(f)$, the so-called empirical risk

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

Let $f_{\mathbf{z}}$ be the minimizer of empirical risk $\mathcal{E}_m(f)$ over the function set $\mathcal{H}$, i.e.,

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_m(f) = \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell(f, z_i). \tag{1}$$

Let $f_{\mathcal{H}}$ be the minimizer of expected risk $\mathcal{E}(f)$ over the function set $\mathcal{H}$, i.e., $f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}(f)$. According to the principle of ERM, we then consider the function $f_{\mathbf{z}}$ as an approximation of the target function $f_{\mathcal{H}}$. Thus our purpose in this paper is to estimate the difference $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}})$ between the value of achieved risk $\mathcal{E}(f_{\mathbf{z}})$ and the value of minimal possible risk $\mathcal{E}(f_{\mathcal{H}})$ in the function set $\mathcal{H}$. Since the minimization (1) is taken over the discrete quantity $\mathcal{E}_m(f)$, we have to regulate the capacity of the function set $\mathcal{H}$. Here the capacity is measured by the covering number.

Definition 2 For a subset $\mathcal{F}$ of a metric space $(\mathcal{B}, d)$ and $\varepsilon > 0$, the covering number $\mathcal{N}(\mathcal{F}, \varepsilon)$ of the function set $\mathcal{F}$ is the minimal $n_1 \in \mathbf{N}$ such that there exist n_1 disks in $\mathcal{F}$ with radius ε covering $\mathcal{F}$.

Now we close this section by giving some basic assumptions on the hypothesis 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 $B(\mathcal{C}^q(\mathcal{X}))$ of a Hölder space $\mathcal{C}^q(\mathcal{X})$ on a compact subset of a Euclidean space $\mathbf{R}^d$ for some $q > 0$. Here the Hölder space $\mathcal{C}^q(\mathcal{X})$ is defined as the space of all continuous functions on $\mathcal{X}$ with the following norm (see [12, 24]):

$$||f||_{\mathcal{C}^q(\mathcal{X})} := ||f||_\infty + |f|_{\mathcal{C}^q(\mathcal{X})}, \quad |f|_{\mathcal{C}^q(\mathcal{X})} := \sup_{x_1\neq x_2, x_1, x_2\in\mathcal{X}} \frac{|f(x_1) - f(x_2)|}{(d(x_1, x_2))^q},$$

where $d(\cdot, \cdot)$ is the metric defined on $\mathcal{X}$.

(ii) Assumption on the loss function: We define

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

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

By the basic assumption (i), there exists a constant $C_0 > 0$ such that for any $\varepsilon > 0$, the covering number of $\mathcal{H}$ with the metric $|| \cdot ||_{\mathcal{C}(\mathcal{X})}$ satisfies (see [24])

$$\mathcal{N}(\mathcal{H}, \varepsilon) \leq \exp\left\{C_0 \varepsilon^{\frac{-2d}{q}}\right\}. \tag{2}$$

Remark 2 Note that reproducing kernel Hilbert spaces (RKHS) plays an essential role in the analysis of learning theory (see e.g.[4, 5, 21, 24]). But Zhou [24] proved that if a Mercer kernel is $\mathcal{C}^q(\mathcal{X})(q > 0)$, then the RKHS associated with this kernel can be embedded into $\mathcal{C}^{q/2}(\mathcal{X})$. This is the reason why we consider the function space $\mathcal{C}^q(\mathcal{X})$ in this paper.

3 Main results

To measure the generalization performance of ERM algorithm, Bousquet [2], Cucker and Smale [4], Vapnik [20], Bartlett and Lugosi [1] first obtained the bounds on the rate of the empirical risks uniform convergence to their expected risks in a given function set $\mathcal{F}$ (or Q) based on i.i.d. sequences respectively, that is, for any $\varepsilon > 0$, they bounded the term

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_m(f)| > \varepsilon\right\}. \tag{3}$$

In addition, Vidyasagar [19] also established the bound on the term (3) based on α-mixing sequences. Zou and Li [25] established the bound on the term (3) based on exponentially strongly mixing observations. For more inequalities on probabilities of uniform deviations, the interested readers can consult [7] and [18] for the details. To establish the bound on the generalization ability of the ERM algorithm with V-geometrically ergodic Markov chains, we should estimate the term (3) for V-geometrically ergodic Markov chains. For this purpose, we first establish the following concentration inequality for V-geometrically ergodic Markov chains.

Theorem 1 *Let $\{z_i\}_{i=1}^m$ be a V-geometrically ergodic Markov chain. Set*

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

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 $\varepsilon, 0 < \varepsilon \leq 3M$,

$$\text{Prob}\big\{|\mathcal{E}(f) - \mathcal{E}_m(f)| \geq \varepsilon\big\} \leq 2\left(1 + \gamma Be^{-2}\right) \exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2M^2}\right\}. \tag{4}$$

Remark 3 Inequality (4) is a Hoeffding's inequality for V-geometrically ergodic Markov chains. To establish this inequality, we introduce the quantity $m^{(\beta)}$, which is called the "effective number of observations" for V-geometrically ergodic Markov chains. From Theorem 1, we can find that $m^{(\beta)}$ plays the same role in our analysis as that played by the number m of observations in the i.i.d. case. To our knowledge, this inequality here is the first inequality for V-geometrically ergodic Markov chains in this topic.

By Theorem 1 and using the similar argument conducted as Theorem B in [4], we can easily establish the following theorem on the term (3) for V-geometrically ergodic Markov chains.

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

$$\mathrm{Prob}\left\{\sup_{f\in\mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_m(f)| \geq \varepsilon\right\} \leq 2\left(1+\gamma Be^{-2}\right)\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{4L}\right)\exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{8M^2}\right\}.$$

Remark 4 (i) Since $m^{(\beta)} \to \infty$ as $m \to \infty$, by Theorem 2, we have that for any $\varepsilon, 0 < \varepsilon \leq 3M$,

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

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. [2, 4, 20]). Thus we have generalized this classical result of i.i.d. samples in [2, 4] and [20] to the case of V-geometrically ergodic Markov chain samples.

As an application of Theorems 1 and 2, we establish the following generalization bounds of the ERM algorithm (1) with V-geometrically ergodic Markov chain samples.

Proposition 1 *With all notations as in Theorem 1, then for any* $\eta_1 \in (0, 1]$, *provided that*

$$m^{(\beta)} \geq \max\left\{\frac{16\ln[(1+\gamma Be^{-2})/\eta_1]}{9}, \frac{4^{2+\frac{2d}{q}}C_0L^{\frac{2d}{q}}}{3^{2+\frac{2d}{q}}M^{\frac{2d}{q}}}\right\},$$

(i) *with probability at least* $1-\eta_1$, *the inequality*

$$\mathcal{E}(f_{\mathbf{z}}) \leq \mathcal{E}_m(f_{\mathbf{z}}) + M\sqrt{\frac{2\ln[(1+\gamma Be^{-2})/\eta_1]}{m^{(\beta)}}} \tag{5}$$

is valid.

(ii) *with probability at least* $1-2\eta_1$, *the inequality*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq M\sqrt{\frac{2\ln[(1+\gamma Be^{-2})/\eta_1]}{m^{(\beta)}}} + \varepsilon(m, \eta_1) \tag{6}$$

holds true, where

$$\varepsilon(m,\eta_1)\le \max\left\{4M\left[\frac{\ln[(1+\gamma Be^{-2})/\eta_1]}{m^{(\beta)}}\right]^{\frac{1}{2}},4\left[\frac{C_0M^2L^{\frac{2d}{q}}}{m^{(\beta)}}\right]^{\frac{q}{2q+2d}}\right\}.$$

Remark 5 (i) Bounds (5) and (6) describe the generalization performance of the ERM algorithm (1) with V-geometrically ergodic Markov chain samples for the given function set $\mathcal{H}$: Bound (5) evaluates the risk for the chosen function in the target function set $\mathcal{H}$, and bound (6) evaluates how close this risk is to the smallest possible risk for the ERM algorithm (1) with V-geometrically ergodic Markov chain samples over the target functions set $\mathcal{H}$.

(ii) Since $m^{(\beta)}\to\infty,$ *as* $m\to\infty$, we have $\varepsilon(m,\eta_1)\to 0$. By inequality (6), we get

$$\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})\to 0,\ \ \textit{as}\ m\to\infty.$$

This shows that the ERM algorithm (1) with V-geometrically erodic Markov chain samples is consistent. This implies that although the output of the ERM algorithm (1) is found via minimizing the empirical risk $\mathcal{E}_m(f)$, it can eventually predict as well as the optimal predictor $f_{\mathcal{H}}$, or it can give the best (the lowest risk) prediction for any unlabeled samples.

By Proposition 1, we can easily establish the following bound on the learning rate of the ERM algorithm (1) with V-geometrically ergodic Markov chain samples.

Corollary 1 *With all notations as in Theorem 1, then for any* $\eta_2\in(0,1]$, *with probability at least* $1-\eta_2$, *the inequality*

$$\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})\le M\sqrt{\frac{2\ln[2(1+\gamma Be^{-2})/\eta_2]}{m^{(\beta)}}}+4\left[\frac{C_0M^2L^{\frac{2d}{q}}}{m^{(\beta)}}\right]^{\frac{q}{2q+2d}}\tag{7}$$

holds true, provided that

$$m^{(\beta)}\ge\max\left\{\frac{16\ln[2C_1/\eta_2]}{9},\frac{4^{2+\frac{2d}{q}}C_0L^{\frac{2d}{q}}}{3^{2+\frac{2d}{q}}M^{\frac{2d}{q}}},\left[\frac{M^{\frac{2d}{q}}(\ln[2C_1/\eta_2])^{2+\frac{2d}{q}}}{C_0^2L^{\frac{2d}{q}}}\right]^{\frac{q}{2d}}\right\},$$

where $C_1=1+\gamma Be^{-2}$.

4 Proofs of main results

In this section, our aim is to prove these main results presented in the last section. For this purpose, we first present our main tools used in this paper.

In this paper we explore to use the β-mixing property of V-geometrically ergodic Markov chains to study the generalization of the ERM algorithm with V-geometrically ergodic Markov chain samples. Thus we present the definition of β-mixing as follows: let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary process defined on a probability space $(X^\infty, \mathcal{S}^\infty, \tilde{P})$. For $-\infty < i < \infty$, let $\mathcal{A}_{-\infty}^{k}$ denote the σ-algebra generated by the random variables $X_i, i \leq k$, and similarly let $\mathcal{A}_k^\infty$ denote the σ-algebra generated by the random variables $X_i, i \geq k$. Let $\tilde{P}_{-\infty}^{k}$ and $\tilde{P}_k^\infty$ denote the corresponding marginal probability measures, respectively. Let $\tilde{P}_0$ denote the marginal probability of each of the X_i. Let $\bar{\mathcal{A}}_1^{k-1}$ denote the σ-algebra generated by the random variables $X_i, i \leq 0$ as well as $X_j, j \geq k$.

Definition 3 ([19]) The sequence $\{X_t\}$ is called geometrically β-mixing, if there exist constants ν and $\lambda_1 < 1$ such that

$$\sup_{C \in \bar{\mathcal{A}}_1^{k-1}} \left| \tilde{P}(C) - \left(\tilde{P}_{-\infty}^{0} \times \tilde{P}_1^\infty \right)(C) \right| = \beta(k) \leq \nu \lambda_1^k, \quad \forall k \geq 1,$$

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

Lemma 1 ([19]) *Suppose X_i is a β-mixing process on a probability space $(X^\infty, \mathcal{S}^\infty, \tilde{P})$. Suppose $g : X^\infty \to R$ is essentially bounded and depends only on the variables $x_{ik}, 0 \leq i \leq l$. Let $\tilde{P}_0$ denote the one-dimensional marginal probability of each of the X_i. Then*

$$\left| \mathrm{E}\left(g, \tilde{P}\right) - \mathrm{E}\left(g, \tilde{P}_0^\infty\right) \right| \leq l\beta(k) || f ||_\infty,$$

where $\mathrm{E}(g, \tilde{P})$, $\mathrm{E}(g, \tilde{P}_0^\infty)$ are the expectations of g with respect to $\tilde{P}$, $\tilde{P}_0^\infty$, respectively.

Lemma 2 ([9]) *Suppose that ζ is a zero-mean random variable assuming values in the interval $[c, d]$. Then for any $r_1 > 0$, we have*

$$\mathrm{E}[\exp(r_1 \zeta)] \leq \exp\left(r_1^2 (d-c)^2 / 8\right).$$

Lemma 3 ([19]) *Suppose a Markov chain $\{\xi_t\}$ is V-geometrically ergodic. Then the sequence $\{\xi_t\}$ is geometrically β-mixing, and the β-mixing coefficient $\beta(n)$ is given by*

$$\beta(n) = \mathrm{E}\left\{ || P^n(\xi_j | \xi) - \pi(\xi_j) ||_{TV}, \pi \right\} = \int || P^n(\xi_j | \xi) - \pi(\xi_j) ||_{TV} \pi(d\xi).$$

Lemma 4 ([5]) *Let $c_1,\ c_2 > 0$, and $s_1 > s_2 > 0$. Then the equation*

$$x^{s_1} - c_1 x^{s_2} - c_2 = 0$$

has a unique positive zero x^*. In addition

$$x^* \le \max\left\{(2c_1)^{1/(s_1-s_2)}, (2c_2)^{(1/s_1)}\right\}.$$

Proof of Theorem 1 We decompose the proof into three steps.

Step 1 By Lemma 3, we have that a V-geometrically ergodic Markov chain is geometrically β-mixing. To exploit the β-mixing property, we then decompose the index set $I = \{1, 2, \cdots, m\}$ into different parts by following the idea of [19], that is, 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. Let $r = m - kl$, and define

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

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

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

Then we have

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

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

$$\exp(s\varpi_m(\mathbf{z})) = \exp\left[\sum_{i=1}^{k} p_i s b_i(\mathbf{z})\right] \le \sum_{i=1}^{m} p_i \exp[s b_i(\mathbf{z})].$$

It follows that

$$\mathrm{E}\left(e^{s\varpi_m(\mathbf{z})}, \tilde{P}\right) \le \sum_{i=1}^{k} p_i \mathrm{E}\left(e^{s b_i(\mathbf{z})}, \tilde{P}\right). \tag{8}$$

Since

$$\exp[s b_i(\mathbf{z})] = \exp\left[\frac{s}{|I_i|}\sum_{j\in I_i} T_j\right] = \prod_{j\in I_i} \exp\left(\frac{sT_j}{|I_i|}\right)$$
$$\le \left[\exp\left(\frac{sM}{|I_i|}\right)\right]^{|I_i|} \le e^{sM},$$

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

By Lemma 1, we have

$$\mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}\right) \leq (|I_i| - 1)\beta(k)||e^{sb_i(\mathbf{z})}||_\infty + \mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}_0^\infty\right).$$

Since under the measure $\tilde{P}_0^\infty$, the various z_i are independent, we have

$$\mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}_0^\infty\right) = \mathrm{E}\left[\prod_{j \in I_i} \exp(sT_j/|I_i|), \tilde{P}_0^\infty\right] = \left\{\mathrm{E}\left[\exp(sT_j/|I_i|), \tilde{P}_0\right]\right\}^{|I_i|}.$$

Apply Lemma 2 to the function T_j, we get $\mathrm{E}[\exp(sT_j/|I_i|), \tilde{P}_0] \leq \exp(s^2M^2/2|I_i|^2)$. Thus we have that for any $s > 0$

$$\mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}\right) \leq \exp\left(\frac{s^2M^2}{2|I_i|}\right) + (|I_i| - 1)\beta(k)e^{sM}.$$

By inequality (8) and the inequality above, we have that for any $s > 0$

$$\mathrm{E}\left(e^{s\varpi_m(\mathbf{z})}, \tilde{P}\right) \leq \sum_{i=1}^{k} p_i \left[\exp\left(\frac{s^2M^2}{2|I_i|}\right) + (|I_i| - 1)\beta(k)e^{sM}\right]. \qquad (9)$$

Step 2 We now bound the second term on the right-hand side of inequality (9) which is denoted henceforth by ϕ. By Lemma 3 and Definition 1, we have

$$\beta(k) = \mathrm{E}\{||P^n(\cdot|x) - \pi(\cdot)||_{TV}, \pi\} \leq \mathrm{E}\left[\gamma\rho^k V(x), \pi\right] \leq \gamma B\rho^k.$$

We suppose $s \leq \frac{3|I_i|}{M}$, then we have that

$$\begin{aligned}
\phi &= \exp\left(\frac{s^2M^2}{2|I_i|}\right) + (|I_i| - 1)\beta(k)e^{sM} \\
&\leq \exp\left(\frac{s^2M^2}{2|I_i|}\right) + e^{|I_i|}e^{-2}\gamma B\rho^k \cdot e^{sM} \\
&\leq \exp\left(\frac{s^2M^2}{2|I_i|}\right) + \gamma Be^{-2}\exp\{k\ln(\rho) + 4|I_i|\}.
\end{aligned}$$

We require $\exp\{k\ln(\rho) + 4|I_i|\} \leq 1$. But $|I_i| \leq (\frac{m}{k} + 1)$, thus the bound holds if $4(\frac{m}{k} + 1) \leq k\ln(1/\rho)$. Since $m + k \leq 2m$, then the bound holds if $\left\{\frac{8m}{\ln(1/\rho)}\right\}^{\frac{1}{2}} \leq k$. Let

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

Then we have

$$\phi \le \exp\left(\frac{s^2 M^2}{2l}\right) + \gamma B e^{-2}. \tag{10}$$

Since inequality (10) is true for all s, $0 < s \le \frac{3|I_i|}{M}$. To make the constraint uniform over all i, we then require s satisfies $0 < s < \frac{3l}{M} \le \frac{3|I_i|}{M}$. Since $\frac{s^2 M^2}{2l} > 0$, we have

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

Returning to inequality (9), we have that for any s, $0 < s < \frac{3l}{M}$,

$$\mathrm{E}\left(e^{s\varpi_m(\mathbf{z})}, \tilde{P}\right) \le \left(1 + \gamma B e^{-2}\right) \exp\left(\frac{s^2 M^2}{2l}\right). \tag{11}$$

Step 3 By Markov's inequality and inequality (11), we have that for any s, $0 < s \le \frac{3l}{M}$,

$$\begin{aligned}\mathrm{Prob}\left\{\mathcal{E}(f) - \mathcal{E}_m(f) \ge \varepsilon\right\} &= \mathrm{Prob}\left\{e^{s[\mathcal{E}(f) - \mathcal{E}_m(f)]} \ge e^{s\varepsilon}\right\} \\ &\le \frac{\mathrm{E}\{e^{s[\mathcal{E}(f) - \mathcal{E}_m(f)]}\}}{e^{s\varepsilon}} \\ &\le \left(1 + \gamma B e^{-2}\right) \exp\left\{-s\varepsilon + \frac{s^2 M^2}{2l}\right\}.\end{aligned}$$

Substituting $s = \frac{l\varepsilon}{M^2}$, and noting that for any $\varepsilon \le 3M$, s satisfies $s < \frac{3l}{M}$, we obtain

$$\mathrm{Prob}\left\{\mathcal{E}(f) - \mathcal{E}_m(f) \ge \varepsilon\right\} \le \left(1 + \gamma B e^{-2}\right) \exp\left\{\frac{-l\varepsilon^2}{2M^2}\right\}.$$

By symmetry, we also have

$$\mathrm{Prob}\left\{\mathcal{E}_m(f) - \mathcal{E}(f) \ge \varepsilon\right\} \le \left(1 + \gamma B e^{-2}\right) \exp\left\{\frac{-l\varepsilon^2}{2M^2}\right\}.$$

Combining these two inequalities above and replacing l by $m^{(\beta)}$ in these inequalities, we can complete the proof of Theorem 1. □

Proof of Proposition 1 For any $\eta_1 \in (0, 1]$, let

$$\left(1 + \gamma B e^{-2}\right) \exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2M^2}\right\} = \eta_1.$$

Solving the equation with respect to ε, we get

$$\varepsilon = M\sqrt{\frac{2\ln[(1+\gamma Be^{-2})/\eta_1]}{m^{(\beta)}}}.$$

Thus by Theorem 1, we have that for any $\eta_1 \in (0, 1]$, and for the function $f_{\mathcal{H}}$ that minimizes the expected risk $\mathcal{E}(f)$ over the set $\mathcal{H}$, the inequality

$$\mathcal{E}(f_{\mathcal{H}}) > \mathcal{E}_m(f_{\mathcal{H}}) - M\sqrt{\frac{2\ln[(1+\gamma Be^{-2})/\eta_1]}{m^{(\beta)}}} \tag{12}$$

holds true with probability $1-\eta_1$ provided that $m^{(\beta)} \geq \frac{2\ln[(1+\gamma Be^{-2})/\eta_1]}{9}$.

In addition, by the assumption (2), we have

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

Thus by Theorem 2, we have that for any ε, $0 < \varepsilon \leq 3M$,

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_m(f)| \geq \varepsilon\right\} \leq 2\left(1+\gamma Be^{-2}\right)\exp\left\{C_0\Big(\frac{\varepsilon}{4L}\Big)^{\frac{-2d}{q}} - \frac{m^{(\beta)}\varepsilon^2}{8M^2}\right\}.$$

Let us rewrite the above inequality in an equivalent form. For the same η_1 as above, let

$$\left(1+\gamma Be^{-2}\right)\exp\left\{C_0\Big(\frac{\varepsilon}{4L}\Big)^{\frac{-2d}{q}} - \frac{m^{(\beta)}\varepsilon^2}{8M^2}\right\} = \eta_1.$$

It follows that

$$\varepsilon^{2+\frac{2d}{q}} - \frac{8M^2\ln\left[\left(1+\gamma Be^{-2}\right)/\eta_1\right]}{m^{(\beta)}} \cdot \varepsilon^{\frac{2d}{q}} - \frac{8C_0M^2(4L)^{\frac{2d}{q}}}{m^{(\beta)}} = 0.$$

By Lemma 4, we can solve this equation with respect to ε. This equation has a unique positive zero $\varepsilon*$, and

$$\varepsilon* \doteq \varepsilon(m, \eta_1) \leq \max\left\{4M\left[\frac{\ln\left[\left(1+\gamma Be^{-2}\right)/\eta_1\right]}{m^{(\beta)}}\right]^{\frac{1}{2}}, 4\left[\frac{C_0M^2L^{\frac{2d}{q}}}{m^{(\beta)}}\right]^{\frac{q}{2q+2d}}\right\}.$$

Then we deduce that with probability at least $1-\eta_1$, for any function $f \in \mathcal{H}$, the inequality

$$\mathcal{E}(f) \leq \mathcal{E}_m(f) + \varepsilon(m, \eta_1)$$

holds true provided that $m^{(\beta)} \geq m_1$, where

$$m_1 = \max\left\{\frac{16\ln\left[\left(1+\gamma Be^{-2}\right)/\eta_1\right]}{9}, \frac{4^{2+\frac{2d}{q}}C_0L^{\frac{2d}{q}}}{3^{2+\frac{2d}{q}}M^{\frac{2d}{q}}}\right\}.$$

In particular, for the function $f_{\mathbf{z}}$ that minimizes the empirical risk $\mathcal{E}_m(f)$ over $\mathcal{H}$, with probability at least $1 - \eta_1$ the inequality

$$\mathcal{E}(f_{\mathbf{z}}) \leq \mathcal{E}_m(f_{\mathbf{z}}) + \varepsilon(m, \eta_1) \tag{13}$$

holds true provided that $m^{(\beta)} \geq m_1$.

Note that

$$\mathcal{E}_m(f_{\mathbf{z}}) \leq \mathcal{E}_m(f_{\mathcal{H}}). \tag{14}$$

Combining inequalities (13), (14) and (12), we can finish the proof of Proposition 1. □

5 Conclusions

Like i.i.d. sampling, the Markov sampling is a naturally and extensively appeared random sampling mechanism, such as time sequence, content-based pattern recognition and biological sequence analysis and so on. To study the generalization performance of the ERM algorithm with Markov chain samples, in this paper we first established the bound on the rate of uniform convergence of the ERM algorithm with V-geometrically ergodic Markov chain samples. As the application of the bound on the rate of uniform convergence, we obtained the generalization bounds of the ERM algorithm with V-geometrically ergodic Markov chain samples. We proved that the ERM algorithm with V-geometrically ergodic Markov chain samples is consistent. The main results obtained in this paper extended the previously known results (see e.g.[1, 2, 20]) of i.i.d. observations to the case of V-geometrically ergodic Markov chain samples. To our knowledge, these results here are the first explicit results on the generalization ability of the ERM algorithm with V-geometrically ergodic Markov chain samples.

Along the line of the present work, several open problems deserve further research. For example, establishing the better learning rates of the ERM algorithm with V-uniformly ergodic markov chains, and establishing the generalization bounds of regularized algorithms (e.g. regularized regression algorithms) with V-uniformly ergodic Markov chain 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 improve the original version of this paper. The authors would like to thank Professor D. X. Zhou for his careful reading and helpful comments on the paper.

References

1. Bartlett, P.L., Lugosi, G.: An inequality for uniform deviations of sample averages from their means. Stat. Probab. Lett. **4**, 55–62 (1999)

2. Bousquet, O.: New approaches to statistical learning theory. Ann. Inst. Stat. Math. **55**, 371–389 (2003)
3. Chen, D.R., Wu, Q., Ying, Y.M., Zhou, D.X.: Support vector machine soft margin classifiers: error analysis. J. Mach. Learn. Res. **5**, 1143–1175 (2004)
4. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Am. Math. Soc. **39**, 1–49 (2001)
5. Cucker, F., Smale, S.: Best choices for regularization parameters in learning theory: on the bias-variance problem. Found. Comput. Math. **2**, 413–428 (2002)
6. Cucker, F., Zhou, D.X.: Learning theory: An approximation theory viewpoint. Cambridge University Press, Cambridge (2007)
7. Devroye, L.: Bounds for the uniform deviation of empirical measures. J. Multivar. Anal. **12**, 72–79 (1982)
8. Gamarnik, D.: Extension of the PAC framework to finite and countable Markov chains. IEEE Trans. Inf. Theory **49**, 338–345 (2003)
9. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **5**8, 13–30 (1963)
10. Modha, S., Masry, E.: Minimum complexity regression estimation with weakly dependent observations. IEEE Trans. Inf. Theory **42**, 2133–2145 (1996)
11. Meyn, S.P., Tweedie, R.L.: Markov chains and stochastic stability. Springer (1993)
12. Smale, S., Zhou, D.X.: Estimating the approximation error in learning theory. Anal. Appl. **1**, 17–41 (2003)
13. Smale, S., Zhou, D.X.: Shannon sampling and function reconstruction from point values. Bull. Am. Math. Soc. **41**, 279–305 (2004)
14. Smale, S., Zhou, D.X.: Online learning with Markov sampling. Anal. Appl. **7**, 87–113 (2009)
15. Steinwart, I., Christmann, A.: Fast learning from non-i.i.d. observations. Adv. Neural Inf. Process. Syst. **22**, 1768–1776 (2009)
16. Steinwart, I., Hush, D., Scovel, C.: Learning from dependent observations. Multivariate Anal. **100**, 175–194 (2009)
17. Sun, H.W., Wu, Q.: Regularized least square regression with dependent samples. Adv. Comput. Math. **32**, 175–189 (2010)
18. Talagrand, M.: Sharper bounds for Gaussian and empirical processes. Ann. Probab. **22**, 28–76 (1994)
19. Vidyasagar, M.: Learning and generalization with applications to neural networks, 2nd edn. Springer, London (2003)
20. Vapnik, V.: Statistical learning theory. John Wiley, New York (1998)
21. Wu, Q., Zhou, D.X.: SVM soft margin classifiers: linear programming versus quadratic programming. Neural Comput. **17**, 1160–1187 (2005)
22. Xu, Y.L., Chen, D.R.: Learning rates of regularized regression for exponentially strongly mixing sequence. J. Statist. Plann. **138**, 2180–2189 (2008)
23. Yu, B.: Rates of convergence for empirical processes of stationary mixing sequences. Ann. Probab. **22**, 94–114 (1994)
24. Zhou, D.X.: Capacity of reproducing kernel spaces in learning theory. IEEE Trans. Inf. Theory **49**, 1743–1752 (2003)
25. Zou, B., Li, L.Q.: The performance bounds of learning machines based on exponentially strongly mixing sequence. Comput. Math. Appl. **53**, 1050–1058 (2007)
26. Zou, B., Li, L.Q., Xu, Z.B.: The generalization performance of ERM algorithm with strongly mixing observations. Mach. Learn. **75**, 275–295 (2009)