J. Math. Anal. Appl. 388 (2012) 333–343

Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

Generalization performance of least-square regularized regression algorithm with Markov chain samples ☆

Bin Zou [a,b,*], Luoqing Li [a], 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 October 2010
Available online 16 November 2011
Submitted by V. Pozdnyakov

Keywords:
Least-square regularized regression
Uniformly ergodic
Markov chain
Generalization
Learning theory

ABSTRACT

The previously known works describing the generalization of least-square regularized regression 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 studying the generalization of least-square regularized regression algorithm with Markov chain samples. We first establish a novel concentration inequality for uniformly ergodic Markov chains, then we establish the bounds on the generalization of least-square regularized regression algorithm with uniformly ergodic Markov chain samples, and show that least-square regularized regression algorithm with uniformly ergodic Markov chains is consistent.

1. Introduction

Learning from samples can be regarded as the regression problem of approximating a multivariate function from finite data. The problem of approximating a function from finite data is usually ill-posed and then a classical method to solve it is regularization technique (see e.g. [1–3]). The previously known results on the learning performance and consistency of regularized regression algorithm are usually based on the assumption that training samples are independent and identically distributed (i.i.d.) (see e.g. [3–7]). However, independence is a very restrictive concept (see [8,9]). 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 [10] established the rates of convergence for empirical processes of stationary mixing sequences. Modha and Masry [11] established the minimum complexity regression estimation with m-dependent observations and strongly mixing observations respectively. Samson [12] studied the concentration of measure inequalities for Markov chains and ϕ-mixing processes. Vidyasagar [9] 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. Glynn and Ormoneit [13] established the Hoeffding's inequality for uniformly ergodic Markov chains based on the equivalent definition of uniformly ergodic Markov chains. Gamarnik [14] 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, Smale and Zhou [16] considered online learning algorithm based on Markov sampling. Kontorovich and Ramanan [17] established the concentration inequalities for dependent random variables via the martingale method. Steinwart et al. [8] 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. WLLNE, SLLNE). Mohri and Rostamizadeh

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

* Corresponding author at: Faculty of Mathematics and Computer Science, Hubei University, Wuhan, 430062, China.
E-mail address: zoubin0502@hubu.edu.cn (B. Zou).

0022-247X/$ – see front matter
doi:10.1016/j.jmaa.2011.11.032

[20] studied the Rademacher complexity bounds for non-i.i.d. processes. Steinwart and Christmann [18] considered the fast learning rates of regularized empirical risk minimizing algorithm for α-mixing process. Zou et al. [19] established the bounds on the generalization performance of the ERM algorithm with strongly mixing observations.

There are many definitions of non-independent sequences in [8], but in this paper we focus only on an analysis in the case when the training samples of least-square regularized regression algorithms are Markov chains, the reasons are as follows: First, 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. We can present two examples of Markov chain input samples as follows [15]:

Example 1. Consider the problem of an insurance company wanting to draft the amount of insurance money and claim settlement according to the health condition of insurance applicants. In the simplest case, the health condition of an insurance applicant consists of healthy and ill. For an insurance applicant during given age stage, we suppose that the probability that he/she is healthy this year and also next year is given. The probability that he/she is ill this year but healthy next year is also known. Let x_i be the health condition given by the i-th year, and y_i be the corresponding profit or loss the insurance company made. Then $\{x_i\}$ is a sequence with Markov property. The insurance company had a data set of past insurance applicants and the profit or loss of the company. To draft the amount of insurance money and claim settlement, one should learn the unknown functional dependency between x_i and y_i from the Markov chain samples $\{\mathbf{z}_i = (x_i, y_i)\}_{i\geqslant 1}$.

Example 2. 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, $x_1 = 1$, $x_2 = 1$, $x_3 = 5$, $x_4 = 3$, ... Through target function $y = f(x) = x^2 + 10x + 3$, we also can produce the corresponding values of x_i, that is, $y_1 = 14$, $y_2 = 14$, $y_3 = 78$, $y_4 = 42$, ... Then a problem is posed: how can we learn the target function $f(x) = x^2 + 10x + 3$ from these Markov chain input samples and the corresponding output samples $\{z_1 = (1, 14), z_2 = (1, 14), z_3 = (5, 78), z_4 = (3, 42), \ldots\}$? In addition, many empirical evidences show that a learning algorithm very often performs well with Markov chain samples. Why it is so, however, has been unknown (particularly, it is unknown how well it performs in terms of consistency and generalization) [15]. Answering those questions is the purpose of the present paper. In this paper we first establish a novel concentration inequality for uniformly ergodic Markov chains, and then we establish the bound on the generalization of least-square regularized regression algorithm with uniformly ergodic Markov chain samples. We prove that least-square regularized regression algorithm with uniformly 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 on the generalization and consistency of least-square regularized regression algorithm with uniformly ergodic Markov chain samples. In Section 4 we prove our main results. We conclude this paper in Section 5.

2. Preliminaries

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

2.1. Stochastic input process

Suppose $(\mathcal{Z}, \mathcal{S})$ is a measurable space, a Markov chain is a sequence of random variables $\{Z_t\}_{t\geqslant 1}$ 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

$$P^n(A|z_i) \doteq \text{Prob}\{Z_{n+i} \in A \mid Z_j,\ j < i,\ Z_i = z_i\}.$$

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

$$P^1(A|z_i) \doteq \text{Prob}\{Z_{i+1} \in A \mid Z_j,\ j < i,\ Z_i = z_i\}.$$

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

$$\text{Prob}\{Z_{n+i} \in A \mid Z_j,\ j < i,\ Z_i = z_i\} = \text{Prob}\{Z_{n+i} \in A \mid Z_i = z_i\}.$$

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 $(\mathcal{Z}, \mathcal{S})$, the total variation distance between the two measures ν_1, ν_2 is defined as

$$\|\nu_1 - \nu_2\|_{TV} \doteq \sup_{A\in\mathcal{S}} \big|\nu_1(A) - \nu_2(A)\big|.$$

Thus we have the following definition of uniformly ergodic Markov chain (see e.g. [21,22]).

Definition 1. A Markov chain $\{Z_t\}_{t\geqslant 1}$ is said to be uniformly ergodic if there exist constants $\gamma < \infty$ and $\rho < 1$ such that for any $z \in \mathcal{Z}$, and for any $n \geqslant 1$,

$$\big\| P^n(\cdot|z) - \pi(\cdot)\big\|_{TV} \leqslant \gamma\rho^n,$$

where $\pi(\cdot)$ is the stationary distribution of Markov chain $\{Z_t\}_{t\geqslant 1}$.

Remark 1. (i) A weaker condition than uniformly ergodic is geometrically ergodic (see e.g. [9,22]). The difference between geometrically ergodic and uniformly ergodic is that here the constant γ does not depend on the initial state z. In particular, if the state space $\mathcal{Z}$ is finite, then all irreducible and aperiodic Markov chains are geometrically (in fact, uniformly) ergodic. In addition, by the theory of Markov chains in [21], we have that the Markov chain presented in Example 2 is a uniformly ergodic Markov chain.

(ii) By Proposition 7 in [22], we have that a Markov chain with stationary distribution $\pi(\cdot)$ is uniformly ergodic if and only if $\sup_{z\in\mathcal{Z}} \|P^n(\cdot|z) - \pi(\cdot)\|_{TV} < \alpha$ for some integer n with $\alpha < \frac{1}{2}$. Therefore, in this paper we assume that the constants γ and ρ in Definition 1 satisfy $\gamma\rho < \frac{1}{2}$. We have $d(1) := \beta_1 = \sup_{z\in\mathcal{Z}} \|P(\cdot|z) - \pi(\cdot)\|_{TV} \leqslant \gamma\rho$. By Proposition 3 in [22], we have that for any $k \in \mathbf{N}$, $d(k) \leqslant (d(1))^k \leqslant \beta_1^k$. Hence $\|P^k(\cdot|z) - \pi(\cdot)\|_{TV} \leqslant \frac{1}{2}\beta_1^k$, so the chain is uniform ergodic with $\gamma = \frac{1}{2}$ and $\rho = \beta_1$.

To develop further conditions which ensure uniform ergodicity, Meyn and Tweedie [21] presented the following definition.

Definition 2. Let $\{Z_t\}_{t\geqslant 1}$ be a uniformly ergodic Markov chain. Then there exists a probability measure ψ on $\mathcal{S}$, a positive number $\tau \in (0, 1)$, and an integer $n_1 \geqslant 1$ such that

$$P^{n_1}(z, A) \geqslant \tau\psi(A)$$

for every $z \in \mathcal{Z}$ and any $A \in \mathcal{S}$, where $P^{n_1}(z, A) \doteq \text{Prob}\{Z_{n_1+i} \in A \mid Z_i = z\}$.

2.2. Regularized regression algorithms

We consider a problem of estimating a continuous function f in $\mathcal{C}(\mathcal{X}, \mathbb{R})$, where $\mathcal{X}$ is a compact subset of $\mathbb{R}^N$ $(N \geqslant 1)$ and $\mathcal{C}(\mathcal{X})$ is a class of continuous functions on $\mathcal{X}$. The observed output y for $x \in \mathcal{X}$ can be represented by $y = f^*(x) + \epsilon_0$, where $f^*(x)$ represents the target function and ϵ_0 represents random noise with a mean of zero and a variance of $\sigma_{\epsilon_0}^2$. Let

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

be a uniformly ergodic Markov chain sample set of size m in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ drawn from an unknown distribution D. The goal of learning from the sample set $\mathbf{z}$ is to choose a function $f : \mathcal{X} \to \mathcal{Y}$ such that it is a good approximation of the target function f^*, which is a minimizer of the error (or risk)

$$\mathcal{E}(f) \doteq \mathrm{E}\big[\ell(f, z)\big] = \int_{\mathcal{Z}} \big(f(x) - y\big)^2 dD.$$

Since one knows only the training sample set $\mathbf{z}$, the minimizer of $\mathcal{E}(f)$ cannot be computed directly. According to the principle of Empirical Risk Minimization (ERM) [2], we minimize, instead of the error $\mathcal{E}(f)$, the so-called empirical error

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

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

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

By the principle of ERM, we then consider the function $f_{\mathbf{z}}$ as an approximation of the target function f^*. However, when the complexity of the function set $\mathcal{H}$ is high, the ERM algorithm (1) is usually ill-posed and overfitting may happen. Thus

regularized techniques are frequently adopted (see [1,3]). These include least-square regularized regression algorithm. The least-square regularized regression algorithm is a discrete least-square problem associated with a Mercer kernel.

let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be continuous, symmetric, and positive semidefinite, i.e., for any finite set of distinct points $\{x_1, x_2, \ldots, x_l\} \subset \mathcal{X}$, the matrix $(K(x_i, x_j))_{i,j=1}^{l}$ is positive semidefinite, such a function is called a Mercer kernel. The (RKHS) $\mathcal{H}$ associated with the kernel K is defined to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in \mathcal{X}\}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}} = \langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$. The reproducing property takes the form

$$\langle K_x, f \rangle_K = f(x), \quad \forall x \in \mathcal{X}, \ \forall f \in \mathcal{H}.$$

Denote $\mathcal{C}(\mathcal{X})$ as the space of continuous functions on $\mathcal{X}$ with the norm $\| \cdot \|_\infty$. Let $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x, x)}$, then the above reproducing property tells us that

$$\|f\|_\infty \leqslant \kappa \|f\|_K, \quad \forall f \in \mathcal{H}.$$

The least-square regularized regression algorithm is to solve

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}} \left\{ \mathcal{E}_m(f) + \lambda \|f\|_K^2 \right\} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \big(f(x_i) - y_i\big)^2 + \lambda \|f\|_K^2 \right\} \tag{2}$$

with $\lambda > 0$ a constant. The constant λ is called the regularization parameter. It often depends on the sample size m: $\lambda = \lambda(m)$, and satisfies $\lim_{m \to \infty} \lambda(m) = 0$ (see [7]).

Throughout this paper, we assume that for some $M \geqslant 0$, $|y| \leqslant M$ almost surely and $|f^*(x)| \leqslant M$.

Our purpose in this paper is to study the learning ability of least-square regularized regression algorithm (2) with uniformly ergodic Markov chain samples. In other words, we expect that the minimizer of the regularized empirical error, $f_{\mathbf{z},\lambda}$ is a good approximation of the minimizer f^* of the error $\mathcal{E}(f)$, as $m \to \infty$ and $\lambda = \lambda(m) \to 0$. Therefore, we have to estimate the difference $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*)$ between the value of achieved risk $\mathcal{E}(f_{\mathbf{z},\lambda})$ and the value of minimal possible risk $\mathcal{E}(f^*)$. Since the minimization (2) is taken over the discrete quantity $\mathcal{E}_m(f)$, we should regulate the capacity of function set $\mathcal{H}$. Here the capacity is measured by the covering number.

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

For same $R > 0$, let

$$\mathcal{B}_R = \big\{ f \in \mathcal{H} : \|f\|_K \leqslant R \big\}.$$

It can be regarded as a subset in $(\mathcal{C}(\mathcal{X}), \| \cdot \|_\infty)$. Denote the covering number of $\mathcal{B}_1$ in $\mathcal{C}(\mathcal{X})$ with the metric $\| \cdot \|_\infty$ by $\mathcal{N}(\varepsilon)$.

Definition 4. We say that the reproducing kernel Hilbert space has polynomial complexity exponent $s > 0$ if

$$\ln \mathcal{N}(\varepsilon) \leqslant C_s \varepsilon^{-s}, \quad \forall \varepsilon > 0.$$

Remark 2. Definition 4 may be found in [7] and [24]. The covering number $\mathcal{N}(\varepsilon)$ has been extensively studied, see, e.g. [26,27] and [25]. In addition, the definition of $\mathcal{B}_R$ is a general definition in learning theory, which was used to study the learning rates of least-square regularized regression in [4,6] and [7].

3. Main results

To study the learning performance of least-square regularized regression algorithm (2) with uniformly ergodic Markov chain samples, we introduce a regularizing function $\widetilde{f}_\lambda \in \mathcal{H}$. This is arbitrarily chosen and depends on λ. A special and standard choice is (see [7])

$$f_\lambda = \arg\min_{f \in \mathcal{H}} \big\{ \mathcal{E}(f) - \mathcal{E}\big(f^*\big) + \lambda \|f\|_K^2 \big\}.$$

By the definition of the output function $f_{\mathbf{z},\lambda}$, for any $\widetilde{f}_\lambda \in \mathcal{H}$, there holds

$$\mathcal{E}_m(f_{\mathbf{z},\lambda}) + \lambda \|f_{\mathbf{z},\lambda}\|_K^2 \leqslant \mathcal{E}_m(\widetilde{f}_\lambda) + \lambda \|\widetilde{f}_\lambda\|_K^2.$$

Hence we have (see [7])

$$\begin{aligned} \mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}\big(f^*\big) &\leqslant \mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}\big(f^*\big) + \lambda \|f_{\mathbf{z},\lambda}\|_K^2 \\ &\leqslant \big\{ \underbrace{\mathcal{E}_m(\widetilde{f}_\lambda) - \mathcal{E}(\widetilde{f}_\lambda)}_{T_1} + \underbrace{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_m(f_{\mathbf{z},\lambda})}_{T_2} \big\} + \big\{ \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}\big(f^*\big) + \lambda \|\widetilde{f}_\lambda\|_K^2 \big\}. \end{aligned} \tag{3}$$

In this way we decompose the excess error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*)$ into two parts: the sample error (the first term) and the regularization error (the second term).

Definition 5. The regularization error for a regularizing function $\widetilde{f}_\lambda \in \mathcal{H}$ is defined by

$$\widetilde{D}(\lambda) := \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}\big(f^*\big) + \lambda \|\widetilde{f}_\lambda\|_K^2. \tag{4}$$

Since the regularization error is independent of the learning samples, in order to estimate the excess error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*)$, our main aim is to estimate the sample error: T_1 and T_2. The function $f_{\mathbf{z},\lambda}$ in T_2 changed with the sample $\mathbf{z}$ runs over a set of functions, and should not be a fixed function. Let us begin with the estimate for T_1. In doing so, we first establish a new concentration inequality for uniformly ergodic Markov chains.

Theorem 1. *Let ξ be a random variable on a probability space $\mathcal{Z}$ and $\{z_i\}_{i=1}^m$ be a uniformly ergodic Markov chain. If $|\xi(z)| \leqslant B$ for all $z \in \mathcal{Z}$, then for any $\varepsilon > 0$,*

$$\mathrm{Prob}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathrm{E}(\xi)\right| \geqslant \varepsilon\right\} \leqslant 2\exp\left\{\frac{-m\varepsilon^2}{2B^2A_m^2}\right\}, \tag{5}$$

where $A_m = \frac{(2\gamma\rho)^m - 1}{2\gamma\rho - 1}$.

Remark 3. (i) By Remark 1, we have

$$A_m = \frac{(2\gamma\rho)^m - 1}{2\gamma\rho - 1} \to \frac{1}{1 - 2\gamma\rho}, \quad \text{as } m \to \infty.$$

This implies that bound (5) has the same convergence rate $O(\exp(-m))$ as those bounds (see e.g. [2,23]) for i.i.d. sample. In particular, when the sequence $\{z_i\}_{i=1}^m$ is i.i.d., by Definition 1, we have $A_m = 1$ for any $m \in \mathbb{N}$.

(ii) Compared inequality (5) with those results (see e.g. [18,28]) based on α-mixing sequences, we can find that these bounds for α-mixing sequences have the rate $O(\exp(-m^{(\alpha)}))$, where m is the number of samples and $m^{(\alpha)} < m$ is the "effective number of observations". This implies that these bounds for α-mixing sequences in [28] and [18] have worse convergence rate than that for i.i.d. sequences and uniformly ergodic Markov chains.

Remark 4. To have a better understanding the significance and value of the obtained result in Theorem 1, now we compare Theorem 1 with the previously known results in [17] and [13] respectively as follows: First, different from the concentration inequality (see Theorem 1.2) in [17] , inequality (5) is a generalization of Hoeffding's inequality to partial sums that are derived from a uniformly ergodic Markov chain, and inequality (5) depends on the constants γ and ρ (see Definition 1) of uniformly ergodic, does not dependent on the contraction coefficient θ_k (see definition (8)).

In addition, Glynn and Ormoneit [13] established a concentration inequality based on Definition 2. Compared inequality (5) with the inequality obtained by Glynn and Ormoneit in [13], we can find that although these two concentration inequalities have the same convergence rates, the difference is obvious, that is, the inequality obtained by Glynn and Ormoneit [13] depends on two constants n_1 and τ of Definition 2, while inequality (5) depends on two other constants γ and ρ of Definition 1, and Definition 1 is a more general definition of uniformly ergodic Markov chains (see [9]).

By Theorem 1, and using the similar arguments conducted as that in Theorem B established by Cucker and Smale [23], we obtain the following bound on the rate of the empirical error uniform convergence to the error for uniformly ergodic Markov chains.

Theorem 2. *Let $\{z_i\}_{i=1}^m$ be a uniformly ergodic Markov chain, then for any $\varepsilon > 0$,*

$$\mathrm{Prob}\Big\{\sup_{f\in\mathcal{B}_R}\big|\mathcal{E}(f) - \mathcal{E}_m(f)\big| \geqslant \varepsilon\Big\} \leqslant 2\mathcal{N}\left(\mathcal{B}_R, \frac{\varepsilon}{8(\kappa R + M)}\right)\exp\left\{\frac{-m\varepsilon^2}{8C_1^2A_m^2}\right\},$$

where A_m is as defined in Theorem 1, $C_1 = B_1(\kappa R + M)$ and $B_1 = \max\{\kappa R, M\}$.

Remark 5. Theorem 2 shows that as long as the covering number of the function space $\mathcal{B}_R$ is finite, the empirical error $\mathcal{E}_m(f)$ will uniformly converge to the error $\mathcal{E}(f)$, and the convergence speed may be exponential. Then we generalized these i.i.d. classical results in [2,23] to uniformly ergodic Markov chains.

As an application of Theorems 1 and 2, we also establish the generalization bound of least-square regularized regression algorithm (2) with uniformly ergodic Markov chains.

Theorem 3. *Let $\widetilde{D}(\lambda) = \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f^*) + \lambda\|\widetilde{f}_\lambda\|_K^2$ for any $\widetilde{f}_\lambda \in \mathcal{H}$. Suppose that $\{z_i\}_{i=1}^m$ is a uniformly ergodic Markov chain. Then for any $\eta \in (0,1)$ and $R \geqslant M$, with probability at least $1-\eta$, there holds*

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}\big(f^*\big) \leqslant A_m\Big(\kappa\sqrt{\widetilde{D}(\lambda)/\lambda} + M\Big)^2\sqrt{\frac{2\ln(2/\eta)}{m}} + 4(\kappa+1)R^2\varepsilon(m,\eta) + \widetilde{D}(\lambda),$$

where A_m is as defined in Theorem 1, $C_2 = \max\{\kappa, 1\}$ *and*

$$\varepsilon(m,\eta) \doteq \max\left\{C_2 A_m\left[\frac{\ln(2/\eta)}{m}\right]^{\frac{1}{2}}, \left[\frac{2^s C_s C_2^2 A_m^2}{m}\right]^{\frac{1}{2+s}}\right\}.$$

By Theorem 3, we also obtain the following corollary on the error bound.

Corollary 1. *Suppose that $\{z_i\}_{i=1}^m$ is a uniformly ergodic Markov chain. Let $0 < \lambda \leqslant 1$, $\widetilde{f}_\lambda \in \mathcal{H}$ and $f_{\mathbf{z},\lambda}$ be defined by* (2)*. Then for any $0 < \delta < 1$, with confidence $1-\delta$, inequality*

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*) \leqslant A_m\big(\kappa\sqrt{\widetilde{D}(\lambda)/\lambda} + M\big)^2\sqrt{\frac{2\ln(2/\delta)}{m}} + \frac{4(\kappa+1)M^2}{\lambda}\left[\frac{2^s C_s C_2^2 A_m^2}{m}\right]^{\frac{1}{2+s}} + \widetilde{D}(\lambda)$$

is valid provided that

$$m \geqslant \frac{C_2^2 A_m^2 \ln(2/\delta)}{4}\left[\frac{\ln(2/\delta)}{C_s^2}\right]^{\frac{1}{s}},$$

where A_m is as defined in Theorem 1 *and $C_2 = \max\{\kappa, 1\}$.*

By Corollary 1, we can easily establish the following bound on the learning rate of the least-square regularized regression algorithm with uniformly ergodic Markov chain samples.

Proposition 1. *Suppose that $\{z_i\}_{i=1}^m$ is a uniformly ergodic Markov chain. Assume $D(\lambda) \leqslant C_0\lambda^\beta$ for some $0 < \beta \leqslant 1$ and $C_0 > 0$, and $\lambda = \lambda(m) = m^{-\frac{1}{(1+\beta)(2+s)}}$. For any $0 < \delta < 1$, with confidence $1-\delta$, the bound*

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*) \leqslant \widetilde{C}\ln(2/\delta)\left(\frac{1}{m}\right)^{\frac{\beta}{(1+\beta)(2+s)}}$$

holds provided that

$$m \geqslant \frac{C_2^2 A_m^2 \ln(2/\delta)}{4}\left[\frac{\ln(2/\delta)}{C_s^2}\right]^{\frac{1}{s}},$$

where $\widetilde{C}$ is a constant depending on C_0, κ, M, s, β.

Remark 6. By Proposition 1, we have

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*) \to 0, \quad \text{as } m \to \infty.$$

This shows that least-square regularized regression algorithm (2) with uniformly ergodic Markov chain samples is consistent. This implies that although the output of the least-square regularized regression algorithm (2) is found via minimizing the regularized empirical error, it can eventually predict as well as the optimal predictor f^*, or it can give the best (the lowest risk) prediction for any unlabeled samples. Then we have generalized this classical results of least-square regularized regression algorithm with i.i.d. samples (see e.g. [7]) to uniformly ergodic Markov chain samples.

4. Proof of main results

To estimate the excess error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*)$, our approach is based on the following four useful lemmas. The first one is due to Kontorovich and Ramanan [17]. The second one and the third one may be found in Azuma [29]. The fourth one is due to Cucker and Smale [4].

Given $1 \leqslant i < j \leqslant m$, ψ_i^j is used to denote the sequence $(\psi_i, \psi_{i+1}, \ldots, \psi_j)$, and Ψ_i^j represents the random vector $(\Psi_i, \Psi_{i+1}, \ldots, \Psi_j)$. For simplicity, ψ_1^j and Ψ_1^j will be sometimes written simply as ψ^j and Ψ^j respectively. Given a probability space $(\mathcal{A}^m, d, \mathbb{P})$, for $1 \leqslant i < j \leqslant m$, define (see (1.1) in [17])

$$\bar{\eta}_{ij} \doteq \sup_{\psi^{i-1}\in\mathcal{B}^{i-1},\,\omega,\hat{\omega}\in\mathcal{B}} \eta_{ij}\big(\psi^{i-1},\omega,\hat{\omega}\big), \tag{6}$$

where, for $\psi^{i-1} \in \mathcal{A}^{i-1}$ and $\omega, \hat{\omega} \in \mathcal{A}$

$$\eta_{ij}\big(\psi^{i-1},\omega,\hat{\omega}\big) = \big\|P\big(\Psi_j^n\big|\Psi^i = \psi^{i-1}\omega\big) - P\big(\Psi_j^n\big|\Psi^i = \psi^{i-1}\hat{\omega}\big)\big\|_{TV}.$$

Let $\mathcal{F}$ be the set of all subsets of $\mathcal{A}^m$. For $i = 1, 2, \ldots, m$, we set $\mathcal{F}_0 = \{\emptyset, \mathcal{A}^m\}$, $\mathcal{F}_m = \mathcal{F}$ and for $1 \leqslant i \leqslant m-1$, let $\mathcal{F}_i$ be the σ-algebra generated by $\Psi^i = (\Psi_1, \Psi_2, \ldots, \Psi_i)$. Then

$$\{\emptyset, \mathcal{A}^m\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_m = \mathcal{F}.$$

For a function $g : \mathcal{A}^m \to \mathbb{R}$, we define the associated Martingale differences by

$$V_i(g) \doteq \mathrm{E}[g|\mathcal{F}_i] - \mathrm{E}[g|\mathcal{F}_{i-1}], \quad i = 1, 2, \ldots, m. \tag{7}$$

A Hamming metric $d : \mathcal{A}^m \times \mathcal{A}^m \to [0, \infty)$ on $\mathcal{A}^m$ is defined by

$$d(u, v) \doteq \sum_{i=1}^{m} \mathbf{1}_{\{\psi_i \neq \psi_i'\}},$$

where $u = (\psi_1, \psi_2, \ldots, \psi_m) \in \mathcal{A}^m$ and $v = (\psi_1', \psi_2', \ldots, \psi_m') \in \mathcal{A}^m$. For $1 \leqslant k \leqslant m$, let $P^{(k)}$ be the $\mathcal{A} \times \mathcal{A}$ transition probability matrix associated with the k-th step of the Markov chain $\{\Psi_t\}_{t \geqslant 1}$, that is, for $1 \leqslant k \leqslant m$,

$$P_{ij}^{(k)} \doteq P^k(\psi_j|\psi_i) = \mathrm{Prob}(\Psi_{k+1} = \psi_j | \Psi_k = \psi_i), \quad \psi_i, \psi_j \in \mathcal{A}.$$

The contraction coefficients θ_k are defined as

$$\theta_k \doteq \sup_{\psi_i, \psi_{i'} \in \mathcal{B}} \| P_{ij}^{(k)} - P_{i'j}^{(k)} \|_{TV}. \tag{8}$$

Kontorovich and Ramanan [17] established the following bound on the Martingale difference $V_i(g)$ (see Theorem 2.1 in [17]).

Lemma 1. *Let $\{\Psi_t\}_{t \geqslant 1}$ be a Markov chain with countable state space $\mathcal{A}$. Assume $g : \mathcal{B}^m \to \mathbb{R}$ is a c-Lipschitz function with respect to the Hamming metric on $\mathcal{A}^m$ for some constant $c > 0$. Then for $1 \leqslant i \leqslant m$,*

$$\| V_i(g) \|_\infty \leqslant c \left(1 + \sum_{j=i+1}^{m} \overline{\eta}_{ij} \right), \quad \overline{\eta}_{ij} \leqslant \theta_i \theta_{i+1} \cdots \theta_{j-1},$$

where $V_i(g)$, $\overline{\eta}_{ij}$ and θ_k are as defined in (7), (6) and (8)*, respectively.*

Lemma 2. *Let ξ_1 be such that the expected value of ξ_1, $\mathrm{E}(\xi_1) = 0$ and $-a \leqslant \xi_1 \leqslant b$. Then for any convex function f*

$$\mathrm{E}\big[f(\xi_1)\big] \leqslant \frac{b}{a+b} f(-a) + \frac{a}{a+b} f(b).$$

Lemma 3. *For any θ, $0 \leqslant \theta_1 \leqslant 1$,*

$$\theta_1 e^{(1-\theta_1)\zeta} + (1-\theta_1) e^{-\theta_1 \zeta} \leqslant e^{(\zeta^2/8)}.$$

Lemma 4. *Let $c_1, c_2 > 0$, and $p_1 > p_2 > 0$. Then the equation*

$$x^{p_1} - c_1 x^{p_2} - c_2 = 0$$

has a unique positive zero x^*. In addition $x^* \leqslant \max\{(2c_1)^{1/(p_1-p_2)}, (2c_2)^{(1/p_1)}\}$.

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

Step 1: Let $g = \frac{1}{m} \sum_{i=1}^{m} \xi(z_i)$. For any i, $1 \leqslant i \leqslant m$, we define $U_i \doteq \mathrm{E}[g|\mathcal{G}_i] - \mu$, where $\mathcal{G}_i$ is the σ-algebra generated by $z^i = (z_1, z_2, \ldots, z_i)$, and $\mu \doteq \mathrm{E}(\xi)$.

For any $\tau > 0$, note that $e^{\tau U_0} = \exp\{\tau[\mathrm{E}(g|\mathcal{G}_0) - \mu]\} = 1$. Thus for any $m \geqslant 1$, $e^{\tau U_m} = e^{\tau U_{m-1}} \cdot e^{\tau(U_m - U_{m-1})}$. It follows that

$$\mathrm{E}\big[e^{\tau U_m} \big| \mathcal{G}_{m-1}\big] = e^{\tau U_{m-1}} \cdot \mathrm{E}\big[e^{\tau(U_m - U_{m-1})} \big| \mathcal{G}_{m-1}\big]. \tag{9}$$

To estimate the second term in Eq. (9), we assume that for all $i \in \{1, 2, \ldots, m\}$, there exist two constants a_i and b_i such that $-a_i \leqslant U_i - U_{i-1} \leqslant b_i$, which will be determined in the sequel. Since $\exp(\cdot)$ is a convex function, by Lemma 2, we have that for any $\tau > 0$

$$\mathrm{E}\big[e^{\tau(U_m - U_{m-1})}\big] \leqslant \frac{b_m e^{-\tau a_m} + a_m e^{\tau b_m}}{a_m + b_m}.$$

Combining Eq. (9) with the above inequality, we have

$$\mathrm{E}\big[e^{\tau U_m}\big] \leqslant \mathrm{E}\big[e^{\tau U_{m-1}}\big] \cdot \frac{b_m e^{-\tau a_m} + a_m e^{\tau b_m}}{a_m + b_m}.$$

Iterating this inequality, we obtain that for any $\tau > 0$

$$\mathrm{E}\big[e^{\tau U_m}\big] \leqslant \prod_{i=1}^{m} \frac{b_i e^{-\tau a_i} + a_i e^{\tau b_i}}{a_i + b_i}. \tag{10}$$

By Lemma 3, we also have that for any $\tau > 0$

$$\frac{b_i e^{-\tau a_i} + a_i e^{\tau b_i}}{a_i + b_i} \leqslant \exp\left\{\frac{\tau^2 (a_i + b_i)^2}{8}\right\}.$$

Returning to inequality (10), we have that for any $\tau > 0$

$$\mathrm{E}\big[e^{\tau U_m}\big] \leqslant \exp\left\{\sum_{i=1}^{m} \frac{\tau^2 (a_i + b_i)^2}{8}\right\}. \tag{11}$$

Step 2: Now we begin to estimate the quantity $\|V_i(g)\|_\infty \doteq \|\mathrm{E}[g|\mathcal{G}_i] - \mathrm{E}[g|\mathcal{G}_{i-1}]\|_\infty$ for any $i, 1 \leqslant i \leqslant m$. By Definition 1, we have that for any k, $1 \leqslant k \leqslant m-1$,

$$\begin{aligned}
\theta_k &= \sup_{z_i, z_{i'} \in \mathcal{Z}} \big\| P_{ij}^{(k)} - P_{i'j}^{(k)} \big\|_{TV} \\
&= \sup_{z_i, z_{i'} \in \mathcal{Z}} \big\| P^k(z_j|z_i) - P^k(z_j|z_{i'}) \big\|_{TV} \\
&\leqslant \sup_{z_i, z_{i'} \in \mathcal{Z}} \big(\big\| P^k(z_j|z_i) - \pi(z_j) \big\|_{TV} + \big\| P^k(z_j|z_{i'}) - \pi(z_j) \big\|_{TV} \big) \\
&\leqslant 2\gamma\rho.
\end{aligned}$$

In addition, for any $u_1 = (z_1, z_2, \ldots, z_m) \in \mathcal{Z}^m$ and $v_1 = (z_1', z_2', \ldots, z_m') \in \mathcal{Z}^m$, we have

$$\begin{aligned}
\big|g(u_1) - g(v_1)\big| &= \left| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \frac{1}{m} \sum_{i=1}^{m} \xi\big(z_i'\big) \right| \\
&\leqslant \frac{1}{m} \sum_{i=1}^{m} \big| \xi(z_i) - \xi\big(z_i'\big) \big| \\
&\leqslant \frac{2B}{m} \sum_{i=1}^{m} \mathbf{1}_{\{z_i \neq z_i'\}}.
\end{aligned} \tag{12}$$

This implies that $g = \frac{1}{m} \sum_{i=1}^{m} \xi(z_i)$ is a $\frac{2B}{m}$-Lipschitz function with respect to the Hamming metric on $\mathcal{Z}^m$. By Lemma 1, we have that for any i, $1 \leqslant i \leqslant m$,

$$\begin{aligned}
\big\| V_i(g) \big\|_\infty &\leqslant c \left(1 + \sum_{j=i+1}^{m} \overline{\eta}_{ij} \right) \\
&\leqslant \frac{2B}{m} (1 + \overline{\eta}_{i(i+1)} + \overline{\eta}_{i(i+2)} + \cdots + \overline{\eta}_{i(m-1)}) \\
&\leqslant \frac{2B}{m} (1 + \theta_i + \theta_i \theta_{i+1} + \cdots + \theta_i \theta_{i+1} \cdots \theta_{m-1}) \\
&\leqslant \frac{2B A_m}{m},
\end{aligned}$$

where $A_m = \frac{(2\gamma\rho)^m - 1}{2\gamma\rho - 1}$.

Thus by inequality (11), we have that for any $\tau > 0$,

$$\mathrm{E}\big[e^{\tau U_m}\big] \leqslant \exp\left\{\frac{\tau^2 B^2 A_m^2}{2m}\right\}. \tag{13}$$

Step 3: By Markov's inequality, we have that for any $\varepsilon > 0$ and $\tau > 0$

$$\mathrm{Prob}\{U_m \geqslant \varepsilon\} = \mathrm{Prob}\big\{e^{\tau U_m} \geqslant e^{\tau\varepsilon}\big\} \leqslant \exp\left\{-\tau\varepsilon + \frac{\tau^2 B^2 A_m^2}{2m}\right\}.$$

Taking $\tau = \frac{m\varepsilon}{B^2 A_m^2}$, we conclude that for any $\varepsilon > 0$,

$$\operatorname{Prob}\{U_m \geqslant \varepsilon\} \leqslant \exp\left\{\frac{-m\varepsilon^2}{2B^2 A_m^2}\right\}.$$

It follows that for any $\varepsilon > 0$,

$$\operatorname{Prob}\left\{\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathrm{E}(\xi) \geqslant \varepsilon\right\} \leqslant \exp\left\{\frac{-m\varepsilon^2}{2B^2 A_m^2}\right\}.$$

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

$$\operatorname{Prob}\left\{\mathrm{E}(\xi) - \frac{1}{m}\sum_{i=1}^{m}\xi(z_i) \geqslant \varepsilon\right\} \leqslant \exp\left\{\frac{-m\varepsilon^2}{2B^2 A_m^2}\right\}.$$

Combining these two inequalities above, we then complete the proof of Theorem 1. □

Proof of Theorem 2. Let $g_1 = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$. For any $f \in \mathcal{B}_R$ and any $z_1 = (x_1, y_1), z_1' = (x_1', y_1') \in \mathcal{Z}$, we have

$$\begin{aligned}\left|\left(f(x_i) - y_i\right)^2 - \left(f\left(x_i'\right) - y_i'\right)^2\right| &= \left|\left[f(x_i) - y_i + f\left(x_i'\right) - y_i'\right]\left[f(x_i) - y_i - f\left(x_i'\right) + y_i'\right]\right| \\ &\leqslant 2(\kappa R + M)\left[\left|f(x_i) - f\left(x_i'\right)\right| + \left|y_i - y_i'\right|\right] \\ &\leqslant 2(\kappa R + M)B_1 \cdot \mathbf{1}_{\{z_i \neq z_i'\}},\end{aligned}$$

where $B_1 := \max\{\kappa R, M\}$. It follows that for any $u_1 = (z_1, z_2, \ldots, z_m) \in \mathcal{Z}^m$ and $v_1 = (z_1', z_2', \ldots, z_m') \in \mathcal{Z}^m$,

$$\begin{aligned}\left|g_1(u_1) - g_1(v_1)\right| &= \left|\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i) - y_i\right)^2 - \frac{1}{m}\sum_{i=1}^{m}\left(f(x_i) - y_i'\right)^2\right| \\ &\leqslant \frac{1}{m}\sum_{i=1}^{m}\left|\left(f(x_i) - y_i\right)^2 - \left(f\left(x_i'\right) - y_i'\right)^2\right| \\ &\leqslant \frac{2(\kappa R + M)B_1}{m}\sum_{i=1}^{m}\mathbf{1}_{\{z_i \neq z_i'\}}. \qquad (14)\end{aligned}$$

This implies that $g_1 = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$ is a $\frac{2(\kappa R + M)B_1}{m}$-Lipschitz function with respect to the Hamming metric on $\mathcal{Z}^m$. Then by Theorem 1, we have that for any $\varepsilon > 0$,

$$\operatorname{Prob}\left\{\left|\mathcal{E}_m(f) - \mathcal{E}(f)\right| \geqslant \varepsilon\right\} \leqslant 2\exp\left\{\frac{-m\varepsilon^2}{2C_1^2 A_m^2}\right\}, \qquad (15)$$

where A_m is defined as in Theorem 1, and $C_1 = (\kappa R + M)B_1$.

In addition, let $\mathcal{L}(f) = \mathcal{E}(f) - \mathcal{E}_m(f)$, we have that for any $f_1, f_2 \in \mathcal{B}_R$,

$$\begin{aligned}\left|\mathcal{L}(f_1) - \mathcal{L}(f_1)\right| &\leqslant \left|\mathcal{E}(f_1) - \mathcal{E}(f_2)\right| + \left|\mathcal{E}_m(f_1) - \mathcal{E}_m(f_2)\right| \\ &\leqslant \mathrm{E}\left\{\left|\left(f_1(x) - y\right)^2 - \left(f_2(x) - y\right)^2\right|\right\} + \frac{1}{m}\sum_{i=1}^{m}\left|\left(f_1(x_i) - y_i\right)^2 - \left(f_2(x_i) - y_i\right)^2\right| \\ &\leqslant 4(\kappa R + M) \cdot \left\|f_1(x) - f_2(x)\right\|_\infty.\end{aligned}$$

Thus by inequality (15) and using the similar arguments conducted as that in Theorem B established by Cucker and Smale [23], we can complete the proof of Theorem 2. □

Proof of Theorem 3. By the definition of $\widetilde{D}(\lambda)$, we have that

$$\lambda\|\widetilde{f}_\lambda\|_K^2 \leqslant \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}\left(f^*\right) + \lambda\|\widetilde{f}_\lambda\|_K^2 = \widetilde{D}(\lambda).$$

It follows that

$$\|\widetilde{f}_\lambda\|_\infty \leqslant \kappa\|\widetilde{f}_\lambda\|_K \leqslant \kappa\sqrt{\frac{\widetilde{D}(\lambda)}{\lambda}}$$

and

$$\big(\widetilde{f}_\lambda(x)-y\big)^2 \leqslant C_2 := \big(\kappa\sqrt{\widetilde{D}(\lambda)/\lambda}+M\big)^2.$$

By Theorem 1, we have that for any $\varepsilon>0$,

$$\mathrm{Prob}\big\{\big|\mathcal{E}_m(\widetilde{f}_\lambda)-\mathcal{E}(\widetilde{f}_\lambda)\big|\geqslant\varepsilon\big\}\leqslant 2\exp\left\{\frac{-m\varepsilon^2}{2C_2^2A_m^2}\right\}. \tag{16}$$

For any $\eta\in(0,1]$, let

$$\exp\left\{\frac{-m\varepsilon^2}{2C_2^2A_m^2}\right\}=\eta$$

and solve the equation above with respect to ε, we have

$$\varepsilon = C_2A_m\sqrt{\frac{2\ln(1/\eta)}{m}}.$$

Then by inequality (16), we have that for any $\eta\in(0,1]$, there exists a subset V_1 of $\mathcal{Z}^m$ such that for any $\widetilde{f}_\lambda\in\mathcal{H}$ and for any $\mathbf{z}\in V_1$, inequality

$$\mathcal{E}_m(\widetilde{f}_\lambda)-\mathcal{E}(\widetilde{f}_\lambda)\leqslant A_m\big(\kappa\sqrt{\widetilde{D}(\lambda)/\lambda}+M\big)^2\sqrt{\frac{2\ln(1/\eta)}{m}} \tag{17}$$

is valid with probability at least $1-\eta$.

By Theorem 2 and Definition 4, we have that for any $\varepsilon>0$,

$$\mathrm{P}\Big\{\sup_{f\in\mathcal{B}_R}\big|\mathcal{E}(f)-\mathcal{E}_m(f)\big|>\varepsilon\Big\}\leqslant 2\exp\left\{C_s\left(\frac{\varepsilon}{8R(\kappa R+M)}\right)^{-s}-\frac{m\varepsilon^2}{8C_1^2A_m^2}\right\}.$$

Let us rewrite the above inequality in the equivalent form. We equate the right-hand side of the above inequality to the same η above

$$\exp\left\{C_s\left(\frac{\varepsilon}{8R(\kappa R+M)}\right)^{-s}-\frac{m\varepsilon^2}{8C_1^2A_m^2}\right\}=\eta.$$

It follows that

$$\varepsilon^{2+s}-\frac{8C_1^2A_m^2\ln(1/\eta)}{m}\cdot\varepsilon^s-\frac{8C_1^2A_m^2C_s[8R(\kappa R+M)]^s}{m}=0.$$

By Lemma 4, this equation with respect to ε has a unique positive zero ε^*, and

$$\varepsilon^*\leqslant\varepsilon'(m,\eta)\doteq\max\left\{4C_1A_m\left[\frac{\ln(1/\eta)}{m}\right]^{\frac{1}{2}},4\left[\frac{C_sC_1^2A_m^2(2R)^s(\kappa R+M)^s}{m}\right]^{\frac{1}{2+s}}\right\}.$$

Then we deduce that for any $f\in\mathcal{B}_R$, there exists a subset $V(R)$ of $\mathcal{Z}^m$, inequality

$$\mathcal{E}(f)-\mathcal{E}_m(f)\leqslant\varepsilon'(m,\eta) \tag{18}$$

holds true with probability at least $1-\eta$.

Let

$$W(R)=\big\{\mathbf{z}\in V_1\colon\|f_{\mathbf{z},\lambda}\|_K\leqslant R\big\}.$$

By inequalities (17) and (18), we deduce that for any $\mathbf{z}\in V(R)\cap W(R)$, with probability at least $1-2\eta$,

$$\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}_m(f_{\mathbf{z},\lambda})+\mathcal{E}_m(\widetilde{f}_\lambda)-\mathcal{E}(\widetilde{f}_\lambda)\leqslant\varepsilon'(m,\eta)+A_m\big(\kappa\sqrt{\widetilde{D}(\lambda)/\lambda}+M\big)^2\sqrt{\frac{2\ln(1/\eta)}{m}}.$$

Thus by inequality (3), we have

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}\big(f^*\big)&\leqslant\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}\big(f^*\big)+\lambda\|f_{\mathbf{z},\lambda}\|_K^2\\&\leqslant A_m\big(\kappa\sqrt{\widetilde{D}(\lambda)/\lambda}+M\big)^2\sqrt{\frac{2\ln(1/\eta)}{m}}+\varepsilon'(m,\eta)+\widetilde{D}(\lambda).\end{aligned} \tag{19}$$

Replacing η by $\eta/2$ in inequality (19), we can complete the proof of Theorem 3. □

Proof of Corollary 1. For all $\lambda > 0$, and almost all $\mathbf{z} \in \mathcal{Z}^m$, by the definition of $f_{\mathbf{z},\lambda}$, we have that for $f = 0$,

$$\begin{aligned}\lambda \|f_{\mathbf{z},\lambda}\|_K^2 &\leqslant \mathcal{E}_m(f_{\mathbf{z},\lambda}) + \lambda \|f_{\mathbf{z},\lambda}\|_K^2 \\ &\leqslant \mathcal{E}_m(0) + 0 \\ &= \frac{1}{m}\sum_{i=1}^{m}(0 - y_i)^2 \leqslant M^2.\end{aligned}$$

Then we have $\|f_{\mathbf{z},\lambda}\|_K \leqslant M/\sqrt{\lambda}$ for almost all $\mathbf{z} \in \mathcal{Z}^m$. This implies that $f_{\mathbf{z},\lambda} \in \mathcal{B}_R$ with $R = M/\sqrt{\lambda}$. Replacing R by $M/\sqrt{\lambda}$ in Theorem 3, we can easily finish the proof of Corollary 1. □

5. Conclusions

In order to study the learning performance of least-square regularized regression algorithm with uniformly ergodic Markov chain samples, we first established a new concentration inequality for uniformly ergodic Markov chains, then we established the bound on the generalization performance of least-square regularized regression algorithm with uniformly ergodic Markov chain samples, and proved that least-square regularized regression algorithm with uniformly ergodic Markov chain samples is consistent. These results extended the i.i.d. classical results on the learning performance of least-square regularized regression algorithm to the case of Markov chain samples. To our knowledge, these studies here are the first works on this topic.

Along the line of the present work, several open problems deserve further research. For example, establishing the bound on the fast learning rates of least-square regularized regression algorithm with uniformly ergodic Markov chain, and establishing the bound on the consistency and generalization of regularized regression algorithms with uniformly ergodic Markov chain based on the measure of Rademacher average. All these problems are under our current investigation.

Acknowledgments

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

References

[1] A.N. Tikhonov, Solution of incorrectly for mulated problems and regularization, Soviet Math. Dokl. 4 (1963) 1035–1038.
[2] V. Vapnik, Statistical Learning Theory, John Wiley, New York, 1998.
[3] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.
[4] F. Cucker, S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, Found. Comput. Math. 2 (2002) 413–428.
[5] I. Steinwart, Consistency of support vector machines and other regularized kernel classifiers, IEEE Trans. Inform. Theory 51 (2005) 128–142.
[6] D.R. Chen, Q. Wu, Y.M. Ying, D.X. Zhou, Support vector machine soft margin classifiers: Error analysis, J. Mach. Learn. Res. 5 (2004) 1143–1175.
[7] Q. Wu, Yiming Ying, D.X. Zhou, Learning rates of least-square regularized regression, Found. Comput. Math. 6 (2006) 171–192.
[8] I. Steinwart, D. Hush, C. Scovel, Learning from dependent observations, J. Multivariate Anal. 100 (2009) 175–194.
[9] M. Vidyasagar, Learning and Generalization with Applications to Neural Networks, Springer, London, 2003.
[10] B. Yu, Rates of convergence for empirical processes of stationary mixing sequences, Ann. Probab. 22 (1994) 94–116.
[11] S. Modha, E. Masry, Minimum complexity regression estimation with weakly dependent observations, IEEE Trans. Inform. Theory 42 (1996) 2133–2145.
[12] P.M. Samson, Concentration of measure inequalities for Markov chains and ϕ-mixing processes, Ann. Probab. 28 (2000) 416–461.
[13] P.W. Glynn, D. Ormoneit, Hoeffding's inequality for uniformly ergodic Markov chains, Statist. Probab. Lett. 56 (2002) 143–146.
[14] D. Gamarnik, Extension of the PAC framework to finite and countable Markov chains, IEEE Trans. Inform. Theory 49 (2003) 338–345.
[15] B. Zou, H. Zhang, Z.B. Xu, Learning from uniformly ergodic Markov chain samples, J. Complexity 25 (2009) 188–200.
[16] S. Smale, D.X. Zhou, Online learning with Markov sampling, Anal. Appl. 7 (2009) 87–113.
[17] L. Kontorovich, K. Ramanan, Concentration inequalities for dependent random variables via the martingale method, Ann. Probab. 36 (2008) 2126–2158.
[18] I. Steinwart, A. Christmann, Fast learning from non-i.i.d. observations, Adv. Neural Inf. Process. Syst. 22 (2009) 1768–1776.
[19] B. Zou, L.Q. Li, Z.B. Xu, The generalization performance of ERM algorithm with strongly mixing observations, Machine Learning 75 (2009) 275–295.
[20] M. Mohri, A. Rostamizadeh, Rademacher complexity bounds for non-i.i.d. processes, in: Advances in Neural Information Processing Systems, NIPS, 2008, MIT Press, Canada, 2009.
[21] S.P. Meyn, R.L. Tweedie, Markov Chains and Stochastic Stability, Springer-Verlag, 1993.
[22] G.O. Roberts, J.S. Rosenthal, General state space Markov chains and MCMC algorithms, Probab. Surv. 1 (2004) 20–71.
[23] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2001) 1–49.
[24] N. Alon, S. Ben-David, N. Cesa-Bianchi, Scale-sensitive dimensions, uniform convergence and learnability, J. ACM 44 (1997) 615–631.
[25] P.L. Bartlett, The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network, IEEE Trans. Inform. Theory 44 (1998) 525–536.
[26] D.X. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.
[27] D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003) 1743–1752.
[28] B. Zou, L.Q. Li, The performance bounds of learning machines based on exponentially strongly mixing sequence, Comput. Math. Appl. 53 (2007) 1050–1058.
[29] K. Azuma, Weighted sums of certain dependent random variables, Tohoku Math. J. 19 (1967) 357–367.