Neural Networks 53 (2014) 40–51

Contents lists available at ScienceDirect

Neural Networks

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

Generalization performance of Gaussian kernels SVMC based on Markov sampling

Jie Xu [a], Yuan Yan Tang [c], Bin Zou [a,*], Zongben Xu [b], Luoqing Li [a], Yang Lu [c]

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

ARTICLE INFO

Article history:
Received 20 June 2013
Received in revised form 18 January 2014
Accepted 24 January 2014

Keywords:
Gaussian RBF kernels
SVMC
Uniformly ergodic Markov chain (u.e.M.c.)
Generalization performance
Markov sampling

ABSTRACT

In this paper we consider Gaussian RBF kernels support vector machine classification (SVMC) algorithm with uniformly ergodic Markov chain (u.e.M.c.) samples in reproducing kernel Hilbert spaces (RKHS). We analyze the learning rates of Gaussian RBF kernels SVMC based on u.e.M.c. samples and obtain the fast learning rate of Gaussian RBF kernels SVMC based on u.e.M.c. samples by using the strongly mixing property of u.e.M.c. samples. We also present the numerical studies on the learning performance of Gaussian RBF kernels SVMC based on Markov sampling for real-world datasets. These experimental results show that Gaussian RBF kernels SVMC based on Markov sampling has better learning performance compared to randomly independent sampling.

1. Introduction

Support Vector Machine (SVM) is one of the most widely used machine learning algorithms for classification problems (Vapnik, 1998). Besides their good performance in practical applications, they also enjoy a good theoretical justification in terms of both universal consistency (Steinwart, 2001; Steinwart & Christmann, 2008) and learning rates (Chen, Wu, Ying, & Zhou, 2004; Steinwart & Scovel, 2007) if the training samples come from an independent and identically distributed (i.i.d.) process. However, independence is a very restrictive concept (Steinwart, Hush, & Scovel, 2009; Vidyasagar, 2003) and this i.i.d. assumption cannot 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 (Steinwart et al., 2009). Therefore, relaxations of such i.i.d. assumption have been considered for quite a while in both machine learning and statistics literatures. For example, Yu (1994) established the rates of convergence for empirical processes of stationary mixing sequences. Modha and Masry (1996) established the minimum complexity regression estimation with m-dependent observations and strongly mixing observations. Smale and Zhou (2009) considered online learning algorithm based on Markov sampling. Steinwart et al. (2009) proved that the SVM for both classification and regression are consistent only if the data-generating process satisfies a certain type of law of large numbers. Zou, Li, and Xu (2009) established the generalization bounds of empirical risk minimization (ERM) algorithm with strongly mixing observations. Mohri and Rostamizadeh (2010) studied the stability bounds of learning algorithms for non-i.i.d. processes.

In this paper, we focus only on an analysis in the case when the input samples are Markov chains, the reasons are as follows: First, in real-world problems, Markov chain samples appear so often and naturally in applications, such as biological (DNA or protein) sequence analysis, content-based web search and marking prediction, and so on. Second, many empirical evidences (Curnow, 1988; Laarhouen & Aarts, 1987; Zou, Li, Xu, Luo, & Tang, 2013) show that learning algorithms very often perform well with Markov chain samples. Why it is so, however, has been unknown (particularly, it is unknown how well it performs in terms of learning rate and generalization). For these reasons, Zou, Peng, and Xu (2013) introduced a Markov sampling algorithm and presented the numerical studies on the learning performance of SVMC with Markov chain samples based on linear prediction models. Since Gaussian RBF kernels are the most widely used kernels in practice (Steinwart & Scovel, 2007), in this paper we consider Gaussian RBF kernels SVMC algorithm with u.e.M.c. samples. We not only

* Corresponding author.
E-mail addresses: jiexu@mail.hust.edu.cn (J. Xu), yytang@umac.mo (Y.Y. Tang), zoubin0502@gmail.com (B. Zou), zbxu@mail.xjtu.edu.cn (Z. Xu), humcli@gmail.com (L. Li), lylylytc@gmail.com (Y. Lu).

http://dx.doi.org/10.1016/j.neunet.2014.01.013
0893-6080/

give an error analysis for Gaussian RBF kernels SVMC algorithm with u.e.M.c. samples, but also obtain the fast learning rate for Gaussian RBF kernels SVMC algorithm with u.e.M.c. samples. In addition, we give a slightly modified version of Markov sampling introduced in Zou, Peng et al. (2013) such that it suits the setting of Gaussian RBF kernels, and then we present the numerical studies on the learning performance of Gaussian RBF kernels SVMC based on Markov sampling for real-world datasets. The experimental results show that Gaussian RBF kernels SVMC based on Markov sampling has better learning performance compared to randomly independent sampling.

This paper is organized as follows: in Section 2, we give some definitions and notations. In Section 3, we present the main results on the learning rates of Gaussian RBF kernels SVMC based on u.e.M.c. samples. In Section 4, we give the numerical studies on the learning performance of Gaussian RBF kernels SVMC algorithm based on Markov sampling. Finally, we conclude this paper in Section 5.

2. Preliminaries

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

2.1. SVMC algorithm

Let (X, d) be a compact metric space and $Y = \{-1, 1\}$. A binary classifier is a function $h : X \to Y$ which labels every point $x \in X$ with some $y \in Y$. The misclassification error for the classifier $h : X \to Y$ is defined to be the probability of the event $\{h(X) \neq Y\}$, that is, $\mathcal{R}(h) = \text{Prob}\{h(X) \neq Y\}$. In this paper, our hypothesis space is a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ (Aronszajn, 1950). Namely, let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite, i.e., for any finite set of distinct points $\{x_1, x_2, \ldots, x_l\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^l$ is positive semidefinite. Such a function is called a Mercer kernel. The RKHS $\mathcal{H}_K$ associated with the kernel K is defined to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K} = \langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_{x'} \rangle_K = K(x, x')$, that is, $\langle \sum_i \alpha_i K_{x_i}, \sum_j \beta_j K_{x_j} \rangle_K = \sum_{i,j} \alpha_i \beta_j K(x_i, x_j)$. The reproducing property takes the form $\langle K_x, f \rangle_K = f(x), \ \forall x \in X, \ \forall f \in \mathcal{H}_K$. Denote $\mathcal{C}(X)$ as the space of continuous functions on X with the norm $\|f\|_\infty = \sup_{x \in X} |f(x)|$. Let $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$; then the reproducing property tells us that $\|f\|_\infty \leq \kappa \|f\|_K, \forall f \in \mathcal{H}_K$.

For a function $f : X \to \mathbb{R}$, the sign function of f is defined as $sgn(f) = 1$ if $f(x) \geq 0$ and $sgn(f) = -1$ if $f(x) < 0$. The soft margin SVM classifier associated with the Mercer kernel K is defined as $sgn(f_S)$ (Vapnik, 1998), where f_S is a minimizer of the following optimization problem involving a set of random samples $S = (x_i, y_i)_{i=1}^m \in \mathcal{Z}^m$:

$$f_S = \arg\min_{f \in \mathcal{H}_K} \ \frac{1}{2}\|f\|_K^2 + \frac{C}{m}\sum_{i=1}^m \xi_i, \quad \text{subject to } y_i f(x_i) \geq 1 - \xi_i, \ \xi_i \geq 0, \tag{1}$$

where C is a constant which depends on $m : C = C(m)$ and often $\lim_{m \to \infty} C(m) = \infty$ (Chen et al., 2004).

A good classifier should produce decision functions whose risks converge to the best classifier, the Bayes classifier, as m and hence $C(m)$ tend to infinity. Let ψ be a probability distribution on $\mathcal{Z} = X \times Y$. The regression function of ψ is defined as $f_\psi(x) = \int_Y y d\psi(y|x)$. Then the Bayes classifier is given by the sign of regression function $f_c = sgn(f_\psi)$. In this paper, we assume that there is a constant B such that for any $y \in Y$, $|y| \leq B$, which implies that $|f_\varphi(x)| \leq B$ for any $x \in X$ (Cucker & Smale, 2002).

To analyze the generalization ability of algorithm (1), we rewrite (1) as a regularization scheme (Chen et al., 2004; Zhang, 2004): define loss function $\ell(f, z)$ as

$$\ell(f, z) = \begin{cases} 0, & f(x)y > 1 \\ 1 - f(x)y, & f(x)y \leq 1. \end{cases} \tag{2}$$

The generalization error is $\mathcal{E}(f) = \mathrm{E}[\ell(f, z)]$. If we define the empirical error as $\mathcal{E}_m(f) = \frac{1}{m}\sum_{i=1}^m \ell(f, z_i)$, then algorithm (1) can be written as

$$f_{S,\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_m(f) + \lambda \|f\|_K^2 \right\}. \tag{3}$$

Here $\lambda = 1/(2C)$ is the regularization parameter.

Although we sometimes use generic kernels and RKHSs, in this paper we are mainly interested in Gaussian RBF kernels, which are the most widely used kernels in practice (Steinwart & Scovel, 2007). Recall that these kernels are of the form $K_\sigma(x, x') = \exp(-\sigma^2 \|x - x'\|_2^2), \ x, x' \in X$, where $\|x - x'\|_2^2$ is the squared Euclidean distance between x and x', $\sigma > 0$ is a free parameter whose inverse $1/\sigma$ is called the width of K_σ (Steinwart & Scovel, 2007). We denote the corresponding RKHS by $\mathcal{H}_K^\sigma$. Different from the previously known works on SVMC algorithm in Steinwart and Christmann (2008), Chen et al. (2004) and Steinwart and Scovel (2007), our goal of this paper is to bound the generalization ability of Gaussian RBF kernels SVMC based on u.e.M.c. samples.

2.2. Uniformly ergodic Markov chains

Suppose $(\mathcal{Z}, \mathcal{S})$ is a measurable space, a Markov chain is a sequence of random variables $\{Z_t\}_{t \geq 1}$ together with a set of transition probability measures $P^n(A|z_i), A \in \mathcal{S}, z_i \in \mathcal{Z}$, which is defined as

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

Thus $P^n(A|z_i)$ denotes the probability that the state z_{n+i} will belong to the set A after n-steps, starting from the initial state z_i at time i. It is common to denote the one-step transition probability by $P^1(A|z_i) = \text{Prob}\{Z_{i+1} \in A | Z_j, j < i, Z_i = z_i\}$. The fact that the transition probability does not depend on the values of Z_j prior to time i is the Markov property, that is, $P^n(A|z_i) = \text{Prob}\{Z_{n+i} \in A | Z_i = z_i\}$. This is expressed in words as "given the present state, the future and past states are independent".

Given two probabilities ν_1, ν_2 on the measure space $(\mathcal{Z}, \mathcal{S})$, we define the total variation distance between the two measures ν_1, ν_2 as $\|\nu_1 - \nu_2\|_{TV} = \sup_{A \in \mathcal{S}} |\nu_1(A) - \nu_2(A)|$. Thus we have the following definition of u.e.M.c. (Meyn & Tweedie, 1993; Vidyasagar, 2003).

Definition 1. A Markov chain $\{Z_t\}_{t \geq 1}$ is said to be uniformly ergodic if for some $\gamma < \infty$ and $0 < \rho < 1$,

$$\|P^n(\cdot|z) - \pi(\cdot)\|_{TV} \leq \gamma \rho^n, \quad \forall n \geq 1, n \in \mathbb{N}$$

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

Remark 1. A weaker condition than uniformly ergodic is V-geometrically ergodic (Meyn & Tweedie, 1993; Vidyasagar, 2003). The difference between V-geometrically ergodic and uniformly ergodic is that here the total variation distance between the n-step transition probability $P^n(\cdot|z)$ and the invariant measure π approaches zero at a geometric rate multiplied by $V(z)$ (Vidyasagar, 2003). Thus the rate of geometric convergence is independent of z, but the multiplicative constant is allowed to depend on z. Especially, if the space $\mathcal{Z}$ 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.

3. Estimating learning rates

To estimate the generalization ability of Gaussian RBF kernels SVMC algorithm, we should bound the excess misclassification error $\mathcal{R}(sgn(f_{S,\lambda})) - \mathcal{R}(f_c)$. Zhang (2004) established the relation between the excess misclassification error and excess generalization error for loss function (2)

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \le \mathcal{E}(f) - \mathcal{E}(f_\psi), \quad f : X \to \mathbb{R}. \tag{4}$$

This implies that the excess misclassification error $\mathcal{R}(sgn(f_{S,\lambda})) - \mathcal{R}(f_c)$ can be bounded by the excess generalization error $\mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_\psi)$. For the excess generalization error $\mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_\psi)$, the following error decomposition method is standard (Chen et al., 2004).

Lemma 1. *Let $f_\lambda = \arg\min_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) + \lambda\|f\|_K^2\right\}$, and $f_{S,\lambda}$ be defined as (3). Then we have $\mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_\psi) \le \mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_\psi) + \lambda\|f_{S,\lambda}\|_K^2$, which can be bounded by*

$$\{\mathcal{E}(f_{S,\lambda}) - \mathcal{E}_m(f_{S,\lambda}) + \mathcal{E}_m(f_\lambda) - \mathcal{E}(f_\lambda)\} + D(\lambda). \tag{5}$$

The first term of (5) is called the sample error (Cucker & Smale, 2001), which can be written as (Wu, Ying, & Zhou, 2006)

$$\left\{\mathrm{E}\zeta_1 - \frac{1}{m}\sum_{i=1}^m \zeta_1(z_i)\right\} + \left\{\frac{1}{m}\sum_{i=1}^m \zeta_2(z_i) - \mathrm{E}\zeta_2\right\}, \tag{6}$$

where $\zeta_1 = \ell(f_{S,\lambda}, z) - \ell(f_\psi, z)$, $\zeta_2 = \ell(f_\lambda, z) - \ell(f_\psi, z)$. The second term of (5) is called the approximation error (Cucker & Smale, 2001), which is defined as $D(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f_\psi) + \lambda\|f_\lambda\|_K^2$.

Steinwart and Scovel (2007) introduced the geometric noise exponent condition for the distribution on $\mathcal{Z}$, and then they established the bound of $D(\lambda)$ for the space $\mathcal{H}_K^\sigma$.

Definition 2 (*Steinwart & Scovel, 2007*)**.** Let $X \subset \mathbb{R}^d$ be compact and ψ be a probability measure on $X \times Y$. We say that ψ has geometric noise exponent $\alpha > 0$ if there exists a constant $C_0 > 0$ such that

$$\int_X |2\eta(x) - 1| \exp\left(\frac{-\tau_x^2}{t}\right)\psi_X(dx) \le C_0 t^{\alpha d/2}, \quad t > 0,$$

where ψ_X is the marginal probability measure of ψ on X, τ_x is defined as

$$\tau_x = \begin{cases} d(x, X_0 \cup X_1), & x \in X_{-1}, \\ d(x, X_0 \cup X_{-1}), & x \in X_1, \\ 0, & otherwise. \end{cases}$$

Here $d(x, A)$ denotes the distance of x to a set A with respect to the Euclidean norm, $X_{-1} = \{x \in X : \eta(x) < \frac{1}{2}\}$, $X_1 = \{x \in X : \eta(x) > \frac{1}{2}\}$, $X_0 = \{x \in X : \eta(x) = \frac{1}{2}\}$ and $\eta(x) = \mathrm{Prob}(y = 1|x)$.

Lemma 2 (*Steinwart & Scovel, 2007*)**.** *Let $\sigma > 0$, X be the closed unit ball of the Euclidean space $\mathbb{R}^d$ and $D(\lambda)$ be the approximation error function with respect to $\mathcal{H}_K^\sigma$. Furthermore, let ψ be a distribution on $X \times Y$ that has geometric noise exponent $0 < \alpha < \infty$ with constant C_0 in Definition 2. Then there is a constant $C_d > 0$ depending only on d such that for all $\lambda > 0$,*

$$D(\lambda) \le C_d\left(\sigma^d\lambda + C_0(2d)^{\alpha d/2}\sigma^{-\alpha d}\right).$$

In particular, if σ satisfies $\sigma(\lambda) = \lambda^{-1/[(\alpha+1)d]}$, then there exists a constant C_1 such that $D(\lambda) \le C_1\lambda^{\frac{\alpha}{\alpha+1}}$.

To estimate the sample error (6), we have to regulate the capacity of function set since the minimization (3) is taken over the discrete quantity $\mathcal{E}_m(f)$. Here the capacity is measured by the covering number (De Vito, Caponnetto, & Rosasco, 2005; Smale & Zhou, 2005; van der Vaart & Wellner, 1996; Zhang, 2004).

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

Let $B_{\mathcal{H}}(R) := \{f \in \mathcal{H}_K : \|f\|_K \le R\}$ be the closed ball of $\mathcal{C}(X)$. Then the covering number of $B_{\mathcal{H}}(1)$ is well defined (Steinwart & Scovel, 2007; van der Vaart & Wellner, 1996). Let $B_{\mathcal{H}}^\sigma(R) := \{f \in \mathcal{H}_K^\sigma : \|f\|_K \le R\}$. For any $\epsilon > 0$, we denote the covering number of $B_{\mathcal{H}}^\sigma(1)$ as $\mathcal{N}(B_{\mathcal{H}}^\sigma(1), \epsilon)$. Steinwart and Scovel (2007) established the following bound on the covering number $\mathcal{N}(B_{\mathcal{H}}(1), \epsilon)$.

Lemma 3. *Let $\sigma \ge 1$, $0 < p < 1$ and $X \subset \mathbb{R}^d$ be a compact subset with nonempty interior. Then there is a constant $C_{p,d} > 0$ independent of σ such that for all $\epsilon > 0$,*

$$\ln \mathcal{N}(B_{\mathcal{H}}^\sigma(1), \epsilon) \le C_{p,d} \cdot \sigma^{(1-p/4)d}\epsilon^{-p}.$$

Then our main results are stated as follows:

Proposition 1. *Set $m^{(\beta)} = \lfloor m\lceil\{8m/\ln(1/\rho)\}^{\frac{1}{2}}\rceil^{-1}\rfloor$, where $\lfloor u\rfloor(\lceil u\rceil)$ denotes the greatest (least) integer less (greater) than or equal to u. Assume that $\{z_i\}_{i=1}^m$ is a u.e.M.c. sample and $R \ge B$. Then for any $0 < \delta < 1$,*

$$\mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_\psi) + 2\lambda\|f_{S,\lambda}\|_K^2 \le 4D(\lambda) + \frac{7(\kappa\sqrt{D(\lambda)/\lambda} + B)\ln(C_2/\delta)}{3m^{(\beta)}} + 8R\cdot\varepsilon(m,\delta)$$

holds true with probability at least $1 - 2\delta$, where $C_2 = 1 + \gamma e^{-2}$, and $\varepsilon(m,\delta) \le \max\{\overline{m}, \widetilde{m}\}$,

$$\overline{m} = \frac{80(\kappa+1)\ln(C_2/\delta)}{3m^{(\beta)}},$$

$$\widetilde{m} = \left[\frac{80C_{p,d}\sigma^{(1-p/4)d}(\kappa+1)}{3m^{(\beta)}}\right]^{\frac{1}{1+p}}.$$

For the proof of Proposition 1, refer to Appendix B. As an application of Proposition 1, we establish the learning rate of Gaussian RBF kernels SVMC for u.e.M.c. samples.

Theorem 1. *Let $\{z_i\}_{i=1}^m$ be a u.e.M.c. sample. Taking $\lambda = \left(\frac{1}{m}\right)^\vartheta$, for any $\epsilon > 0$, $0 < \delta < 1$ and $m \ge m_\delta$, there exists a constant $\widehat{C}$ independent of m such that*

$$\mathcal{R}(sgn(f_{S,\lambda})) - \mathcal{R}(f_c) \le \widehat{C}\left(\frac{1}{m}\right)^\theta,$$

holds true with confidence at least $1 - \delta$, where $m_\delta = \max\{m'_\delta, m''_\delta\}$, $m'_\delta = \max\{\ln(1/\rho)/8, 128/(\ln 1/\rho)\}$,

$$m''_\delta = \frac{5\cdot 2^{11}(\kappa+1)^2(\ln(2C_2/\delta))^2}{9\ln(1/\rho)}\cdot\left[\frac{\ln(2C_2/\delta)}{C_{p,d}\sigma^{(1-p/4)d}}\right]^{\frac{2}{p}},$$

$$\vartheta = \min\left\{\alpha + 1, \frac{(\alpha+1)d}{2 + (1+p)(2\alpha+1)d}\right\},$$

$$\theta = \min\left\{\alpha, \frac{\alpha d}{2 + (1+p)(2\alpha+1)d} - \epsilon\right\}.$$

For the proof of Theorem 1, refer to Appendix B. By Theorem 1, we can find that for $p \to 0$ and sufficiently large α, the learning rate obtained in Theorem 1 is arbitrarily close to the best kernel independent learning rate $m^{-\frac{1}{2}}$ (see, e.g., De Vito et al., 2005; Smale & Zhou, 2005). In order to improve the learning rate obtained in Theorem 1, we use the strongly mixing property of uniformly ergodic Markov chains. That is, Rosenblatt (1972) proved that if a stationary Markov chain satisfies both uniform ergodicity and mixing (in the ergodic-theoretic sense), then it is strongly mixing.

Definition 4 (*Strongly Mixing*)**.** The sequence $\{\xi_t\}$ is called α-mixing, or strongly mixing, if for any $k \to \infty$,

$$\sup_{A\in\mathcal{A}_{-\infty}^{0}, B\in\mathcal{A}_{k}^{\infty}} \{|P(A\cap B) - P(A)P(B)|\} = \alpha(k) \to 0,$$

where $\alpha(k)$ is called the α-mixing coefficient, $\mathcal{A}_{-\infty}^{0}$ and $\mathcal{A}_{k}^{\infty}$ denote the σ-algebra generated by random variables ξ_i, $i \le 0$ and ξ_i, $i \ge k$, respectively.

Assumption 1 (*Geometrically α-mixing (Vidyasagar, 2003)*)**.** Assume that the α-mixing coefficient of sequence $\{\xi_t\}$ satisfies $\alpha(k) \le \overline{\alpha}\exp(-ck^{\beta})$, $k \ge 1, k \in \mathbb{N}$ for some $\overline{\alpha} > 0$, $\beta > 0$, and $c > 0$.

Remark 2. Assumption 1 is satisfied by a large class of processes (Modha & Masry, 1996), for example, certain linear processes (which include certain ARMA processes) satisfy the assumption with $\beta = 1$ (Withers, 1981), and many Markov processes (which includes certain bilinear processes, nonlinear ARX processes, and ARH processes) satisfy Assumption 1 (Davydov, 1973; Steinwart et al., 2009). As a trivial example, i.i.d. random variables satisfy Assumption 1 with $\beta = \infty$.

By Assumption 1, we establish the following learning rate of Gaussian RBF kernels SVMC for u.e.M.c. samples.

Theorem 2. *Let $\{z_i\}_{i=1}^{m}$ be a u.e.M.c. sample. Take $\lambda = \left(\frac{1}{m}\right)^{\vartheta}$, for any $\epsilon > 0, 0 < \eta < 1$ and $m \ge m_\eta$, there exists a constant $\widetilde{C}$ independent of m such that*

$$\mathcal{R}(sgn(f_{S,\lambda})) - \mathcal{R}(f_c) \le \widetilde{C}\left(\frac{1}{m}\right)^{\theta}$$

holds true with confidence at least $1-\eta$, where m_η is a constant given by $m_\eta \ge \max\{m'_\eta, m''_\eta\}$, $m'_\eta = \max\{/c8, 2^{2+5/\beta}/c^{\beta}\}$,

$$m''_\eta = \left(\frac{80(\kappa+1)\ln(C_4/\eta)}{3\cdot 2^{\frac{-2\beta+5}{\beta+1}} c^{\frac{1}{\beta+1}}}\right)^{\frac{\beta+1}{\beta}} \left(\frac{\ln(C_4/\eta)}{C_{p,d}\sigma^{(1-p/4)d}}\right)^{\frac{\beta+1}{p\beta}},$$

$$\vartheta = \min\left\{\alpha+1, \frac{(\alpha+1)d}{2+(1+p)(2\alpha+1)d}\right\},$$

$$\theta = \min\left\{\alpha, \frac{2\alpha d\beta}{(\beta+1)[2+(1+p)(2\alpha+1)d]} - \epsilon\right\},$$

and $C_4 = 1 + 4e^{-2}\overline{\alpha}$.

For the proof of Theorem 2, refer to Appendix B. To have a better understanding of Theorem 2, we compare Theorem 2 with the previously known results as follows: Zhang and Tao (2012) studied the generalization bounds of ERM learning processes for continuous-time Markov chains under three assumptions (see Conditions C1, C2 and C3 in Zhang and Tao (2012)), and obtained the learning rate $m^{-\frac{1}{1.3}}$ (see inequality (47) in Zhang and Tao (2012)). Zou, Peng et al. (2013) obtained the weak learning rate $(m^{(\beta)})^{-\frac{1}{4}}$ with $m^{(\beta)} = O(m^{-\frac{1}{2}})$. While by Theorem 2, we can find that for sufficiently large α, β and $p \to 0$, θ is arbitrarily close to 1. This implies that for sufficiently small p and larger α and β, the learning rate obtained in Theorem 2 is arbitrarily close to m^{-1}, which is the optimal learning rate of i.i.d. samples in statistical learning theory (Steinwart and Scovel (2007) established the similar learning rate for Gaussian RBF kernels SVMC with i.i.d. samples. Chen et al. (2004) obtained the similar learning rate for SVMC with i.i.d. samples. Tong, Chen, and Peng (2009) established the similar learning rate for SVM regression with i.i.d. samples). This implies that the results obtained in this paper extend the classical results of Gaussian RBF kernels SVMC with i.i.d. samples in Steinwart and Scovel (2007) to the case of u.e.M.c. samples.

Table 1
5 real-world datasets.

Dataset	Training size	Test size	Input dimension
Abalone	2 089	2 088	8
Shuttle	43 500	14 500	9
Magic	12 680	6 340	10
Waveform	4 600	400	21
Splice	2 175	1 000	60

4. Numerical studies

Inspired by the idea from MCMC methods (Curnow, 1988; Laarhouen & Aarts, 1987), Zou, Peng et al. (2013) introduced a Markov sampling algorithm such that Markov chain samples can be generated from a given dataset D, and then they studied the learning performance of SVMC based on Markov sampling for linear prediction models. In this paper we generalize the study on the learning performance of SVMC algorithm with Markov chain samples based on linear prediction models to the case of nonlinear prediction models, Gaussian RBF kernels. We give here a slightly modified version of Markov sampling in Zou, Peng et al. (2013) that suits our needs.

Remark 3. Since we have only the dataset D, to generate u.e.M.c. samples, we introduce a technical condition f_0 and two technical parameters k and q: first, to define the transition probability, in this paper we introduce the preliminary learning model f_0. The reason is that under the technical condition, we can compute easily the transition probabilities P (or P', P'') and P, P' and P'' are always positive. Thus by the theory of Markov chain in (Qian & Gong, 1998), we can conclude that the generated sequence $\{z_1, z_2, \ldots, z_t\}$ by Algorithm 1 is a u.e.M.c. sequence. Second, to generate quickly Markov chain samples, we also introduce the continuously reject number k and the constant q. Since for some datasets, generating Markov chain samples is very time-consuming by using the sampling method in Zou, Peng et al. (2013). Namely, as the loss $\ell(f, z_t)$ of current sample z_t is very small, then the candidate sample z_* will not be accepted since the acceptance probability $P = \min\{1, e^{-\ell(f_0,z_*)}/e^{-\ell(f_0,z_t)}\}$ is very small. In the following experiments, we take $k = 5$ and $q = 1.2$. In addition, to generate the balance training samples, we introduce the notions m_+, m_- and $m\%2$. The case of SVMC with unbalance training samples is under our current investigation. Compared the above Markov sampling with randomly independent sampling, we can find that randomly independent sampling can be regarded as the special case of Algorithm 1, that is, all the acceptance probabilities P, P' and P'' in Algorithm 1 are always 1.

4.1. Experimental results

We present the numerical study on the learning performance of Gaussian RBF kernels SVMC for 5 real-world datasets: Abalone, Magic, Shuttle (http://archive.ics.uci.edu/ml/datasets.html), Splice, Waveform (http://www.fml.tuebingen.mpg.de/Members/raetsch/benchmark). We present the information of these datasets in Table 1.

For randomly independent sampling, we decompose the experiment into two steps: first, a training set S_T of m training samples was generated randomly from a given dataset. We use Gaussian RBF kernels SVMC to train the set S_T, and then we test it on the given test set. Second, after the experiment had been repeated for 50 times, the misclassification rates were presented in Tables 2 and 3, where MR (i.i.d.) denotes the misclassification rates based on randomly independent sampling.

For Markov sampling, we first generate a training set S'_T of m training samples by Algorithm 1. Then we use Gaussian RBF

Algorithm 1 Markov sampling for Gaussian kernels SVMC

Step 1: Draw randomly $N_1 (N_1 \leq m)$ training samples $\{z_i\}_{i=1}^{N_1}$ from a dataset D. Use the Gaussian RBF kernels SVMC algorithm to train these samples, and obtain a preliminary learning model f_0. Set $m_+ = 0$ and $m_- = 0$. m is the number of training samples, and m_+ and m_- denote the number of training samples which label are $+1$ and -1, respectively.

Step 2: Draw randomly a sample from D and denote it the current sample z_t. If $m\%2 = 0$, $m\%2$ denotes the remainder of m divided by 2. Then set $m_+ = m_+ + 1$ if the label of z_t is $+1$, or set $m_- = m_- + 1$ if the label of z_t is -1.

Step 3: Draw randomly a sample from D and denote it the candidate sample z_*.

Step 4: Calculate the ratio P of $e^{-\ell(f_0,z)}$ at the sample z_* and the sample z_t, $P = e^{-\ell(f_0,z_*)}/e^{-\ell(f_0,z_t)}$.

Step 5: If $P = 1, y_t = -1$ and $y_* = -1$ accept the candidate sample z_* with probability $P' = e^{-y_* f_0}/e^{-y_t f_0}$. If $P = 1, y_t = 1$ and $y_* = 1$ accept the candidate sample z_* with probability $P' = e^{-y_t f_0}/e^{-y_* f_0}$. If $P = 1$ and $y_t y_* = -1$ or $P > 1$ or $P < 1$, accept the candidate sample z_* with probability P. If there are k candidate samples z_* cannot be accepted continuously, then set $P'' = qP$ and then with probability P'' accept the sample z_*. Set $z_{t+1} = z_*, m_+ = m_+ + 1$ if the label of z_t is $+1$, or set $m_- = m_- + 1$ if the label of z_t is -1.

Step 6: If $m_+ < \frac{m}{2}$ or $m_- < \frac{m}{2}$ then return to Step 3, else stop it.

Table 2
Misclassification rates for 1000 training samples.

Dataset	MR (i.i.d.)	MR (Markov)
Abalone	0.3632 ± 0.1379	0.2216 ± 0.0023
Shuttle	0.0723 ± 0.0114	0.0619 ± 0.0042
Magic	0.2167 ± 0.0043	0.2133 ± 0.0023
Waveform	0.2116 ± 0.0025	0.2036 ± 0.0019
Splice	0.2650 ± 0.0039	0.2216 ± 0.0023

Table 3
Misclassification rates for 1500 training samples.

Dataset	MR (i.i.d.)	MR (Markov)
Abalone	0.4646 ± 0.1280	0.2231 ± 0.0069
Shuttle	0.0673 ± 0.0074	0.0594 ± 0.0053
Magic	0.2242 ± 0.0049	0.2159 ± 0.0010
Waveform	0.2226 ± 0.0017	0.2119 ± 0.0021
Splice	0.2648 ± 0.0029	0.2567 ± 0.0023

kernels SVMC to train the set S'_T, and test it on the same test set. After the experiment had been repeated for 50 times, the misclassification rates were presented in Tables 2 and 3, where MR (Markov) denotes the misclassification rates based on Markov sampling.

From Tables 2 and 3, we can find that for the same size of training samples and the same test set, all the means of misclassification rates of Gaussian RBF kernels SVMC based on Markov sampling are smaller than that of randomly independent sampling, and all the standard deviations of misclassification rates of Gaussian RBF SVMC based on Markov sampling are also smaller than that of randomly independent sampling except Waveform for 1500 training samples. To simplify the process of these experiments, we take $N_1 = m$ in the above experiments. In addition, the parameters λ and σ of Gaussian RBF kernels SVMC based on randomly independent sampling and Markov sampling are chosen by the method of 5-fold cross-validation, respectively.

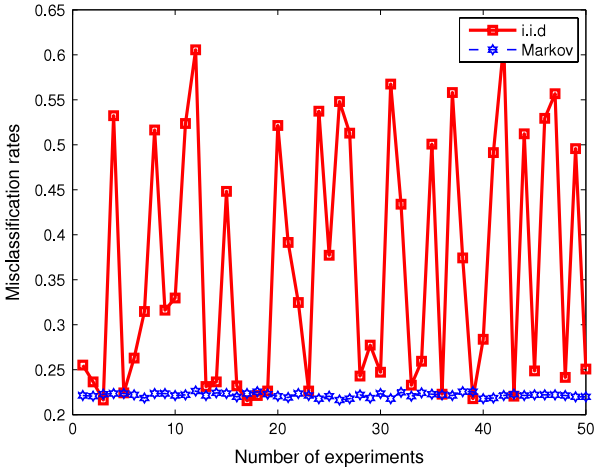

Fig. 1. 50 times experimental misclassification rates for Abalone and $m = 1000$.

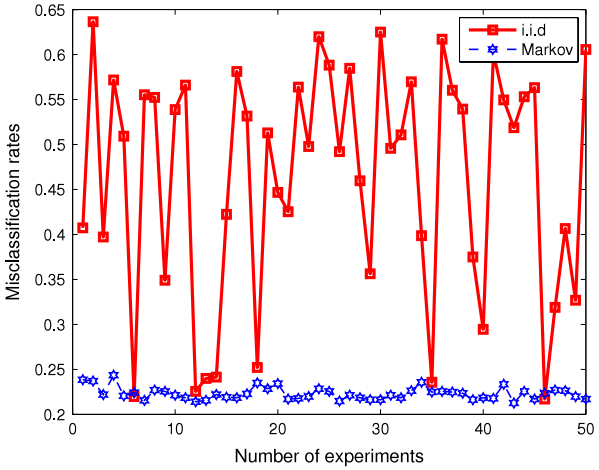

Fig. 2. 50 times experimental misclassification rates for Abalone and $m = 1500$.

4.2. Discussions and comparisons

To have a better understanding of learning performance of Gaussian RBF kernels SVMC based on Markov sampling, we also present the following figures on 50 times experimental results of Gaussian RBF kernels SVMC based on Markov sampling and randomly independent sampling. Here "red square" denotes the results based on randomly independent sampling, "blue hexagram" denotes the results based on Markov sampling. The numbers on the vertical axis of figures denote the misclassification rates, and the numbers on the horizontal axis of figures denote the experimental times.

In Figs. 1 and 2, we can find that for Abalone, 1000 and 1500 training samples, the 50 times misclassification rates of Gaussian RBF kernels SVMC based on Markov sampling are smaller than that of randomly independent sampling except at most 3 times experimental results.

In Figs. 3–5, we can find that for Shuttle, 1000 and 1500 training samples, the 50 times misclassification rates of Gaussian RBF kernels SVMC based on Markov sampling are smaller than that of randomly independent sampling except at most 12 times experimental results. While for 4000 training samples, almost all the 50 times misclassification rates Gaussian RBF kernels SVMC based on Markov sampling are smaller than that of randomly independent sampling.

In Figs. 6 and 7, we can find that for Magic and 1000 training samples, the 50 times misclassification rates of Gaussian RBF

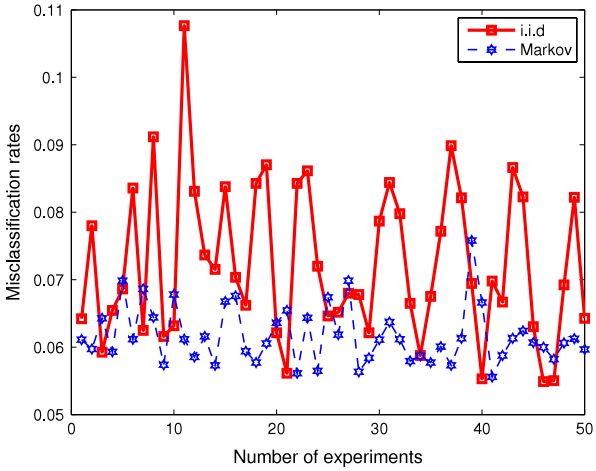

Fig. 3. 50 times experimental misclassification rates for Shuttle and $m = 1000$.

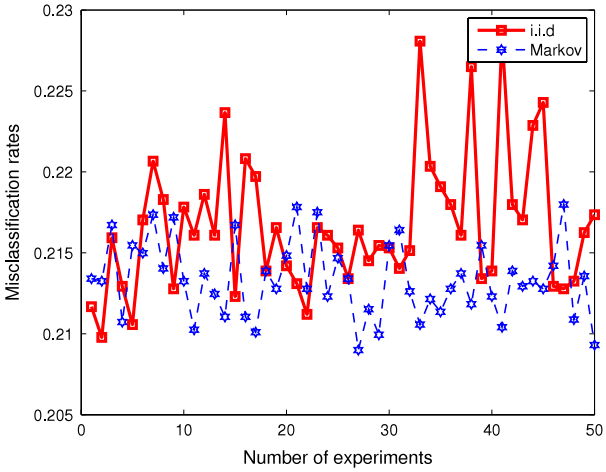

Fig. 6. 50 times experimental misclassification rates for Magic and $m = 1000$.

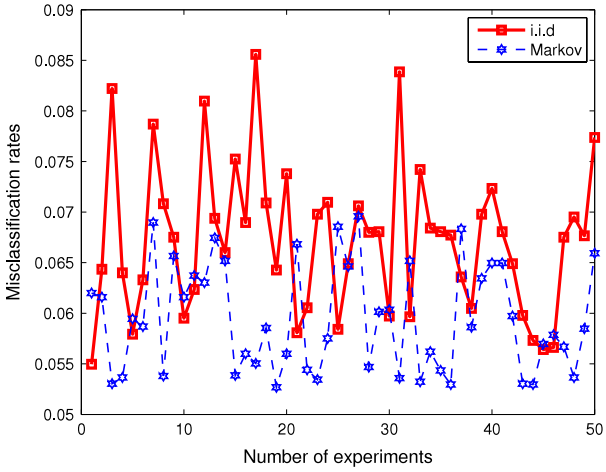

Fig. 4. 50 times experimental misclassification rates for Shuttle and $m = 1500$.

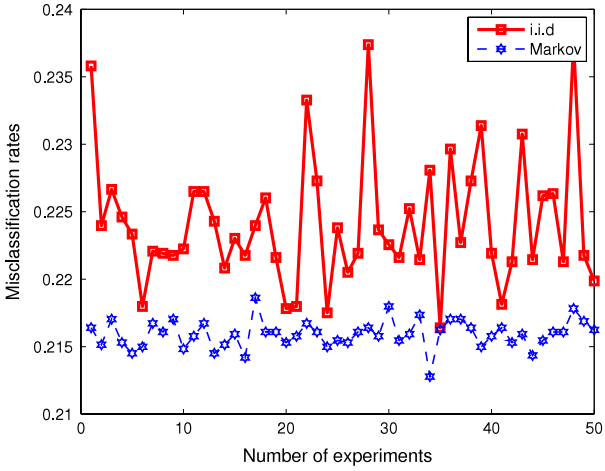

Fig. 7. 50 times experimental misclassification rates for Magic and $m = 1500$.

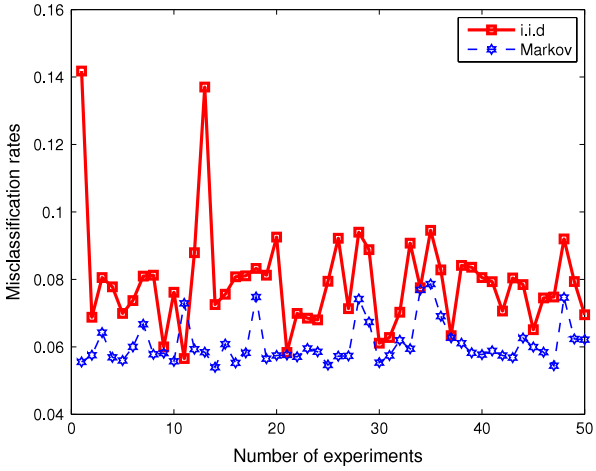

Fig. 5. 50 times experimental misclassification rates for Shuttle and $m = 4000$.

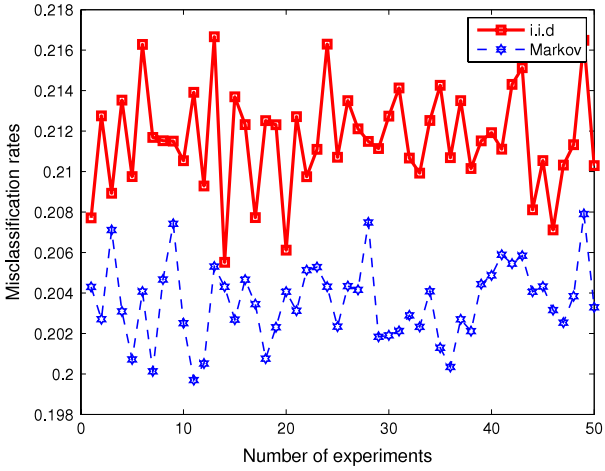

Fig. 8. 50 times experimental misclassification rates for Waveform and $m = 1000$.

kernels SVMC based on Markov sampling are smaller than that of randomly independent sampling except 14 times experimental results. While for 1500 training samples, all the 50 times experimental results of Gaussian RBF kernels SVMC based on Markov sampling are better than that of randomly independent sampling.

In Figs. 8 and 9, we can find that for Waveform, 1000 and 1500 training samples, all the 50 times misclassification rates of Gaussian RBF kernels SVMC based on Markov sampling are smaller than that of randomly independent sampling.

In Figs. 10 and 11, we can find that for Splice, 1000 and 1500 training samples, all the 50 times misclassification rates of Gaussian RBF kernels SVMC based on Markov sampling are smaller than that of randomly independent sampling. In addition, we also compare the total times (second) of training and sampling based on Markov sampling with that of randomly independent sampling

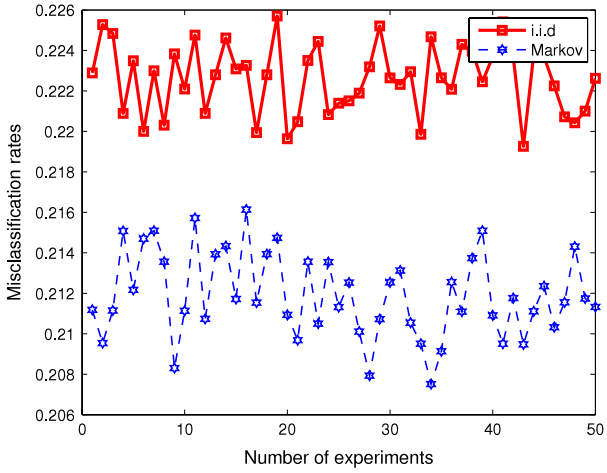

Fig. 9. 50 times experimental misclassification rates for Waveform and $m = 1500$.

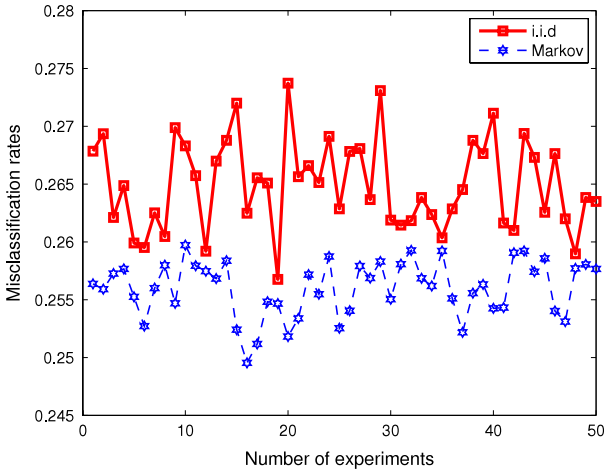

Fig. 10. 50 times experimental misclassification rates for Splice and $m = 1000$.

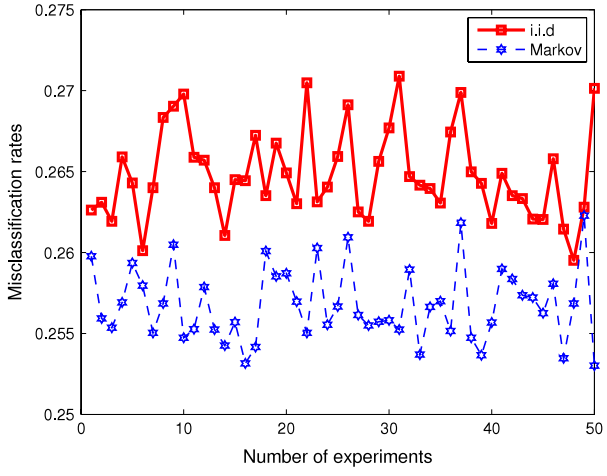

Fig. 11. 50 times experimental misclassification rates for Splice and $m = 1500$.

in Table 4. Here "Times (Markov)" and "Times (i.i.d.)" denote the total times of training and sampling based on Markov sampling and randomly independent sampling, respectively. "Abalone-1000" denotes the average times of training and sampling of 50 times experiments for Abalone with 1000 training samples.

From Table 4, we can find that, the average times of training and sampling based on Markov sampling are less than that of randomly independent sampling except Shuttle with 1000 and 1500 training samples, Waveform with 1000 and 1500 training samples.

Table 4
Comparisons for the total times of training and sampling.

Dataset	Times (i.i.d.)	Times (Markov)
Abalone-1000	38.0543	18.3295
Shuttle-1000	29.3699	33.8851
Magic-1000	40.9011	20.3814
Waveform-1000	57.5039	59.8591
Splice-1000	39.3179	27.8640
Abalone-1500	136.7234	128.9456
Shuttle-1500	107.4241	108.7373
Magic-1500	202.0325	125.6450
Waveform-1500	79.1662	217.7864
Splice-1500	136.1976	77.4119
Shuttle-4000	566.9826	540.9785

Finally, we interpret the learning performance of Gaussian RBF kernels SVMC based on Markov sampling as follows: first, in the process of Markov sampling, the candidate samples z_* are accepted with different acceptance probabilities, while for random sampling, all the candidate samples z_* are accepted with probability 1. Second, by these acceptance probabilities defined in Step 5 of Algorithm 1, we can find that the samples that have the same or similar property (with respect to the loss function $\ell(f, z)$) will be accepted with another probability P', which implies that the Markov chain samples are different compared to random sampling. More importantly, after many times transitions, the samples that close to the interface of two classes data will be sampled and be accepted with high probabilities, which are the reasons that the learning performance of Gaussian RBF kernels SVMC based on Markov sampling is better than that of randomly independent sampling.

5. Conclusions

In this paper we study the generalization performance of Gaussian RBF kernels SVMC algorithm based on u.e.M.c. samples. We not only establish the learning rates of Gaussian RBF kernels SVMC algorithm based on u.e.M.c. samples, but also obtain the fast learning rates for Gaussian RBF kernels SVMC algorithm with u.e.M.c. samples by using the strongly mixing property of u.e.M.c. samples. The learning rate obtained in this paper is same as the optimal learning rate of learning algorithm that established in Chen et al. (2004), Steinwart and Scovel (2007), Tong et al. (2009). This implies that the results obtained in this paper extend the classical results of SVMC based on i.i.d. samples in Chen et al. (2004) and Steinwart and Scovel (2007) to the case of u.e.M.c. samples. To our knowledge, these studies here are the first works on this topic. In order to study the learning performance of Gaussian RBF kernels SVMC based on Markov sampling, we also present the numerical studies on benchmark repository using Gaussian RBF kernels SVMC based on Markov sampling. The experimental results show that the Gaussian RBF kernels SVMC based on Markov sampling can provide smaller misclassification rates compared to randomly independent sampling.

Along the line of the present work, several open problems deserve further research, for example, the study on the learning performance of Gaussian RBF kernels SVMC based on Markov chain samples for the data sets with higher input dimensions, and the study on the Markov sampling algorithm for regression problem and online learning algorithms. 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. This work is supported in part by NSFC project (2013CB329404, 11131006, 11371007, 61370002), Multi-Year Research of University of Macau (No. MYRG205(Y1-L4)-FST11-TYY, No. MYRG187(Y1-L3)-FST11-TYY), Start-up Research of University of Macau (No. SRG010-FST11-TYY).

Appendix A. Main tools

Our main tools are as follows: let $\{\xi_i\}_{i=-\infty}^{\infty}$ be a stationary process defined on a probability space $(\xi^\infty, \mathcal{S}^\infty, \tilde{P})$. For $-\infty < i < \infty$, let $\mathcal{A}_{-\infty}^k$ denote the σ-algebra generated by random variables $\xi_i, i \le k$, and similarly let $\mathcal{A}_k^\infty$ denote the σ-algebra generated by random variables $\xi_i, i \ge k$. 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 ξ_i. Let $\bar{\mathcal{A}}_1^{k-1}$ denote the σ-algebra generated by the random variables $\xi_i, i \le 0$ as well as $\xi_j, j \ge k$.

Definition 5 (*Vidyasagar, 2003*)**.** The sequence $\{\xi_t\}$ is called geometrically β-mixing, if there exist constants $\nu > 0$ and $\lambda_1 < 1$ such that for any $k \ge 1, k \in \mathbb{N}$

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

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

Lemma 4 (*Vidyasagar, 2003*)**.** *Suppose $\{\xi_t\}$ is a β-mixing process on a probability space $(\xi^\infty, \mathcal{S}^\infty, \tilde{P})$. Suppose $g : \xi^\infty \to \mathbb{R}$ is essentially bounded and depends only on the variables ξ_i, $0 \le i \le l$. Let $\tilde{P}_0$ denote the one-dimensional marginal probability of each of ξ_i. Then*

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

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 5 (*Vidyasagar, 2003*)**.** *Let $\{\xi_t\}$ be a V-geometrically ergodic Markov chain. Then the sequence $\{\xi_t\}$ is geometrically β-mixing, and the β-mixing coefficient $\beta(k)$ is given by*

$$\begin{aligned}\beta(k) &= \mathrm{E}\left\{\|P^k(\cdot|\xi) - \pi(\cdot)\|_{TV}, \pi\right\} \\ &= \int \|P^k(\cdot|\xi) - \pi(\cdot)\|_{TV} \pi(d\xi).\end{aligned}$$

Lemma 6 (*Saunders, Gammerman, & Vovk, 1998*)**.** *Let W be a random variable such that $\mathrm{E}(W) = 0$, and W satisfies the Bernstein moment condition, that is, for some $K_1 > 0$,*

$$\mathrm{E}|W|^k \le \frac{Var(W)}{2} k! K_1^{k-2} \tag{7}$$

for all $k \ge 2$. Then for all $0 < \zeta < 1/K_1$,

$$\mathrm{E}[\exp(\zeta W)] \le \exp\left[\frac{\zeta^2 \mathrm{E}|W|^2}{2(1 - \zeta K_1)}\right].$$

*In particular, If $|W| \le 3K_1$ almost everywhere, then the Bernstein moment condition (7) holds true (*Modha & Masry, 1996*).*

By Definition 5, Lemmas 4–6, we establish the following concentration inequality for u.e.M.c. samples.

Lemma 7. *Let $\{z_i\}_{i=1}^m$ be a u.e.M.c. sample. Denote $V_i = \phi(z_i)$, where ϕ is a real-valued measure function and $m^{(\beta)} = \lfloor m\lceil\{8m/\ln(1/\rho)\}^{\frac{1}{2}}\rceil^{-1}\rfloor$, where $\lfloor u\rfloor(\lceil u\rceil)$ denotes the greatest (least) integer less (greater) than or equal to u. Assume that $|V_i| \le d_1$ for any $1 \le i \le m$ and $\mathrm{E}[V_1] = 0$. Then for any $\varepsilon > 0$,*

$$\mathrm{Prob}\left\{\frac{1}{m}\sum_{i=1}^m V_i \ge \varepsilon\right\} \le (1 + \gamma e^{-2}) \exp\left\{\frac{-\varepsilon^2 m^{(\beta)}}{2(\mathrm{E}|V_1|^2 + \varepsilon d_1/3)}\right\}.$$

Proof. We decompose the proof into three steps.

Step 1: By Remark 1, uniformly ergodic Markov chain is V-geometrically ergodic. Then by Lemma 5, we have that uniformly ergodic Markov chain is geometrically β-mixing,

$$\beta(k) = \mathrm{E}\left\{\|P^k(\cdot|z) - \pi(\cdot)\|_{TV}, \pi\right\} \le \gamma\rho^k. \tag{8}$$

Thus we can use the β-mixing property of u.e.M.c. to prove Lemma 7 as follows: we decompose the index set $I = \{1, \ldots, m\}$ into different parts by following the idea from Vidyasagar (2003), 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 notational clutter. Let $r = m - kl$, and define

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

Let $p_i = |I_i|/m$ for $i = 1, 2, \ldots, k$, and define $a_m(z) = \frac{1}{m}\sum_{i=1}^m V_i$, $b_i(z) = \frac{1}{|I_i|}\sum_{j\in I_i} V_j$. Then we have $\frac{1}{m}\sum_{i=1}^m V_i = a_m(z) = \sum_{i=1}^k p_i b_i(z)$.

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

$$\exp[\tau a_m(z)] = \exp\left[\sum_{i=1}^k p_i \tau b_i(z)\right] \le \sum_{i=1}^k p_i \exp[\tau b_i(z)].$$

It follows that

$$\mathrm{E}(e^{\tau a_m(z)}, \tilde{P}) \le \sum_{i=1}^k p_i \mathrm{E}(e^{\tau b_i(z)}, \tilde{P}). \tag{9}$$

Since

$$\begin{aligned}\exp[\tau b_i(z)] &= \exp\left[\frac{\tau}{|I_i|}\sum_{j\in I_i} V_j\right] = \prod_{j\in I_i} \exp\left(\frac{\tau V_j}{|I_i|}\right) \\ &\le \left[\exp\left(\frac{\tau d}{|I_i|}\right)\right]^{|I_i|} \le e^{\tau d},\end{aligned}$$

where in the last step we use the assumption $|V_j| \le d$. Note that the quantities $\mathrm{E}(e^{\tau b_i(z)}, \tilde{P})$ are all the same since the stochastic is stationary. Moreover, since the components in the index set I_i are separated by at least k. By Lemma 4, we have that for any $\tau > 0$

$$\begin{aligned}\mathrm{E}(e^{\tau b_i(z)}, \tilde{P}) &\le (|I_i| - 1)\beta(k)\|e^{\tau b_i(z)}\|_\infty + \mathrm{E}(e^{\tau b_i(z)}, \tilde{P}_0^\infty) \\ &\le (|I_i| - 1)\beta(k)e^{\tau d} + \mathrm{E}(e^{\tau b_i(z)}, \tilde{P}_0^\infty).\end{aligned} \tag{10}$$

Since under the measure $\tilde{P}_0^\infty$, the various z_i are independent, by Lemma 6, we have that for any $0 < \tau < 3|I_i|/d$

$$\begin{aligned}\mathrm{E}(e^{\tau b_i(z)}, \tilde{P}_0^\infty) &= \mathrm{E}\left[\prod_{j\in I_i} \exp(\tau V_j/|I_i|), \tilde{P}_0^\infty\right] \\ &= \prod_{j\in I_i} \mathrm{E}\left[\exp(\tau V_j/|I_i|), \tilde{P}_0^\infty\right] \\ &= \left\{\mathrm{E}\left[\exp\left(\frac{\tau V_1}{|I_i|}\right), \tilde{P}_0\right]\right\}^{|I_i|} \\ &\le \exp\left[\frac{\tau^2 \mathrm{E}|V_1|^2}{2|I_i|(1 - \tau d/3|I_i|)}\right].\end{aligned}$$

By inequality (10), we have that for any $3|I_i|/d > \tau > 0$

$$\mathrm{E}(e^{\tau b_i(z)}, \tilde{P}) \le \exp\left[\frac{\tau^2 \mathrm{E}|V_1|^2}{2|I_i|(1-\tau d/3|I_i|)}\right] + (|I_i|-1)\beta(k)e^{\tau d}.$$

Thus by inequality (9) and the above inequality, we have that for any $3|I_i|/d > \tau > 0$

$$\mathrm{E}(e^{\tau a_m(z)}, \tilde{P}) \le \sum_{i=1}^{k} p_i \left\{\exp\left[\frac{\tau^2 \mathrm{E}|V_1|^2}{2|I_i|(1-\tau d/3|I_i|)}\right] + (|I_i|-1)\beta(k)e^{\tau d}\right\}. \tag{11}$$

Step 2: We now bound the second term on the right-hand side of inequality (11) which is denoted henceforth by ϕ_i, $1 \le i \le k$. By inequality (8), we have that for any $0 < \tau \le 3|I_i|/d$,

$$\begin{aligned}\phi_i &= \exp\left[\frac{\tau^2 \mathrm{E}|V_1|^2}{2|I_i|(1-\tau d/3|I_i|)}\right] + (|I_i|-1)\beta(k)e^{\tau d} \\ &\le \exp\left[\frac{\tau^2 \mathrm{E}|V_1|^2}{2|I_i|(1-\tau d/3|I_i|)}\right] + e^{|I_i|}e^{-2}\gamma\rho^k \cdot e^{\tau d} \\ &\le \exp\left[\frac{\tau^2 \mathrm{E}|V_1|^2}{2|I_i|(1-\tau d/3|I_i|)}\right] + \gamma e^{-2} e^{\{k\ln\rho + 4|I_i|\}}.\end{aligned}$$

The above inequality follows from the fact that $|I_i - 1| \le e^{|I_i|-2}$ for $|I_i| \ge 2$. We require $\exp\{k\ln\rho + 4|I_i|\} \le 1$. But $|I_i| \le (m/k+1)$; thus the bound holds if $4(m/k+1) \le k\ln(1/\rho)$ or $4(m+k) \le k^2\ln(1/\rho)$. Since $m+k \le 2m$, then the bound holds if $\{8m/\ln(1/\rho)\}^{\frac{1}{2}} \le k$. Let $k = \lceil\{8m/\ln(1/\rho)\}^{\frac{1}{2}}\rceil$. Since for all $i = 1, \ldots, k$, $|I_i| \ge l$, and $l = \lfloor m/k \rfloor$, we have

$$\phi_i \le \exp\left[\frac{\tau^2 \mathrm{E}|V_1|^2}{2l(1-\tau d/3l)}\right] + \gamma e^{-2}. \tag{12}$$

Since inequality (12) is true for all τ, $0 < \tau \le 3|I_i|/d$. To make the constraint uniform over all i, we then require τ satisfy $0 < \tau < 3l/d \le 3|I_i|/d$. Since $\tau^2\mathrm{E}|V_1|^2/2l(1-\tau d/3l) > 0$, we have that for any $0 < \tau < 3l/d$, $\phi_i \le (1+\gamma e^{-2})\exp\left[\tau^2\mathrm{E}|V_1|^2/2l(1-\tau d/3l)\right]$. Returning to inequality (11), we have that for $0 < \tau < 3l/d$,

$$\mathrm{E}(e^{\tau a_m(z)}, \tilde{P}) \le (1+\gamma e^{-2})\exp\left[\frac{\tau^2 \mathrm{E}|V_1|^2}{2l(1-\tau d/3l)}\right]. \tag{13}$$

Step 3: By Markov's inequality and inequality (13), we have that for any $0 < \tau \le 3l/d$,

$$\begin{aligned}\mathrm{Prob}\left\{\frac{1}{m}\sum_{i=1}^{m} V_i \ge \varepsilon\right\} &= \mathrm{Prob}\left\{e^{\tau\left[\frac{1}{m}\sum_{i=1}^{m} V_i\right]} \ge e^{\tau\varepsilon}\right\} \\ &\le C_2 \exp\left\{-\tau\varepsilon + \frac{\tau^2 \mathrm{E}|V_1|^2}{2l(1-\tau d/3l)}\right\},\end{aligned}$$

where $C_2 = 1 + \gamma e^{-2}$. Substituting $\tau = \frac{l\varepsilon}{(\mathrm{E}|V_1|^2 + \varepsilon d/3)}$ and noting that the selected value for τ satisfies $\tau \le 3l/d$, then we have that for any $\varepsilon > 0$

$$\mathrm{Prob}\left\{\frac{1}{m}\sum_{i=1}^{m} V_i \ge \varepsilon\right\} \le C_2 \exp\left\{\frac{-l\varepsilon^2}{2(\mathrm{E}|V_1|^2 + \varepsilon d/3)}\right\}.$$

By the inequality above and replacing l by $m^{(\beta)}$, we complete the proof of Lemma 7.

Lemma 8. *Let $\{z_i\}_{i=1}^m$ be a u.e.M.c. sample and $\mathcal{G}$ be a set of functions on $\mathcal{Z}$. Suppose that there is some $a > 0$ such that $\mathrm{E}(g^2) \le a\mathrm{E}(g)$ for any $g \in \mathcal{G}$ and $|g - \mathrm{E}(g)| \le A$ almost everywhere for any $g \in \mathcal{G}$. Then for any $\varepsilon > 0$,*

$$\begin{aligned}&\mathrm{Prob}\left\{\sup_{g\in\mathcal{G}} \frac{\mathrm{E}(g) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)}{\sqrt{\mathrm{E}(g)+\varepsilon}} \ge \sqrt{\varepsilon}\right\} \\ &\le C_2 \mathcal{N}\left(\mathcal{G}, \frac{\varepsilon}{4}\right)\exp\left\{\frac{-\varepsilon m^{(\beta)}}{32(a+A/3)}\right\}.\end{aligned}$$

Proof. Let $\mu = \mathrm{E}(g)$ and $\sigma^2 = \mathrm{E}[(g-\mu)^2]$. For any $\varepsilon > 0$ and any $1 \ge \alpha_1 > 0$, by Lemma 7, we have

$$\begin{aligned}&\mathrm{Prob}\left\{\frac{\mu - \frac{1}{m}\sum_{i=1}^{m} g(z_i)}{\sqrt{\mu+\varepsilon}} \ge \alpha_1\sqrt{\varepsilon}\right\} \\ &\le C_2 \exp\left\{\frac{-\alpha_1^2\varepsilon(\mu+\varepsilon)m^{(\beta)}}{2[\sigma^2 + (A\alpha_1\sqrt{\varepsilon}\sqrt{\mu+\varepsilon})/3]}\right\}.\end{aligned} \tag{14}$$

By assumption, we have $\sigma^2 \le \mathrm{E}[g^2] \le c\mathrm{E}[g] = c\mu$, and

$$\sigma^2 + \frac{A\alpha_1\sqrt{\varepsilon}\sqrt{\mu+\varepsilon}}{3} \le c\mu + \frac{A(\mu+\varepsilon)}{3} \le (\mu+\varepsilon)\left(c + \frac{A}{3}\right).$$

By (14), we have that for any $\varepsilon > 0$, any $1 \ge \alpha_1 > 0$

$$\begin{aligned}&\mathrm{Prob}\left\{\frac{\mu - \frac{1}{m}\sum_{i=1}^{m} g(z_i)}{\sqrt{\mu+\varepsilon}} \ge \alpha_1\sqrt{\varepsilon}\right\} \\ &\le (1+\gamma e^{-2})\exp\left\{\frac{-\alpha_1^2\varepsilon m^{(\beta)}}{2(c+A/3)}\right\}.\end{aligned} \tag{15}$$

Let $\{g_j\}_{j=1}^{n_1} \subset \mathcal{G}$ with $n_1 = \mathcal{N}(\mathcal{G}, \alpha_1\varepsilon)$ such that $\mathcal{G}$ is covered by balls $D_j = \{g \in \mathcal{G} : \|g - g_j\|_\infty \le \alpha_1\varepsilon\}$ centered at g_j with radius $\alpha_1\varepsilon$. Then for any $1 \le j \le n_1$, by inequality (15), we have

$$\begin{aligned}&\mathrm{Prob}\left\{\frac{\mathrm{E}(g_j) - \frac{1}{m}\sum_{i=1}^{m} g_j(z_i)}{\sqrt{\mathrm{E}(g_j)+\varepsilon}} \ge \alpha_1\sqrt{\varepsilon}\right\} \\ &\le (1+\gamma e^{-2})\exp\left\{\frac{-\alpha_1^2\varepsilon m^{(\beta)}}{2(c+A/3)}\right\}.\end{aligned} \tag{16}$$

For any $g \in \mathcal{G}$, there is some $j \in \{1, \ldots, n_1\}$ such that $\|g - g_j\|_\infty \le \alpha_1\varepsilon$. This implies that $|\mathrm{E}(g) - \mathrm{E}(g_j)| \le \|g - g_j\|_\infty \le \alpha_1\varepsilon$, $|\frac{1}{m}\sum_{i=1}^m g(z_i) - \frac{1}{m}\sum_{i=1}^m g_j(z_i)| \le \|g - g_j\|_\infty \le \alpha_1\varepsilon$. It follows that

$$\frac{\left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \frac{1}{m}\sum_{i=1}^{m} g_j(z_i)\right|}{\sqrt{\mathrm{E}(g)+\varepsilon}} \le \alpha_1\sqrt{\varepsilon},$$

$$\frac{\mathrm{E}(g) - \mathrm{E}(g_j)}{\sqrt{\mathrm{E}(g)+\varepsilon}} \le \alpha_1\sqrt{\varepsilon}.$$

The second inequality above implies that $\sqrt{\mathrm{E}(g_j)+\varepsilon} < 2\sqrt{\mathrm{E}(g)+\varepsilon}$ (Chen et al., 2004). Thus we have

$$\mathrm{Prob}\left\{\sup_{g\in\mathcal{G}} \frac{\mathrm{E}(g) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)}{\sqrt{\mathrm{E}(g)+\varepsilon}} \ge 4\alpha_1\sqrt{\varepsilon}\right\}$$

$$\le \sum_{j=1}^{n_1} \text{Prob}\left\{\frac{\mathcal{E}(g_j)-\mathcal{E}_m(g_j)}{\sqrt{\mathcal{E}(g_j)+\varepsilon}} \ge \alpha_1\sqrt{\varepsilon}\right\}$$
$$\le (1+\gamma e^{-2})\mathcal{N}(\mathcal{G},\alpha_1\varepsilon)\exp\left\{\frac{-\alpha_1^2\varepsilon m^{(\beta)}}{2(c+A/3)}\right\}.$$

Taking $\alpha_1 = \frac{1}{4}$ in the above inequality, we finish the proof of Lemma 8.

Lemma 9 (*Cucker & Smale, 2002*). *Let $c_1,\ c_2 > 0$, and $p_1 > p_2 > 0$. Then the equation $x^{p_1} - c_1x^{p_2} - c_2 = 0$ has a unique positive zero x^*. In addition $x^* \le \max\{(2c_1)^{1/(p_1-p_2)}, (2c_2)^{(1/p_1)}\}$.*

To prove Theorem 2, we use the following lemma for strongly mixing (Modha & Masry, 1996).

Lemma 10. *Let $\{z_i\}_{i\ge1}$ be a stationary strongly mixing sequence with the mixing coefficient satisfying Assumption 1. Let an integer $m \ge 1$ be given. For each integer $i \ge 1$, let $U_i = f(z_i)$, where f is some real-valued Borel measurable function. Assume that $|U_1| \le d_2$ a.s. and that $\mathrm{E}[U_1] = 0$. Set $m^{(\alpha)} = \lfloor m\lceil\{8m/c\}^{1/(\beta+1)}\rceil^{-1}\rfloor$. Then for all $\varepsilon > 0$*

$$\text{Prob}\left\{\frac{1}{m}\sum_{i=1}^m U_i \ge \varepsilon\right\} \le (1+4e^{-2}\overline{\alpha})\exp\left\{\frac{-\varepsilon^2 m^{(\alpha)}}{2(\mathrm{E}|U_1|^2+\varepsilon d_2/3)}\right\}.$$

By Lemma 10 and using the similar argument conducted as that in Lemma 8, we establish the relative uniform convergence bound for strongly mixing sequence.

Lemma 11. *Let $\{z_i\}_{i=1}^m$ be strongly mixing and $\mathcal{G}$ be a set of functions on $\mathcal{Z}$. Suppose that there is some $c' \ge 0$ such that $\mathrm{E}(g^2) \le c'\mathrm{E}(g)$ for any $g \in \mathcal{G}$ and $|g - \mathrm{E}(g)| \le A'$ almost everywhere for any $g \in \mathcal{G}$. Then for any $\varepsilon > 0$,*

$$\text{Prob}\left\{\sup_{g\in\mathcal{G}}\frac{\mathrm{E}(g)-\frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{\mathrm{E}(g)+\varepsilon}} \ge \sqrt{\varepsilon}\right\}$$
$$\le (1+4e^{-2}\overline{\alpha})\mathcal{N}\left(\mathcal{G},\frac{\varepsilon}{4}\right)\exp\left\{\frac{-\varepsilon m^{(\alpha)}}{32(c'+A'/3)}\right\}.$$

Appendix B. Proofs of main results

Proof of Proposition 1. We decompose this proof into two steps.

Step 1: Estimate the second term of sample error (6): $\frac{1}{m}\sum_{i=1}^m \zeta_2(z_i) - \mathrm{E}\zeta_2$. By the definition of $D(\lambda)$, we have $\lambda\|f_\lambda\|_K^2 \le \mathcal{E}(f_\lambda) - \mathcal{E}(f_\psi) + \lambda\|f_\lambda\|_K^2 = D(\lambda)$. It follows that $\|f_\lambda\|_K \le \sqrt{D(\lambda)/\lambda}$. By $|f_\psi(x)| \le B$ for any $x \in X$, we have $|\zeta_2| = |(1-yf_\lambda(x))_+ - (1-yf_\psi(x))_+| \le b := \kappa\sqrt{D(\lambda)/\lambda} + B$. It follows that $|\zeta_2(z) - \mathrm{E}\zeta_2| \le 2b$, $\mathrm{E}(\zeta_2^2) \le b\cdot|\mathcal{E}(f_\lambda) - \mathcal{E}(f_\psi)| \le D(\lambda)b$.

Set $V_i = \zeta_2(z_i) - \mathrm{E}\zeta_2$, $m \ge i \ge 1$. By Lemma 7, we have that for any $\varepsilon > 0$

$$\text{Prob}\left\{\frac{1}{m}\sum_{i=1}^m \zeta_2(z_i) - \mathrm{E}\zeta_2 \ge \varepsilon\right\}$$
$$\le (1+\gamma e^{-2})\exp\left\{\frac{-\varepsilon^2 m^{(\beta)}}{2b(D(\lambda)+2\varepsilon/3)}\right\}.$$

Then we have that for any $0 < \delta < 1$, inequality

$$\frac{1}{m}\sum_{i=1}^m \zeta_2(z_i) - \mathrm{E}\zeta_2 \le \varepsilon_1(m,\delta)$$

is valid with probability at least $1-\delta$, where

$$\varepsilon_1(m,\delta) = \frac{2b\ln(C_2/\delta)}{3m^{(\beta)}} + \sqrt{\left(\frac{2b\ln(C_2/\delta)}{3m^{(\beta)}}\right)^2 + \frac{2b\ln(C_2/\delta)D(\lambda)}{m^{(\beta)}}},$$

and $C_2 = 1+\gamma e^{-2}$. Notice that

$$\varepsilon_1(m,\delta) \le \frac{7b\ln(C_2/\delta)}{3m^{(\beta)}} + D(\lambda),$$

and replacing b by $\kappa\sqrt{D(\lambda)/\lambda}+B$, we have that for any $0 < \delta < 1$, the following inequality holds true with probability at least $1-\delta$

$$\frac{1}{m}\sum_{i=1}^m \zeta_2(z_i) - \mathrm{E}\zeta_2 \le \frac{7(\kappa\sqrt{D(\lambda)/\lambda}+B)\ln(C_2/\delta)}{3m^{(\beta)}} + D(\lambda). \quad (17)$$

Step 2: Estimate the first term of sample error (6): $\mathrm{E}\zeta_1 - \frac{1}{m}\sum_{i=1}^m \zeta_1(z_i)$. Let $\mathcal{F}_R = \{(1-yf(x))_+ - (1-yf_\psi(x))_+, f \in B^\sigma_{\mathcal{H}}(R)\}$, $R > 0$, and $g = (1-yf(x))_+ - (1-yf_\psi(x))_+$. We have

$$\mathrm{E}(g) = \mathcal{E}(f) - \mathcal{E}(f_\psi) \ge 0, \qquad \frac{1}{m}\sum_{i=1}^m g(z_i) = \mathcal{E}_m(f) - \mathcal{E}_m(f_\psi).$$

For any $f \in B^\sigma_{\mathcal{H}}(R)$, we have $\|f\|_\infty \le \kappa\|f\|_K \le \kappa R$. It follows that $|g(z)| \le \kappa R + B := b_1$, $|g(z) - \mathrm{E}(g)| \le 2b_1$. Then we have $\mathrm{E}(g^2) \le (\kappa R + B)(\mathcal{E}(f) - \mathcal{E}(f_\psi)) = b_1\mathrm{E}(g)$.

Denote $\mathcal{E}'(f) = \mathcal{E}(f) - \mathcal{E}_m(f)$. By Lemma 8, we have that for any $\varepsilon > 0$

$$\text{Prob}\left\{\sup_{f\in B^\sigma_{\mathcal{H}}(R)}\frac{\mathcal{E}'(f)-\mathcal{E}'(f_\psi)}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\psi)+\varepsilon}} \ge \sqrt{\varepsilon}\right\}$$
$$= \text{Prob}\left\{\sup_{g\in\mathcal{F}_R}\frac{\mathrm{E}(g)-\frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{\mathrm{E}(g)+\varepsilon}} \ge \sqrt{\varepsilon}\right\}$$
$$\le C_2\mathcal{N}\left(\mathcal{F}_R,\frac{\varepsilon}{4}\right)\exp\left\{\frac{-3\varepsilon m^{(\beta)}}{160(\kappa R+B)}\right\}.$$

Since for any $g_1, g_2 \in \mathcal{F}_R$, $|g_1(x) - g_2(x)| \le \|f_1 - f_2\|_\infty$, we have that for any $\varepsilon > 0$, an $\varepsilon/(4R)$-covering of $B^\sigma_{\mathcal{H}}(1)$ provides an $\varepsilon/4$-covering of $\mathcal{F}_R$ (Wu et al., 2006). Then

$$\text{Prob}\left\{\sup_{f\in B^\sigma_{\mathcal{H}}(R)}\frac{\mathcal{E}'(f)-\mathcal{E}'(f_\psi)}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\psi)+\varepsilon}} \ge \sqrt{\varepsilon}\right\}$$
$$\le C_2\mathcal{N}\left(B^\sigma_{\mathcal{H}}(1),\frac{\varepsilon}{4R}\right)\exp\left\{\frac{-3\varepsilon m^{(\beta)}}{160(\kappa R+B)}\right\}.$$

It follows that for $f_{S,\lambda}$ that minimizes the regularized empirical error (3) over $B^\sigma_{\mathcal{H}}(R)$,

$$\text{Prob}\left\{\frac{\mathcal{E}'(f_{S,\lambda})-\mathcal{E}'(f_\psi)}{\sqrt{\mathcal{E}(f_{S,\lambda})-\mathcal{E}(f_\psi)+\varepsilon}} \ge \sqrt{\varepsilon}\right\}$$
$$\le C_2\mathcal{N}\left(B^\sigma_{\mathcal{H}}(1),\frac{\varepsilon}{4R}\right)\exp\left\{\frac{-3\varepsilon m^{(\beta)}}{160(\kappa R+B)}\right\}. \quad (18)$$

Set the right-hand side of inequality (18) to the same value δ above and by Lemma 3, we have

$$C_2\exp\left\{C_{p,d}\sigma^{(1-p/4)d}\left(\frac{4R}{\varepsilon}\right)^p - \frac{3\varepsilon m^{(\beta)}}{160(\kappa R+B)}\right\} = \delta.$$

It follows that

$$\varepsilon^{1+p} - \frac{160(\kappa R + B)\ln(C_2/\delta)}{3m^{(\beta)}} \cdot \varepsilon^p - \frac{160C_{p,d}\sigma^{(1-p/4)d}(\kappa R + B)(4R)^p}{3m^{(\beta)}} = 0.$$

By Lemma 9, we can solve this equation with respect to $\varepsilon := \varepsilon_2(m, \delta)$, and then we have $\varepsilon_2(m, \delta) \le 4R \cdot \max\{m_1, m_2\}$,

$$m_1 = \frac{80(\kappa + 1)\ln(C_2/\delta)}{3m^{(\beta)}},$$

$$m_2 = \left[\frac{80C_{p,d}\sigma^{(1-p/4)d}(\kappa + 1)}{3m^{(\beta)}}\right]^{\frac{1}{1+p}}.$$

Using the fact that $\sqrt{\varepsilon}\sqrt{\mathcal{E}(f) + \varepsilon} \le \frac{1}{2}\mathcal{E}(f) + \varepsilon$, by inequality (18) we have that with probability at least $1-\delta$, the following inequality is valid

$$\mathrm{E}\zeta_1 - \frac{1}{m}\sum_{i=1}^{m}\zeta_1(z_i) \le \frac{1}{2}[\mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_\psi)] + \varepsilon_2(m, \delta).$$

By Lemma 1 and inequality (17) and the above inequality, we complete the proof of Proposition 1.

Proof of Theorem 1. Denote $\Delta_S = \mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_\psi) + 2\lambda\|f_{S,\lambda}\|_K^2$. By the definition of $f_{S,\lambda}$, we have

$$\lambda\|f_{S,\lambda}\|_K^2 \le \mathcal{E}_m(f_{S,\lambda}) + \lambda\|f_{S,\lambda}\|_K^2 \le \mathcal{E}_m(0) = \frac{1}{m}\sum_{i=1}^{m}(1 - y_i 0)_+ \le 1.$$

It follows that $\|f_{S,\lambda}\|_K \le 1/\sqrt{\lambda}$ for almost all S. This implies that for any S, $f_{S,\lambda} \in B^{\sigma}_{\mathcal{H}}(R)$ with $R = 1/\sqrt{\lambda}$. In addition, by the facts that $\lceil t\rceil \le 2t$ for all $t \ge 1$, $\lfloor t\rfloor \ge t/2$ for all $t \ge 2$ (Steinwart et al., 2009), we have $m^{(\beta)} \ge m^{\frac{1}{2}}[\ln(1/\rho)]^{\frac{1}{2}}/(8\sqrt{2})$ as m satisfies $m \ge \max\{\ln(1/\rho)/8, 128/(\ln 1/\rho)\}$.

By Proposition 1, we have that the inequality

$$\Delta_S \le 4D(\lambda) + \frac{7(\kappa\sqrt{D(\lambda)/\lambda} + B)\ln(C_2/\delta)}{3m^{(\beta)}} + 8R\left(\frac{640\sqrt{2}C_{p,d}\sigma^{(1-p/4)d}}{3(\ln(1/\rho))^{\frac{1}{2}}m^{\frac{1}{2}}}\right)^{\frac{1}{1+p}} \tag{19}$$

holds true provided that m satisfies

$$m \ge \max\left\{\frac{\ln(1/\rho)}{8}, \frac{128}{(\ln 1/\rho)}, \frac{5\cdot 2^{11}(\kappa+1)^2(\ln(C_1/\delta))^2}{9\ln(1/\rho)} \cdot \left[\frac{\ln(C_1/\delta)}{C_{p,d}\sigma^{(1-p/4)d}}\right]^{\frac{2}{p}}\right\}.$$

Denote $\mathcal{W}(R) = \{S \in \mathcal{Z}^m : \|f_{S,\lambda}\|_K \le R\}$. Choosing $\lambda = \left(\frac{1}{m}\right)^{\vartheta}$, by inequality (19) and Lemma 2, we have that there is a set $V_R \subseteq \mathcal{Z}^m$ with measure at most δ such that for any $S \in \mathcal{W}(R) \setminus V_R$,

$$\Delta_S \le \lambda^{\frac{\alpha}{\alpha+1}}\left\{C_3 R\lambda^{\frac{1}{2(\alpha+1)}} + 2C_3\right\}, \tag{20}$$

where $C_3 > 1$ is a constant independent of m. Start with $R = R^{(0)} = 1/\sqrt{\lambda}$, by (20), we have $\mathcal{Z}^m = \mathcal{W}(R^{(0)}) \subseteq \mathcal{W}(R^{(1)}) \cup V_{R^{(0)}}$, where

$$R^{(1)} \le \lambda^{\frac{-1}{2(\alpha+1)}}(C_3\lambda^{\frac{-\alpha}{4(\alpha+1)}} + 2C_3).$$

By (20), for $j = 2, 3, \ldots$, we iteratively derive (Tong et al., 2009)

$$\mathcal{Z}^m = \mathcal{W}(R^{(0)}) \subseteq \mathcal{W}(R^{(1)}) \cup V_{R^{(0)}} \subseteq \cdots \subseteq \mathcal{W}(R^{(j)}) \cup \left(\bigcup_{k=0}^{j-1} V_{R^{(k)}}\right),$$

each $V_{R^{(k)}}$ has measure at most δ and $R^{(j)}$ is given by

$$R^{(j)} \le \lambda^{\frac{-1}{2(\alpha+1)}}(C_3\lambda^{\frac{-\alpha}{2(\alpha+1)}}\left(\tfrac{1}{2}\right)^j + 2jC_3). \tag{21}$$

For $\epsilon > 0$, choose $J \in \mathbb{N}$ such that

$$\left(\frac{1}{2}\right)^J \le \frac{2 + (1+p)(2\alpha+1)d}{\alpha d}\cdot\epsilon.$$

Replacing j by J in (21), we have that for $S \in \mathcal{W}(R^{(J)})$

$$\|f_{S,\lambda}\|_K \le \lambda^{\frac{-1}{2(\alpha+1)}}(C_3\lambda^{\frac{-\alpha}{2(\alpha+1)}}\left(\tfrac{1}{2}\right)^J + 2JC_3).$$

This together with (20) gives

$$\mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_\psi) \le \Delta_S \le \widehat{C}\left(\frac{1}{m}\right)^{\theta}, \quad \forall S \in \mathcal{W}(R^{(J)}) \setminus V_{R^J}.$$

Since $\bigcup_{k=0}^{j-1} V_{R^{(k)}}$ has measure at most $J\delta$, replacing δ by δ/J, the measure of $\mathcal{W}(R^{(J)}) \setminus V_{R^{(J)}}$ is at least $1-\delta$. We complete the proof of Theorem 1.

Proof of Theorem 2. Rosenblatt (1972) proved that if a stationary Markov chain satisfies both uniform ergodicity and mixing (in the ergodic-theoretic sense), then it is strongly mixing. Thus by Lemmas 10 and 11, and using the similar argument conducted as that in Proposition 1, we have that for any $0 < \eta < 1$ and all $S \in \mathcal{W}(R) = \{S \in \mathcal{Z}^m : \|f_{S,\lambda}\|_K \le R\}$, inequality

$$\mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_\psi) + 2\lambda\|f_{S,\lambda}\|_K^2 \le 4D(\lambda) + 8R\varepsilon(m,\eta) + \frac{7(\kappa\sqrt{D(\lambda)/\lambda} + B)\ln(C_4/\eta)}{3m^{(\alpha)}} \tag{22}$$

holds true with probability at least $1 - 2\eta$, where $C_4 = 1 + 4e^{-2}\overline{\alpha}$, and $\varepsilon(m,\eta) \le \max\{\widehat{m}, \underline{m}\}$,

$$\widehat{m} = \frac{80(\kappa+1)\ln(C_4/\eta)}{3m^{(\alpha)}},$$

$$\underline{m} = \left[\frac{80C_{p,d}\sigma^{(1-p/4)d}(\kappa+1)}{3m^{(\alpha)}}\right]^{\frac{1}{1+p}}.$$

In addition, by the facts that $\lceil t\rceil \le 2t$ for all $t \ge 1$, $\lfloor t\rfloor \ge t/2$ for all $t \ge 2$ (Steinwart et al., 2009), we have $m^{(\alpha)} \ge 2^{-\frac{2\beta+5}{\beta+1}}c^{\frac{1}{\beta+1}}m^{\frac{\beta}{\beta+1}}$. For $m \ge \max\{c/8, 2^{2+5/\beta}c^{-\beta}\}$. By the similar argument conducted as that in the proof of Theorem 1, we can finish the proof of Theorem 2.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*, 337–404.

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

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

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

Curnow, R. N. (1988). The use of Markov chain models in studying the evolution of the proteins. *Journal of Theoretical Biology*, *134*, 51–57.

Davydov, Y. A. (1973). Mixing conditions for Markov chains. *Theory of Probability and its Applications*, *XVIII*, 312–328.

De Vito, E., Caponnetto, A., & Rosasco, L. (2005). Model selection for regularized least- squares algorithm in learning theory. *Foundations of Computational Mathematics*, *5*, 59–85.
Laarhouen, P. M., & Aarts, E. L. (1987). *Simulated annealing: theory and application*. Norwell, MA: Kluwer Acdemic Publishers.
Meyn, S. P., & Tweedie, R. L. (1993). *Markov chains and stochastic stability*. New York: Springer-Verlag.
Modha, S., & Masry, E. (1996). Minimum complexity regression estimation with weakly dependent observations. *IEEE Transaction on Information Theory*, *42*, 2133–2145.
Mohri, M., & Rostamizadeh, A. (2010). Stability bounds for stationary phi-mixing and beta-mixing processes. *Journal of Machine Learning Research*, *11*, 798–814.
Qian, M. P., & Gong, G. L. (1998). *Applied random processes*. Beijing: Peking University Publisher.
Rosenblatt, M. (1972). Uniform ergodicity and strong mixing. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *24*(1), 79–84.
Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the fifteenth international conference on machine learning* (pp. 515–521). Morgan Kaufmann Publishers Inc.
Smale, S., & Zhou, D. X. (2005). Shannon sampling II. Connections to learning theory. *Applied and Computational Harmonic Analysis*, *19*, 285–302.
Smale, S., & Zhou, D. X. (2009). Online learning with Markov sampling. *Analysis and Applications*, *7*, 87–113.
Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, *2*, 67–93.
Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer.
Steinwart, T., Hush, D., & Scovel, C. (2009). Learning from dependent observations. *Journal of Multivariate Analysis*, *100*, 175–194.
Steinwart, I., & Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, *35*, 575–607.
Tong, H. Z., Chen, D. R., & Peng, L. Z. (2009). Analysis of support vector machine regression. *Foundations of Computational Mathematics*, *9*, 243–257.
van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and emprical processes*. New York: Springer-Verlag.
Vapnik, V. (1998). *Statistical learning theory*. New York: John Wiley.
Vidyasagar, M. (2003). *Learning and generalization with applications to neural networks*. London: Springer.
Withers, C. S. (1981). Conditions for linear processes of sationary mixing sequences. *The Annals of Probability*, *22*, 94–116.
Wu, Q., Ying, Y. M., & Zhou, D. X. (2006). Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, *6*, 171–192.
Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, *22*, 94–116.
Zhang, T. (2004). Statistical behaviour and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, *32*, 56–134.
Zhang, C., & Tao, D. C. (2012). Generalization bounds of ERM-based learning processes for continuous-time Markov chains. *IEEE Transactions on Neural Network and Learning System*, *23*, 1872–1883.
Zou, B., Li, L. Q., & Xu, Z. B. (2009). The generalization performance of ERM algorithm with strongly mixing observations. *Machine Learning*, *75*(3), 275–295.
Zou, B., Li, L. Q., Xu, Z. B., Luo, T., & Tang, Y. Y. (2013). Generalization performance of Fisher linear discriminant based on Markov sampling. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(2), 288–300.
Zou, B., Peng, Z. M., & Xu, Z. B. (2013). The learning performance of support vector machine classification based on Markov sampling. *Science China: Information Sciences*, *56*, 032110(1)–032110(16).