The Generalization Ability of SVM Classification Based on Markov Sampling

Jie Xu, Yuan Yan Tang, *Fellow, IEEE*, Bin Zou, Zongben Xu, Luoqing Li, Yang Lu, and Baochang Zhang

***Abstract*—The previously known works studying the generalization ability of support vector machine classification (SVMC) algorithm are usually based on the assumption of independent and identically distributed samples. In this paper, we go far beyond this classical framework by studying the generalization ability of SVMC based on uniformly ergodic Markov chain (u.e.M.c.) samples. We analyze the excess misclassification error of SVMC based on u.e.M.c. samples, and obtain the optimal learning rate of SVMC for u.e.M.c. samples. We also introduce a new Markov sampling algorithm for SVMC to generate u.e.M.c. samples from given dataset, and present the numerical studies on the learning performance of SVMC based on Markov sampling for benchmark datasets. The numerical studies show that the SVMC based on Markov sampling not only has better generalization ability as the number of training samples are bigger, but also the classifiers based on Markov sampling are sparsity when the size of dataset is bigger with regard to the input dimension.**

***Index Terms*—Generalization ability, learning rate, Markov sampling, support vector machine classification (SVMC).**

I. Introduction

SUPPORT vector machine (SVM) is one of the most widely used machine learning algorithms for classification problems [1]–[3], in particular for classifying high-dimensional data. Besides their good performance in practical applications, they also enjoy a good theoretical justification in terms of both universal consistency [4]–[6] and learning rates [6]–[9], if the training samples come from an independent and identically distributed (i.i.d.) process. However, independence is a very restrictive concept [10]. First, it is often an assumption, rather than a deduction on the basis of observations. Second, it is an all or nothing property, in the sense that two random variables are either independent or they are not—the definition does not permit an intermediate notion of being nearly independent. As a result, many of the proofs based on the assumption that the underlying stochastic sequence is i.i.d. are rather "fragile." In addition, this i.i.d. assumption can not be strictly justified in real-world problems, and many machine learning applications such as market prediction, system diagnosis, and speech recognition are inherently temporal in nature, and consequently not i.i.d. processes [10], [11]. 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 [12] considered the convergence rates of empirical processes for stationary mixing sequences. Vidyasagar [11] studied the notions of mixing and proved that most of the desirable properties (e.g., PAC or UCEMUP) of i.i.d. sequence are preserved when the underlying sequence is mixing sequence. Mohri and Rostamizadeh [13] studied the generalization bounds of stable learning algorithms for non-i.i.d. processes. Smale and Zhou [14] established the learning rates of online learning algorithm with Markov chain samples. Steinwart *et al.* [10] 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 (e.g., WLLNE, SLLNE). Zou *et al.* [15] studied the generalization bounds of empirical risk minimization (ERM) algorithm with strongly mixing observations. Zhang and Tao [16] studied the generalization bounds of ERM-based learning processes for continuous time Markov chains.

There are many dependent sampling mechanisms (e.g., α-mixing, β-mixing and ϕ-mixing) studied in machine learning literatures [10]–[13]. 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, marking prediction, and so on. Second, many empirical evidences [17] show that learning algorithms very often perform well with Markov chain samples (e.g., biological sequence analysis, speech recognition). Why it is so, however, has been unknown (particularly, it is unknown how well it performs in terms of learning rate

Manuscript received November 28, 2013; revised March 10, 2014, July 12, 2014, and July 31, 2014; accepted August 5, 2014. This work was supported in part by the NSFC under Project 11131006, Project 11371007, and Project 61370002, in part by the National Basic Research Program of China under Grant 2013CB329404, in part by the NSF of Hubei Province under Grant 2011CDA410003, in part by the Multiyear Research of the University of Macau under Grant MYRG205(Y1-L4)-FST11-TYY and Grant MYRG187(Y1-L3)-FST11-TYY, and in part by the Start-up Research of the University of Macau under Grant SRG010-FST11-TYY. This paper was recommended by Associate Editor G.-B. Huang. *(Corresponding Author: Bin Zou.)*

J. Xu is with the Faculty of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China (e-mail: jiexu@mail.hust.edu.cn).

B. Zou and L. Li are with the Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China (e-mail: zoubin0502@gmail.com; humcli@gmail.com).

Z. Xu is with the Institute for Information and System Science, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zbxu@mail.xjtu.edu.cn).

Y. Y. Tang and Y. Lu are with the Faculty of Science and Technology, University of Macau 999078, China (e-mail: yytang@umac.mo; lylylytc@gmail.com).

B. Zhang is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China, and also with the Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, 16163, Genova, Italy (e-mail: bczhang@buaa.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2346536

and generalization). Inspired by the idea from [18], in this paper, we first establish two new concentration inequalities for uniformly ergodic Markov chain (u.e.M.c.), and then we establish the optimal learning rate of support vector machine classification (SVMC) for u.e.M.c. samples. In addition, learning from large data is very time-consuming, we usually sample randomly a part of samples from the large dataset and learn from these samples. Then a problem is posed: how to sample a part of samples from the large dataset such that SVMC has better generalization ability? Inspired by the idea from Markov chain Monto Carlo (MCMC) methods [19], we introduce a new Markov sampling algorithm for SVMC to generate u.e.M.c. samples from given dataset. The numerical studies of real-world datasets show that the SVMC based on Markov sampling not only have better learning performance, but also the classifiers are sparsity as the size of data is bigger with regard to the dimension of data.

This paper is organized as follows. In Section II, we introduce some notions and notations used in this paper. In Section III, we present the main results on the learning rates of SVMC with u.e.M.c. samples, and prove our main results. In Section IV, we introduce a new Markov sampling algorithm, and present the numerical studies on the generalization performance of SVMC based on Markov sampling for benchmark datasets. In Section V, we give some useful discussions. Finally, we conclude this paper in Section VI.

II. Preliminaries

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

A. SVMC Algorithm

Let $(\mathcal{X}, d)$ be a compact metric space and $\mathcal{Y} = \{-1, 1\}$. A binary classifier is a function $\hat{f} : \mathcal{X} \to \mathcal{Y}$ which labels every point $x \in \mathcal{X}$ with some $y \in \mathcal{Y}$. Let ψ be a probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and (X, Y) be the corresponding random variable. The misclassification error for a classifier $\hat{f} : \mathcal{X} \to \mathcal{Y}$ is defined to be the probability of the event $\{\hat{f}(X) \neq Y\}$, that is, $\mathcal{R}(\hat{f}) = \Pr\{\hat{f}(X) \neq Y\}$. The SVM classifier [1] is constructed from samples and depends on a reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel [21]. 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 $\{x_i\}_{i=1}^{l} \subset \mathcal{X}$, the matrix $(K(x_i, x_j))_{i,j=1}^{l}$ is positive semidefinite. Such a function is called a Mercer kernel. The RKHS $\mathcal{H}_K$ associated with the kernel K is defined to be the closure of the linear span of the set of functions $\{K_x = K(x, \cdot) : x \in \mathcal{X}\}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K} = \langle \cdot, \cdot \rangle_K$ satisfying $\langle K_{x_i}, K_{x_j} \rangle_K = K(x_i, x_j)$

$$\left\langle \sum_i \alpha_i K_{x_i}, \sum_j \beta_j K_{x_j} \right\rangle_K = \sum_{i,j} \alpha_i \beta_j K\left(x_i, x_j\right).$$

The reproducing property takes the form $\langle K_x, f \rangle_K = f(x)$, $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_K$ [22].

Denote $\mathcal{C}(\mathcal{X})$ as the space of continuous functions on $\mathcal{X}$ with the norm $||f||_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. Let $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x, x)}$, then the above reproducing property tells us that $||f||_\infty \leq \kappa ||f||_K, \ \forall f \in \mathcal{H}_K$. For a function $f : \mathcal{X} \to \mathbb{R}$, the sign function is defined as $sgn(f)(x) = 1$ if $f(x) \geq 0$ and $sgn(f)(x) = -1$ if $f(x) < 0$. Then the SVM classifier associated with the Mercer kernel K is defined as $sgn(f_S)$, where f_S is a minimizer of the following optimization problem involving a set of random sample $S = \{z_i\}_{i=1}^{m} \in \mathcal{Z}^m$:

$$\begin{aligned} f_S = \ & \arg\min_{f \in \mathcal{H}_K} \tfrac{1}{2}||f||_K^2 + \tfrac{C}{m} \textstyle\sum_{i=1}^{m} \xi_i \\ & \text{s.t. } y_i f(x_i) \geq 1 - \xi_i, \xi_i \geq 0, 1 \leq i \leq m \end{aligned} \tag{1}$$

where C is a constant which depends on m: $C = C(m)$ and often $\lim_{m \to \infty} C(m) = \infty$ [6], [8].

We can rewrite algorithm (1) as a regularization scheme as follows [22]. Define the loss function $\ell(f, z)$ as

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

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

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

Here $\lambda = 1/(2C)$ is the regularization parameter [22].

To measure the generalization ability of algorithm (2), we should bound how $\mathrm{sgn}(f_S)$ converges (with respect to the misclassification error) to the best classifier, the Bayes rule, as m and, hence, $C(m)$ tend to infinity. Recall the regression function of ψ, $f_\psi(x) = \int_{\mathcal{Y}} y d\psi(y|x), \ x \in \mathcal{X}$. Then the Bayes rule is given by the sign of the regression function $f_c = sgn(f_\psi)$.

Differ from the previously known works on the generalization ability of SVMC in [5] and [8], in this paper, our aim is to bound the generalization ability of SVMC algorithm (2) for u.e.M.c. sample S.

B. u.e.M.cs

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 $Pr^n(A|z_i)$, $A \in \mathcal{S}, z_i \in \mathcal{Z}$. It is assumed that

$$Pr^n\left(A|z_i\right) := \Pr\left\{Z_{n+i} \in A | Z_j, j < i, Z_i = z_i\right\}.$$

Thus, $Pr^n(A|z_i)$ 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. 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 $Pr^n(A|z_i) = \Pr\{Z_{n+i} \in A | Z_i = z_i\}$. 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. [11].

Definition 1: A Markov chain $\{Z_t\}_{t \geq 1}$ is said to be uniformly ergodic if for some $0 < \gamma_0 < \infty$ and $0 < \rho_0 < 1$

$$||Pr^k(\cdot|z) - \pi(\cdot)||_{TV} \leq \gamma_0 \rho_0^k, \forall k \geq 1, k \in \mathbb{N}$$

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

Meyn and Tweedie [23] proved that the k-step transition probability measures $Pr^k(\cdot|\cdot)$ of u.e.M.c. satisfy the following Doeblin condition [24].

Proposition 1 (Doeblin Condition): Let $\{Z_t\}_{t\geq 1}$ be a Markov chain with transition probability measure $Pr^k(\cdot|\cdot)$, and let μ be some nonnegative measure with nonzero mass μ_0. If there is some integer t such that for all z in $\mathcal{Z}$, and all measurable sets A, $Pr^t(A|z) \leq \mu(A)$, then for any integer k and for any z, z' in $\mathcal{Z}$

$$\left\| Pr^k(\cdot|z) - Pr^k\left(\cdot|z'\right)\right\|_{TV} \leq 2\beta_1^{k/t} \tag{3}$$

where $\beta_1 = 1 - \mu_0$.

III. Estimating Learning Rates

To bound the generalization ability of SVMC algorithm (2), we should estimate the excess misclassification error $\mathcal{R}(sgn(f_S)) - \mathcal{R}(f_c)$. Since the minimization (2) is taken over the discrete quantity $\mathcal{E}_m(f)$, we have to regulate the capacity of the function set. Here the capacity is measured by the covering number [25], [26].

Definition 2: 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 $\mathcal{F}$.

For $R > 0$, let $\mathcal{B}_R = \{f \in \mathcal{H}_K : ||f||_K \leq R\}$. It is a subset of $\mathcal{C}(X)$ and the covering number is well defined [26]. We denote the covering number of $\mathcal{B}_1$ as $\mathcal{N}(\epsilon) = \mathcal{N}(\mathcal{B}_1, \epsilon), \epsilon > 0$.

Definition 3 [8]: The RKHS is said to have polynomial complexity exponent $s > 0$ if there is some $C_s > 0$ such that $\ln \mathcal{N}(\epsilon) \leq C_s(1/\epsilon)^s, \ \forall \epsilon > 0$.

Remark 1: The covering number $\mathcal{N}(\epsilon)$ has been extensively studied, please see [27]–[29]. It was shown in [28] that Definition 3 holds if K is $C^{2n/s}$ on a subset $\mathcal{X}$ of $\mathbb{R}^n$. In particular, for a C^∞ kernel (such as Gaussians), Definition 3 is valid for any $s > 0$ [28].

Zhang [4] established the relation between excess misclassification error and excess generalization error $\mathcal{E}(f) - \mathcal{E}(f_\psi)$ for convex loss, that is

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

For the excess generalization error $\mathcal{E}(f) - \mathcal{E}(f_\psi)$, we have the following error decomposition [8].

Proposition 2: Let f_S be defined as (2) and $f_\lambda = \arg\min_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) + \lambda||f||_K^2\right\}$. Then we have

$$\begin{aligned}\mathcal{E}(f_S) - \mathcal{E}(f_\psi) &\leq \mathcal{E}(f_S) - \mathcal{E}(f_\psi) + \lambda||f_S||_K^2 \\ &\leq \{\mathcal{E}(f_S) - \mathcal{E}_m(f_S) + \mathcal{E}_m(f_\lambda) - \mathcal{E}(f_\lambda)\} + D(\lambda)\end{aligned} \tag{5}$$

where $D(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f_\psi) + \lambda||f_\lambda||_K^2$.

In Proposition 2, we decompose the excess generalization error into two parts. The first term on the right-hand side of (5) is called the sample error, which can be written as

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

where $\zeta_1 = \ell(f_S, z) - \ell(f_\psi, z)$, $\zeta_2 = \ell(f_\lambda, z) - \ell(f_\psi, z)$. The second term $D(\lambda)$ is called the regularization error, which is dependent on the choice of function space. The regularization error has been well understood in learning theory (see [8], [30]).

Definition 4: We say the function f_ψ can be approximated by $\mathcal{H}_K$ with exponent $0 < \beta \leq 1$ if there exists a constant C_β such that for any $\lambda > 0$, $D(\lambda) \leq C_\beta \lambda^\beta$.

In this paper, we assume that there is a constant B such that $|f_\psi| \leq B$ (see [9], [30]).

A. Main Results

Our main results are stated as follows.

Theorem 1: Let $\{z_i\}_{i=1}^m$ be a u.e.M.c. sample, and $R \geq B$. Then for any $0 < \delta < 1$, the inequality

$$\begin{aligned}\mathcal{E}(f_S) - \mathcal{E}(f_\psi) + 2\lambda||f_S||_K^2 \leq 3D(\lambda) + 2R\cdot\varepsilon(m,\delta) \\ + \frac{112\ln(2/\delta)||\Gamma_0||^2\left(\sqrt{D(\lambda)/\lambda} + B\right)}{m}\end{aligned}$$

holds true with probability at least $1 - \delta$, where $\varepsilon(m, \delta) \leq \max\{m_1, m_2\}$, $||\Gamma_0|| = \sqrt{2}/(1 - \beta_1^{1/2t})$

$$m_1 = \frac{112(\kappa+1)\ln(2/\delta)||\Gamma_0||^2}{m}$$

$$m_2 = \left[\frac{112C_s(\kappa+1)||\Gamma_0||^2}{m}\right]^{\frac{1}{1+s}}$$

β_1 and t are defined as that in Proposition 1.

For all $\lambda > 0$, by the definition of f_S, we have

$$\lambda||f_S||_K^2 \leq \mathcal{E}_m(f_S) + \lambda||f_S||_K^2 \leq \mathcal{E}_m(0) + 0 \leq 1.$$

It follows that $||f_S||_K \leq 1/\sqrt{\lambda}$ for almost all $S \in \mathcal{Z}^m$. Thus, by Theorem 1, we have

Corollary 1: Let $\{z_i\}_{i=1}^m$ be a u.e.M.c. sample. Then for any $0 < \delta < 1$ and $0 < \lambda \leq 1/B^2$, the inequality

$$\begin{aligned}\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \leq \frac{2}{\sqrt{\lambda}}\left(\frac{112C_s(\kappa+1)||\Gamma_0||^2}{m}\right)^{\frac{1}{1+s}} \\ + \frac{112\ln(2/\delta)||\Gamma_0||^2\left(\sqrt{D(\lambda)/\lambda} + B\right)}{m} + 3D(\lambda)\end{aligned}$$

is valid with probability at least $1 - \delta$ provided that $m \geq 112(\kappa+1)||\Gamma_0||^2\ln(2/\delta)\left(\ln(2/\delta)/C_s\right)^{1/s}$.

The learning rate in weak form can be obtained from Corollary 1. We improve the error estimate stated in Corollary 1 by using the iteration technique, which was introduced in [31] and improved in [9] and [32]. Our main result can be stated as the following theorem.

Theorem 2: Let $\{z_i\}_{i=1}^m$ be a u.e.M.c. sample. Taking $\lambda = (1/m)^\vartheta$. For any $\epsilon > 0$ and $0 < \delta < 1$, there exists a constant $\widehat{C}$ independent of m such that

$$\mathcal{R}(sgn(f_S)) - \mathcal{R}(f_c) \leq \widehat{C}\left(\frac{1}{m}\right)^\theta$$

holds true with probability at least $1 - \delta$ provided $m \geq 112(\kappa+1)||\Gamma_0||^2\ln(1/\delta)\left(\ln(1/\delta)/C_s\right)^{1/s}$, where

$$\vartheta = \min\left\{\frac{2}{\beta+1}, \frac{2}{(1+\beta)(1+s)}\right\}$$

$$\theta = \min\left\{\frac{2\beta}{\beta+1}, \frac{2\beta}{(1+\beta)(1+s)} - \epsilon\right\}.$$

Remark 2: To have a better understanding of the significance and value of the established result in Theorem 2, we compare Theorem 2 with the previously known results.

1) By Theorem 2, we can find that for $\beta = 1, \theta > (1/2)$ (up to a ϵ). In particular, when $\beta = 1, s \to 0, \theta$ is arbitrarily close to 1. This implies that the learning rate in Theorem 2 is arbitrarily close m^{-1}. Compared the learning rate in Theorem 2 with that of i.i.d. samples in [8], [32], and [33], we can find that this learning rate is same as that obtained in [8], [32], and [33] for i.i.d. setting.
2) For non-i.i.d. setting, Steinwart and Christmann [7] established the fast learning rate [7, eq. (2)] of least squares SVM algorithm for exponentially strongly mixing. The rate in [7] is same as that known in the i.i.d. case (see [8], [32], [33]). Xu and Chen [35] obtained the same learning rate as that in [7] for least square regularized regression with exponentially strongly mixing sequence. These rates are optimal for exponentially strongly mixing sequence under the framework of statistical learning theory. Different from [7] and [35], in this paper, we consider different non-i.i.d. setting, u.e.M.c., and established the same learning rate as that in [7] and [35], which implies that the learning rate obtained in this paper is optimal for Markov chain setting under the framework of statistical learning theory. To our knowledge, the result obtained in Theorem 2 is the first work in this paper.

B. Main Tools

To prove the main results, our main tools are the following four lemmas. Lemmas 1 and 2 are due to Samson [18]. Inspired by the idea from [18], we establish two new concentration inequalities in Lemmas 3 and 4.

To measure the dependence between random variables $Z_1, \ldots, Z_m$, Marton [20] introduced the triangular matrix $\Gamma = (\gamma_i^j)_{1 \le i,j \le m}$. Namely, let $(Z_1, \ldots, Z_m)$ be a sample of real-valued random variables defined on the probability space $\mathcal{Z}$. We assume that the random variables $Z_1, \ldots, Z_m$ are taken out of a sequence $\{Z_t\}_{t \in \mathbb{Z}}$ which is not independent. For $i \ge j$, $\gamma_i^j = 0$ if $i > j$ and $\gamma_i^j = 1$ if $i = j$. For $1 \le i < j \le m$, let Z_i^j represent the random variables $(Z_i, \ldots, Z_j)$, and let $\mathcal{L}(Z_j^n | Z_1^{i-1} = z_1^{i-1}, Z_i = z_i)$ denote the law of Z_j^m conditionally to $Z_1^{i-1} = z_1^{i-1}$ and $Z_i = z_i$. For every $1 \le i < j \le m$ and for $z_1, \ldots, z_i, z_i'$ in $\mathcal{Z}$, let

$$a_j\left(z_1^{i-1}, z_i, z_i'\right) = ||\mathcal{L}\left(Z_j^n | Z_1^{i-1} = z_1^{i-1}, Z_i = z_i'\right) - \mathcal{L}\left(Z_j^n | Z_1^{i-1} = z_1^{i-1}, Z_i = z_i\right)||_{TV}.$$

Then for $1 \le i < j \le m$, γ_i^j is defined as

$$\left(\gamma_i^j\right)^2 = \sup_{z_i, z_i' \in \mathcal{Z}} \sup_{z_1^{i-1} \in \mathcal{Z}^{i-1}} a_j\left(z_1^{i-1}, z_i, z_i'\right).$$

In particular, if $Z_1, \ldots, Z_m$ is a Markov chain, by the Markov property, the coefficient $(\gamma_i^j)^2$ take a simpler form. That is, for $1 \le i < j \le m$

$$\left(\gamma_i^j\right)^2 = \sup_{z_i, z_i' \in \mathcal{Z}} ||\mathcal{L}\left(Z_j | Z_i = z_i\right) - \mathcal{L}\left(Z_j | Z_i = z_i'\right)||_{TV}.$$

Let P denote the law of the sample $(Z_1, \ldots, Z_m)$ on $\mathcal{Z}^m$. For two measures of probability Q and R on $\mathcal{Z}^m$, let $\mathcal{M}(Q, R)$ denote the set of all probability measures on $\mathcal{Z}^m \times \mathcal{Z}^m$ with marginals Q and R. Let $||\Gamma||$ be the usual operator norm of the matrix Γ with respect to the Euclidean topology [18]. For every function g on $\mathcal{Z}^m$, the entropy functional is defined as

$$Ent_P\left(g^2\right) = \int g^2 \log g^2 dP - \int g^2 dP \log \int g^2 dP.$$

By $||\Gamma||$ and the entropy functional, Samson [18] established the bound of $d_2(Q, R)$

$$d_2(Q, R) = \inf_{\Pi \in \mathcal{M}(Q,R)} \left(\int \sum_{i=1}^n Pr^2(Z_i \ne z_i') | Z_i' = z_i')\right)^{\frac{1}{2}}.$$

Lemma 1 (Theorem 1 in [18]): For every probability measure Q on $\mathcal{Z}^m$ with Radon–Nikodym derivative dQ/dP with respect to the measure P

$$d_2(Q, P) = d_2(P, Q) \le ||\Gamma|| \sqrt{2 Ent_P\left(\frac{dQ}{dP}\right)}.$$

By Proposition 1, Samson [18] established the bound of $||\Gamma||$ for u.e.M.c. samples, please see inequality (2.5).

Lemma 2: For u.e.M.c. sample $Z_1, \ldots, Z_m$, we have $||\Gamma|| \le \sqrt{2}/(1 - \beta_1^{1/2t})$, where β_1 and t are defined as that in Proposition 1.

Lemma 3: Let $\mathcal{F}$ be a countable class of bounded measurable functions, and $Z_1, \ldots, Z_m$ be a u.e.M.c. sample. Assume that $0 \le g(z) \le C$ for any $g \in \mathcal{F}$ and for any $z \in \mathcal{Z}$. Then for any $\varepsilon > 0$ ($||\Gamma_0|| = \sqrt{2}/(1 - \beta_1^{1/2t})$)

$$Pr\left\{\left|\frac{1}{m}\sum_{i=1}^m g(z_i) - E(g)\right| \ge \varepsilon\right\} \le 2 \exp\left\{\frac{-m\varepsilon^2}{56C||\Gamma_0||^2 E(g)}\right\}.$$

Proof: Let $\xi = (z_1, \ldots, z_m)$ and $\zeta = (z_1', \ldots, z_m')$. For any $g \in \mathcal{F}$, we define $f(\xi) = \sum_{i=1}^m g(z_i), f(\zeta) = \sum_{i=1}^m g(z_i')$. Then we have

$$\begin{aligned} f(\xi) - f(\zeta) &= \sum_{i=1}^m \left(g(z_i) - g\left(z_i'\right)\right) \\ &\le \sum_{i=1}^m g(z_i) \mathbf{1}_{z_i \ne z_i'} + \sum_{i=1}^m g\left(z_i'\right) \mathbf{1}_{z_i \ne z_i'}. \end{aligned} \tag{7}$$

Let Q be a probability measure on $\mathcal{Z}^m$ with density h with respect to P. For every measure Π on $\mathcal{Z}^m \times \mathcal{Z}^m$ with marginals Q and P, that is, $\Pi \in \mathcal{M}(Q, P)$. Therefore, by inequality (7), we have

$$\begin{aligned} I &:= \int f(\xi) dQ - \int f(\zeta) dP \\ &= \int \int (f(\xi) - f(\zeta)) d\Pi(\xi, \zeta) \\ &\le \int \int \sum_{i=1}^m g(z_i) \mathbf{1}_{z_i \ne z_i'} d\Pi(\xi, \zeta) \\ &\quad + \int \int \sum_{i=1}^m g(z_i') \mathbf{1}_{z_i \ne z_i'} d\Pi(\xi, \zeta). \end{aligned}$$

Integrating with respect to the variable ξ and using the Cauchy–Schwarz inequality, we have

$$\begin{aligned} I \le & \left(\int \sum_{i=1}^{m} g^2(z_i')dQ\right)^{\frac{1}{2}} \\ & \times \left(\int \sum_{i=1}^{m} Pr^2(Z_i \neq z_i'|Z_i' = z_i')dQ\right)^{\frac{1}{2}} \\ & + \left(\int \sum_{i=1}^{m} g^2(z_i)dP\right)^{\frac{1}{2}} \\ & \times \left(\int \sum_{i=1}^{m} Pr^2(Z_i' \neq z_i|Z_i = z_i)dP\right)^{\frac{1}{2}}. \end{aligned} \tag{8}$$

Let $V_1 = \sum_{i=1}^{m} g^2(z_i)$, $V_2 = \sum_{i=1}^{m} g^2(z_i')$. By Lemma 1, we have

$$\left(\int \sum_{i=1}^{m} Pr^2(Z_i \neq z_i'|Z_i' = z_i')dQ\right)^{\frac{1}{2}} \le ||\Gamma||\sqrt{2Ent_P\left(\frac{dQ}{dP}\right)}$$

$$\left(\int \sum_{i=1}^{m} Pr^2(Z_i' \neq z_i|Z_i = z_i)dP\right)^{\frac{1}{2}} \le ||\Gamma||\sqrt{2Ent_P\left(\frac{dQ}{dP}\right)}.$$

By inequality (8), we have

$$\begin{aligned} I \le & \sqrt{2||\Gamma||^2 E_Q(V_1) Ent_P\left(\frac{dQ}{dP}\right)} \\ & + \sqrt{2||\Gamma||^2 E_P(V_2) Ent_P\left(\frac{dQ}{dP}\right)}. \end{aligned} \tag{9}$$

We then use the following variational equality:

$$\begin{aligned} & \sqrt{2||\Gamma||^2 E_Q(V_1) Ent_P\left(\frac{dQ}{dP}\right)} \\ & \quad = \inf_{\tau>0}\left(\frac{\tau||\Gamma||^2 E_Q(V_1)}{2} + \frac{1}{\tau} Ent_P\left(\frac{dQ}{dP}\right)\right). \end{aligned}$$

Thus, we have that for any $\tau > 0$

$$I \le \frac{\tau||\Gamma||^2}{2}\left(E_Q(V_1) + E_P(V_2)\right) + \frac{2}{\tau} Ent_P(h).$$

It follows that:

$$\begin{aligned} & \int \left[\frac{\tau}{2}(f - E_P(f)) - \frac{\tau^2||\Gamma||^2}{4}(V_1 + E_P(V_2))\right] h dP \\ & \le Ent_P(h). \end{aligned} \tag{10}$$

Taking $h = e^l/(E_P(e^l))$, where $l = (\tau/2)(f - E_P(f)) - (\tau^2||\Gamma||^2/4)(V_1 + E_P(V_2))$, we have that for any $\tau > 0$

$$\int \exp\left[\frac{\tau}{2}(f - E_P(f)) - \frac{\tau^2||\Gamma||^2}{4}(V_1 + E_P(V_2))\right] dP \le 1.$$

It follows that:

$$\begin{aligned} & \int \exp\left[\frac{\tau(f - E_P(f))}{2} - \frac{\tau^2||\Gamma||^2 V_1}{4}\right] dP \\ & \qquad \le \exp\left\{\frac{\tau^2||\Gamma||^2 E(V_2)}{4}\right\}. \end{aligned}$$

By Cauchy–Schwarz inequality, we have

$$\begin{aligned} & \int \exp\left[\frac{\tau}{4}(f - E_P(f))\right] dP \\ & \quad \le \exp\left\{\frac{\tau^2||\Gamma||^2 E(V_1)}{8}\right\}\left(\int \exp\left(\frac{\tau^2||\Gamma||^2 V_2}{4}\right) dP\right)^{\frac{1}{2}}. \end{aligned}$$

Since $V_1 \le Cf$, by the above inequality, we have

$$\int \exp\left(\frac{\tau^2||\Gamma||^2}{4} V_2\right) dP \le \exp\left\{\frac{3\tau^2||\Gamma||^2 CE(f)}{4}\right\}.$$

It follows that:

$$\int \exp\left[\frac{\tau}{4}(f - E_P(f))\right] dP \le \exp\left\{\frac{7\tau^2||\Gamma||^2 CE(f)}{8}\right\}.$$

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

$$\begin{aligned} Pr\{f - E(f) \ge \varepsilon\} &= Pr\{e^{\frac{\tau}{4}(f - E(f))} \ge e^{\frac{\tau}{4}\varepsilon}\} \\ &\le \exp\left\{\frac{7\tau^2||\Gamma||^2 CE(f)}{8} - \frac{\tau}{4}\varepsilon\right\}. \end{aligned}$$

Taking $\tau = (\varepsilon/(7C||\Gamma||^2 E(f)))$, we have that for any $\varepsilon > 0$

$$Pr\{f - E(f) \ge \varepsilon\} \le \exp\left\{\frac{-\varepsilon^2}{56C||\Gamma||^2 E(f)}\right\}. \tag{11}$$

By symmetry and replacing ε, $||\Gamma||$ by $m\varepsilon$, $||\Gamma_0||$ in inequality (11), we complete the proof of Lemma 3. ■

Lemma 4: With all notations as that in Lemma 3, then for any $\varepsilon > 0$

$$\begin{aligned} & \Pr\left\{\sup_{g\in\mathcal{F}} \frac{\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \mathrm{E}(g)}{\sqrt{\mathrm{E}(g) + \varepsilon}} \ge 4\sqrt{\varepsilon}\right\} \\ & \qquad \le \mathcal{N}(\mathcal{F}, \varepsilon) \exp\left\{\frac{-\varepsilon m}{56C||\Gamma_0||^2}\right\}. \end{aligned}$$

Proof: By Lemma 3, we have that for any $\varepsilon > 0$

$$\Pr\left\{\frac{\mathrm{E}(g) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)}{\sqrt{\mathrm{E}(g) + \varepsilon}} \ge \sqrt{\varepsilon}\right\} \le \exp\left\{\frac{-\varepsilon m}{56C||\Gamma_0||^2}\right\}.$$

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

$$\Pr\left\{\frac{\mathrm{E}(g_j) - \frac{1}{m}\sum_{i=1}^{m} g_j(z_i)}{\sqrt{\mathrm{E}(g_j) + \varepsilon}} \ge \sqrt{\varepsilon}\right\} \le \exp\left\{\frac{-\varepsilon m}{56C||\Gamma_0||^2}\right\}.$$

For any $g \in \mathcal{F}$, there is some j such that $||g - g_j||_\infty \le \varepsilon$. This implies that $|\mathrm{E}(g) - \mathrm{E}(g_j)| \le ||g - g_j||_\infty \le \varepsilon$, and

$$\left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \frac{1}{m}\sum_{i=1}^{m} g_j(z_i)\right| \le ||g - g_j||_\infty \le \varepsilon.$$

It follows that $\mathrm{E}(g) - \mathrm{E}(g_j)/\sqrt{\mathrm{E}(g) + \varepsilon} \le \sqrt{\varepsilon}$ and:

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

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

$$\begin{aligned}\Pr\left\{\sup_{g\in\mathcal{F}}\frac{\mathrm{E}(g)-\frac{1}{m}\sum_{i=1}^{m}g(z_i)}{\sqrt{\mathrm{E}(g)+\varepsilon}}\geq 4\sqrt{\varepsilon}\right\}&\leq\sum_{j=1}^{n_1}\Pr\left\{\frac{\mathcal{E}(g_j)-\mathcal{E}_m(g_j)}{\sqrt{\mathcal{E}(g_j)+\varepsilon}}\geq\sqrt{\varepsilon}\right\}\\&\leq\mathcal{N}(\mathcal{F},\varepsilon)\exp\left\{\frac{-\varepsilon m}{56C||\Gamma_0||^2}\right\}.\end{aligned}$$

■

Lemma 5 [36]: 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, $x^*\leq\max\{(2c_1)^{1/(p_1-p_2)},(2c_2)^{(1/p_1)}\}$.

C. Proofs of Main Results

Now we give the proofs of our main results as follows.

Proof of Theorem 1: We decompose the proof of Theorem 1 into two steps.

Step 1: Estimate the second term of the sample error: $(1/m)\sum_{i=1}^{m}\zeta_2(z_i)-\mathrm{E}\zeta_2$. By the definition of $D(\lambda)$

$$\lambda||f_\lambda||_K^2\leq\mathcal{E}(f_\lambda)-\mathcal{E}(f_\psi)+\lambda||f_\lambda||_K^2=D(\lambda).$$

It follows that $||f_\lambda||_K\leq\sqrt{D(\lambda)/\lambda}$. By our assumption, $|f_\psi|\leq B$, we have

$$|\zeta_2|:=|(1-yf_\lambda(\mathbf{x}))_+-(1-yf_\psi)_+|\leq d:=\sqrt{D(\lambda)/\lambda}+B.$$

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

$$\Pr\left\{\frac{\frac{1}{m}\sum_{i=1}^{m}\zeta_2(z_i)-\mathrm{E}(\zeta_2)}{\mathrm{E}(\zeta_2)+\varepsilon}\geq\sqrt{\varepsilon}\right\}\leq\exp\left\{\frac{-\varepsilon m}{56||\Gamma_0||^2d}\right\}.$$

Then for any $0<\delta<1$, with probability at least $1-\delta$, the following inequality is valid:

$$\frac{1}{m}\sum_{i=1}^{m}\zeta_2(z_i)-\mathrm{E}(\zeta_2)\leq\frac{1}{2}D(\lambda)+\frac{56\ln(1/\delta)d||\Gamma_0||^2}{m}.$$

Step 2: Estimate the first term of the sample error: $\mathrm{E}\zeta_1-(1/m)\sum_{i=1}^{m}\zeta_1(z_i):=S_1$. For $R>0$, let $\mathcal{F}_R=\{(1-yf(x))_+-(1-yf_\psi(x))_+,\ f\in\mathcal{B}_R\}$ and $g=(1-yf(x))_+-(1-yf_\psi(x))_+$. We have $\mathrm{E}(g)=\mathcal{E}(f)-\mathcal{E}(f_\psi)\geq 0$, $(1/m)\sum_{i=1}^{m}g(z_i)=\mathcal{E}_m(f)-\mathcal{E}_m(f_\psi)$. For any $f\in\mathcal{B}_R$, we have $||f||_\infty\leq\kappa||f||_K\leq\kappa R$. By the assumption $|f_\psi|\leq B$, we have $|g(z)|\leq\kappa R+B:=b$. By Lemma 4, we have that for any $\varepsilon>0$

$$\begin{aligned}&\Pr\left\{\sup_{f\in\mathcal{B}_R}\frac{\mathcal{E}'(f)-\mathcal{E}'(f_\psi)}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\psi)+\varepsilon}}\geq\sqrt{\varepsilon}\right\}\\&\quad=\Pr\left\{\sup_{g\in\mathcal{F}_R}\frac{\mathrm{E}(g)-\frac{1}{m}\sum_{i=1}^{m}g(z_i)}{\sqrt{\mathrm{E}(g)+\varepsilon}}\geq\sqrt{\varepsilon}\right\}\\&\quad\leq\mathcal{N}(\mathcal{F}_R,\varepsilon)\exp\left\{\frac{\varepsilon m}{56b||\Gamma_0||^2}\right\}\end{aligned}\tag{12}$$

where $\mathcal{E}'(f)=\mathcal{E}(f)-\mathcal{E}_m(f)$. Since for any $g_1,g_2\in\mathcal{F}_R$, $|g_1(x)-g_2(x)|\leq||f_1-f_2||_\infty$, by inequality (12), we have that for any $\varepsilon>0$

$$\begin{aligned}&\Pr\left\{\sup_{f\in\mathcal{B}_R}\frac{\mathcal{E}'(f)-\mathcal{E}'(f_\psi)}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\psi)+\varepsilon}}\geq\sqrt{\varepsilon}\right\}\\&\qquad\leq\mathcal{N}\left(\frac{\varepsilon}{R}\right)\exp\left\{\frac{\varepsilon m}{56||\Gamma_0||^2(\kappa R+B)}\right\}.\end{aligned}$$

It follows that for f_S that minimizes the regularized empirical error (2) over $\mathcal{B}_R$:

$$\begin{aligned}&\Pr\left\{\frac{\mathcal{E}'(f_S)-\mathcal{E}'(f_\psi)}{\sqrt{\mathcal{E}(f_S)-\mathcal{E}(f_\psi)+\varepsilon}}\geq\sqrt{\varepsilon}\right\}\\&\qquad\leq\mathcal{N}\left(\frac{\varepsilon}{R}\right)\exp\left\{\frac{\varepsilon m}{56||\Gamma_0||^2(\kappa R+B)}\right\}.\end{aligned}\tag{13}$$

Set the right-hand side of inequality (13) to the same value δ above. By Definition 3, we have

$$\exp\left\{C_s\left(\frac{R}{\varepsilon}\right)^s-\frac{\varepsilon m}{56||\Gamma_0||^2(\kappa R+B)}\right\}=\delta.$$

It follows that:

$$\begin{aligned}\varepsilon^{1+s}-\frac{56(\kappa R+B)\ln(1/\delta)||\Gamma_0||^2}{m}\cdot\varepsilon^s\\-\frac{56C_s(\kappa R+B)R^s||\Gamma_0||^2}{m}=0.\end{aligned}$$

By Lemma 5, we can solve this equation with respect to $\varepsilon:=\varepsilon'(m,\delta)$. The solution is then given by $\varepsilon'(m,\delta)\leq\max\{\widetilde{m},\widehat{m}\}$, where

$$\widetilde{m}=\frac{112(\kappa R+B)\ln(1/\delta)||\Gamma_0||^2}{m}$$

$$\widehat{m}=\left[\frac{112C_s(\kappa R+B)R^s||\Gamma_0||^2}{m}\right]^{\frac{1}{1+s}}.$$

Using the fact that $\sqrt{\varepsilon}\sqrt{\mathcal{E}(f)+\varepsilon}\leq(1/2)\mathcal{E}(f)+\varepsilon$ and inequality (13), we have that with probability at least $1-\delta$, the following inequality is valid:

$$\begin{aligned}S_1&=\mathcal{E}(f_S)-\mathcal{E}(f_\psi)-\left[\mathcal{E}_m(f_S)-\mathcal{E}_m(f_\psi)\right]\\&\leq\frac{1}{2}\left[\mathcal{E}(f_S)-\mathcal{E}(f_\psi)\right]+\varepsilon'(m,\delta).\end{aligned}$$

Thus, we have that for any $0<\delta<1$, with probability at least $1-2\delta$ the inequality

$$\begin{aligned}\mathcal{E}(f_S)-\mathcal{E}(f_\psi)+2\lambda||f_S||_K^2\leq 3D(\lambda)+2\varepsilon'(m,\delta)\\+\frac{112\ln(1/\delta)||\Gamma_0||^2\left(\sqrt{D(\lambda)/\lambda}+B\right)}{m}\end{aligned}\tag{14}$$

is valid. Replacing δ by $\delta/2$ in inequality (14), we complete the proof of Theorem 1. ■

Proof of Theorem 2: By Theorem 1, we have that for any $0<\delta<1$, the inequality

$$\begin{aligned}\Delta:&=\mathcal{E}(f_S)-\mathcal{E}(f_\psi)+2\lambda||f_S||_K^2\\&\leq\frac{112\ln(2/\delta)||\Gamma_0||^2\left(\sqrt{D(\lambda)/\lambda}+B\right)}{m}\\&\quad+3D(\lambda)+2R\cdot\left[\frac{112C_s(\kappa+1)||\Gamma_0||^2}{m}\right]^{\frac{1}{1+s}}\end{aligned}\tag{15}$$

is valid with probability at least $1-\delta$ provided that $m \geq 112(\kappa+1)||\Gamma_0||^2 \ln(2/\delta)\,(\ln(2/\delta)/C_s)^{1/s}$.

Denote $\mathcal{W}(R) = \{S \in \mathcal{Z}^m : ||f_S||_K \leq R\}$. Taking $\lambda = (1/m)^{\vartheta}$, we easily check that

$$\frac{1}{m}\lambda^{\frac{\beta-1}{2}} \leq \lambda^{\beta},\ \lambda^{\frac{1}{\vartheta(s+1)}} \leq \lambda^{\frac{\beta+1}{2}}.$$

By inequality (15), we have that there is a set $V_R \subseteq \mathcal{Z}^m$ with measure at most δ such that for any $S \in \mathcal{W}(R)\backslash V_R$

$$\Delta \leq \lambda^{\beta}\left\{2C_1 + C_1 R\lambda^{\frac{1-\beta}{2}}\right\} \tag{16}$$

where $C_1 > 1$ is a constant independent of m. By using (16) iteratively, we can find a small ball $\mathcal{B}_R$ such that it contains f_S with high confidence. Start with $R = R^{(0)} = 1/\sqrt{\lambda}$, by (16), we have $\mathcal{Z}^m = \mathcal{W}(R^{(0)}) \subseteq \mathcal{W}(R^{(1)}) \cup V_{R^{(0)}}$, where $R^{(1)} \leq \lambda^{(\beta-1)/2}\{2C_1 + C_1\lambda^{-\beta/4}\}$.

By inequality (16), for $j = 2, 3, \ldots,$ we iteratively derive

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

where each $V_{R^{(j)}}$ has measure at most δ and $R^{(j)}$ is given by $R^{(j)} \leq \lambda^{(\beta-1)/2}\{2jC_1 + C_1\lambda^{-(\beta/2)(1/2)^j}\}$. For $\epsilon > 0$, choose $J \in \mathbb{N}$ such that $(1/2)^{J+1} \leq \epsilon \cdot (1+\beta)(1+s)/(2\beta)$. Thus, we have that for $S \in \mathcal{W}(R^{(J)})$ $||f_S||_K \leq \lambda^{(\beta-1)/2}\{2JC_1 + C_1\lambda^{-(\beta/2)(1/2)^J}\}$. This together with inequality (16) gives

$$\mathcal{E}(f_S) - \mathcal{E}(f_\psi) \leq \Delta \leq \widehat{C}\left(\frac{1}{m}\right)^{\theta},\ \forall S \in \mathcal{W}\left(R^{(J)}\right)\backslash V_{R^J}.$$

Since $\bigcup_{k=0}^{j-1} V_{R^{(k)}}$ has measure at most $J\delta$, replacing δ by δ/J, the measure of $\mathcal{W}(R^{(J)})\backslash V_{R^{(J)}}$ is at least $1-\delta$. We complete the proof of Theorem 2. ■

IV. Numerical Studies

In this section, we introduce a new Markov sampling algorithm, and then we give the numerical studies on the learning performance of SVMC with Markov sampling.

A. New Markov Sampling Algorithm

For a given original training sample set D_{tr}, the new Markov sampling algorithm for SVMC is stated as follows.

Algorithm 1 Markov Sampling for SVMC

Step 1: Let m be the size of training samples and $m\%2$ be the remainder of m divided by 2. m_+ and m_- denote the size of training samples which label are $+1$ and -1, respectively. Draw randomly $N_1(N_1 \leq m)$ training samples $\{z_i\}_{i=1}^{N_1}$ from the dataset D_{tr}. Then we can obtain a preliminary learning model f_0 by SVMC and these samples. Set $m_+ = 0$ and $m_- = 0$.

Step 2: Draw randomly a sample from D_{tr} and denote it the current sample z_t. If $m\%2 = 0$, 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 another sample from D_{tr} 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 z_* with probability $P' = e^{-y_* f_0}/e^{-y_t f_0}$. If $P = 1$, $y_t = 1$ and $y_* = 1$ accept z_* with probability $P' = e^{-y_* f_0}/e^{-y_t f_0}$. If $P = 1$ and $y_t y_* = -1$ or $P < 1$, accept z_* with probability P. If there are k candidate samples z_* can not be accepted continuously, then set $P'' = qP$ and with probability P'' accept 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 [if the accepted probability P'(or P'', P) is larger than 1, accept z_* with probability 1].

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

Remark 3: To have better understanding Algorithm 1, we present the following remarks.

1) We introduce the notions $m\%2$, such that Algorithm 1 is suit to all the cases (even or odd) of training sample sizes. To generate quickly Markov chain samples, in Algorithm 1 we introduce the continuously reject number k and the constant q. Since as the loss $\ell(f, z_t)$ of sample z_t is smaller, the acceptance probability $P = e^{-\ell(f_0,z_*)}/e^{-\ell(f_0,z_t)}$ will be smaller. This implies that the candidate sample z_* will always be rejected, and generating u.e.M.c. samples is very time-consuming. In the following experiments, we take $k = 5$ and $q = 1.2$.
2) Since we have only the dataset D_{tr}, to define the transition probability of Markov chain, in Algorithm 1 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', P'' are always positive. Thus, by the theory of Markov chain in [37] (if the size of state space of Markov chain is finite, and the transition probabilities of any two states are always positive, then the Markov chain is u.e.M.c., please see [37, Th. 3.8]), we can conclude that $\{z_i\}_{i=1}^t$ generated by Algorithm 1 is a u.e.M.c. sequence.
3) Different from MCMC method in [19], Algorithm 1 is a method of generating u.e.M.c. samples from a given dataset, and in Algorithm 1 we did not use the information of distribution of training samples since the distribution of samples is unknown. While MCMC is a sampling method of using the probability distribution of training samples. Compared random sampling with Algorithm 1, we can find that random sampling can be regarded as the special case of Algorithm 1, that is, the acceptance probabilities P, P', P'' defined in Algorithm 1 are equal to 1.

B. Experiment Results

We present the numerical study on the learning performance of SVMC based on linear prediction models for 10 real-world

TABLE I
TEN REAL-WORLD DATASETS

Dataset	Training size	Test size	Input dimension
Abalone	2089	2088	8
Shuttle	43500	14500	9
Magic	12680	6340	10
Letter	13333	6667	16
Waveform	4600	400	21
Splice	2175	1000	60
DUSPS(0,2)	1100	1100	256
DUSPS(2,7)	1100	1100	256
Isolet	6238	1559	617
Gisette	6000	6000	5000

TABLE II
MISCLASSIFICATION RATES (%) FOR 500 TRAINING SAMPLES

Dataset	MR(i.i.d.)	MR(Markov)
Abalone	21.22 $\pm$ 0.47	**21.01 $\pm$ 0.38**
Shuttle	4.05 $\pm$ 0.49	**3.10 $\pm$ 0.24**
Magic	21.63 $\pm$ 0.65	**21.00 $\pm$ 0.34**
Letter	28.01 $\pm$ 0.77	**26.58 $\pm$ 0.43**
Waveform	12.71 $\pm$ 0.40	**12.08 $\pm$ 0.31**
Splice	12.90 $\pm$ 0.48	**12.21 $\pm$ 0.29**
DUSPS(0,2)	0.64 $\pm$ 0.18	**0.55 $\pm$ 0.17**
DUSPS(2,7)	3.07 $\pm$ 0.47	**2.49 $\pm$ 0.33**
Isolet	21.87 $\pm$ 1.46	**20.41** $\pm$ 1.49
Gisette	4.21 $\pm$ 0.30	4.57 $\pm$ 0.30

datasets. The information of these datasets are summarized in Table I and all these datasets are 2-classes real-world datasets.

For random sampling, we decompose the experiment into two steps. First, m training samples D_{rand} were generated randomly from the given dataset D_{tr}, and then we obtain a classifier by SVMC with these samples D_{rand}. Then we test it on the given test set. Second, after the experiment had been repeated 50 times, the misclassification rates were presented in Table II. For Markov sampling, we first generate m training samples D_{mar} from D_{tr} by Algorithm 1. Then we obtain a classifier by SVMC with these samples D_{mar}, and test it on the same test set. After the experiment had been repeated 50 times, the misclassification rates were presented in Table II, where "MR (i.i.d.)," "MR (Markov)" denote the misclassification rates of SVMC based on random sampling, Markov sampling, respectively.

To simplify the experimental process, we take $N_1 = m$ in Algorithm 1 for all of these experiments. In the next section, we will present some discussions on the experimental results based on $N_1 < m$. The parameter λ of SVMC is chosen by the method of fivefold cross-validation.

From Table II, we can find that for 500 training samples, the standard deviations and means of average misclassification rates of SVMC based on Markov sampling are smaller than that of random sampling except Isolet and Gisette. To show the learning performance of SVMC based on Markov sampling, we present the average misclassification rates of SVMC based on Markov sampling (Markov) and random sampling (i.i.d.) for different training sizes and four datasets in Figs. 1–4. These average misclassification rates are based on 50 times experimental results.

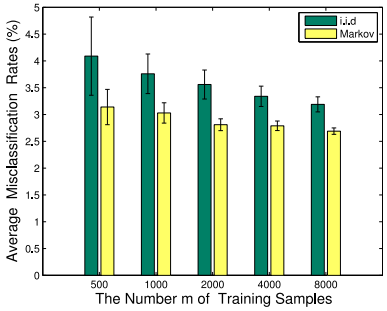

Fig. 1. Average misclassification rates for Shuttle and $m = 500$, 1000, 2000, 4000, 8000.

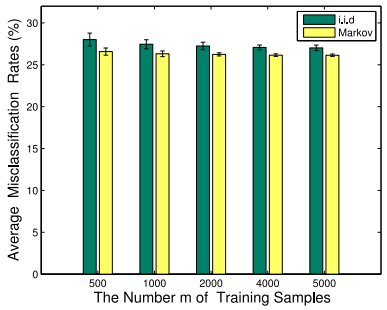

Fig. 2. Average misclassification rates for Letter and $m = 500$, 1000, 2000, 4000, 5000.

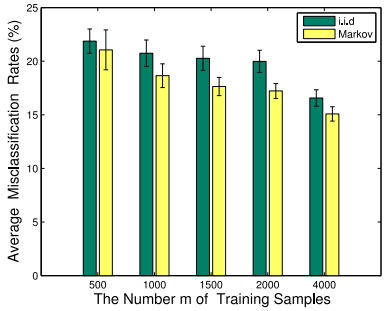

Fig. 3. Average misclassification rates for Isolet and $m = 500$, 1000, 1500, 2000, 4000.

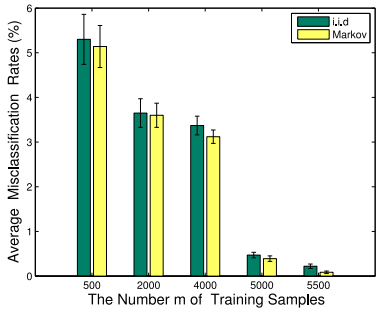

Fig. 4. Average misclassification rates for Gisette and $m = 500$, 2000, 4000, 5000, 5500.

To have a better understanding of learning performance of SVMC based on Markov sampling, we also present the following figures to show the 50 times experimental misclassification rates of SVMC based on Markov sampling.

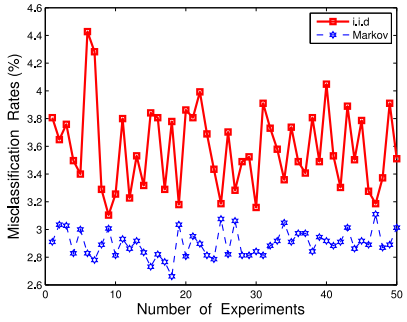

Fig. 5. Fifty times misclassification rates for Shuttle and $m = 2000$.

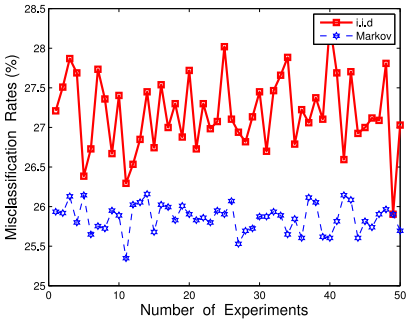

Fig. 6. Fifty times misclassification rates for Letter and $m = 2000$.

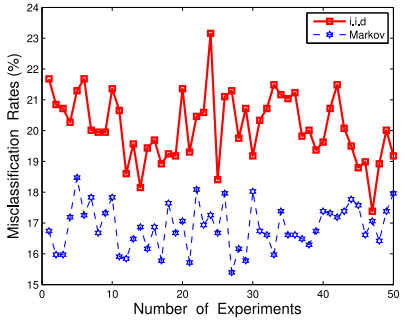

Fig. 7. Fifty times misclassification rates for Isolet and $m = 2000$.

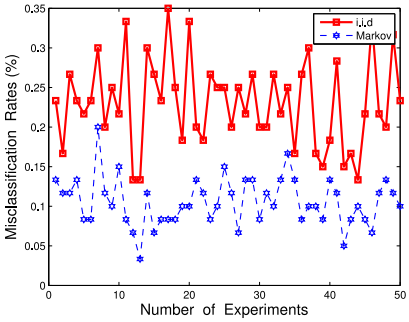

Fig. 8. Fifty times misclassification rates for Gisette and $m = 5500$.

Figs. 1–8 show that for Shuttle, Letter, Isolet, and Gisette, as the size of training samples is bigger, the experimental results of SVMC based on Markov sampling are better than that of random sampling. For other datasets, since the experimental results are similar, we do not present all of them here.

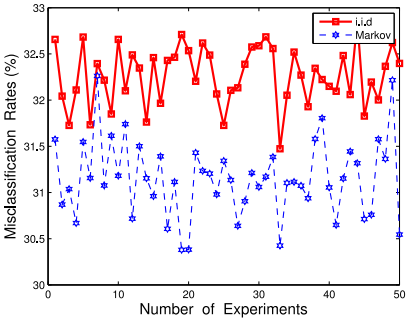

Fig. 9. Fifty times misclassification rates for Splice, Gaussian kernel, and $m = 1000$.

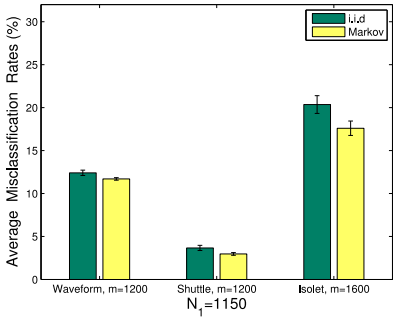

Fig. 10. Average misclassification rates for Waveform, $m = 1200$, Shuttle, $m = 1200$, Isolet, $m = 1600$, and $N_1 = 1150$.

V. Discussion

In this section, we present some discussions on the learning performance of SVMC based on Markov sampling.

A. Nonlinear Prediction Models

For nonlinear prediction models, we consider the case of Gaussian kernel SVMC with Markov sampling. We present the following figure to show the learning performance of Gaussian kernel SVMC with Markov sampling for Splice. The parameters λ and σ of Gaussian kernel SVMC are chosen by the method of fivefold cross-validation.

By Fig. 9, we can find that for Splice and 1000 training samples, all the 50 times misclassification rates of Gaussian kernel SVMC based on Markov sampling are better than that of random sampling. Different from the case of linear prediction models, the experiments of Gaussian kernel SVMC are very time-consuming, in particular for the dataset with high input dimension [38], [39].

B. Preliminary Learning Model Based on Smaller Samples

For the case of $N_1 < m$, we present the following figure (see Fig. 10) to show the learning performance of SVMC with Markov sampling for Waveform, Shuttle, and Isolet.

C. Sparsity of SVM Classifier

For SVMC, the optimal separating function $f(x)$ reduces to a linear combination of kernels on the training samples

$$f(x) = \sum_i k_i y_i K(x_i, x) + b. \tag{17}$$

TABLE III
AVERAGE NUMBERS OF SUPPORT VECTOR

Dataset	SVs(i.i.d.)	SVs(Markov)
Abalone-1800	895.96	**384.22**
Shuttle-700	73.40	**32.96**
Magic-1800	962.02	**429.66**
Letter-2000	1250.5	**720.04**
Waveform-1800	458.34	**205.66**
Splice-800	207.02	**95.18**
DUSPS(0,2)-1000	60.86	61.78
DUSPS(2,7)-1000	92.88	93.18
Isolet-1200	384	**314.74**
Gisette-5500	1186.90	1223.20

In (17), the vectors x_i that correspond to the nonzero coefficients k_i are called to be support vector [1]. If the numbers of support vector are smaller, then the express (17) is said to be "more sparse." In Table III, we present the average numbers of support vector of SVM classifier based on Markov sampling and random sampling for 50 experimental results, respectively. Here "Abalone-1800" denotes that the number of support vector is based on Abalone and 1800 training samples, "SVs(i.i.d.)" and "SVs(Markov)" denote the numbers of support vector for random sampling and Markov sampling, respectively.

By Table III, we can find that as the size of dataset is bigger with regard to the input dimension of data (e.g., Shuttle, Splice, Abalone, Magic, Letter, Waveform, and Isolet), the classifiers based on Markov sampling are more sparse than the classifier based on random sampling.

D. Explanation of Learning Performance

We interpret the learning performance of SVMC based on Markov sampling as follows. First, in the process of Markov sampling, the candidate sample z_* is accepted with different probabilities. While for random sampling, all the candidate samples are accepted with probability 1. Second, by these transition 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 or representative compared to random sampling. For these reasons, as the size of training samples is large, after many times transitions, the samples that closer (or closest) to the interface of two classes data will be sampled and be accepted with high probabilities. By the theory of statistical learning theory, the samples that closest to the interface of two classes data are the support vectors, which are the most "important" samples for classification problem [22]. Therefore, the learning performance of SVMC based on Markov sampling is better than that of random sampling, and the classifier based on Markov sampling is more sparse compared to random sampling as the size of training samples is bigger.

VI. CONCLUSION

To study the generalization performance of SVMC based on u.e.M.c. samples, inspired by the idea from [18], in this paper, we first establish two new concentration inequalities for u.e.M.c. samples, then we analyze the excess misclassification error of SVMC with u.e.M.c. samples, and obtain the optimal learning rate for SVMC with u.e.M.c. samples. These results extend the classical results of SVMC based on i.i.d. samples to the case of u.e.M.c. samples. In addition, in this paper, we also introduced a new Markov sampling algorithm to generate u.e.M.c. samples from given dataset. The numerical studies show that as the number of training samples is large, the learning performance of SVMC based on Markov sampling is better than that of random sampling, and the SVM classifier based on Markov sampling is more sparse compared to that of random sampling as the size of training samples is bigger with regard to the dimension of data. To our knowledge, these studies here are the first works on this paper.

Along the line of the present work, several open problems deserves further research. For example, studying the generalization performance of online learning based on Markov sampling and studying the Markov sampling algorithm for regression problems with nonlinear prediction models. All these problems are under our current investigation.

ACKNOWLEDGMENT

The authors would like to thank the Editor-in-Chief, the handling Associate Editor, and three anonymous referees, whose careful comments and valuable suggestions led to a significant improvement of the presentation of this paper.

REFERENCES

[1] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.
[2] Y. J. Tian, Z. Q. Qi, X. C. Ju, Y. Shi, and X. H. Liu, "Nonparallel support vector machines for pattern classification," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1067–1079, Jul. 2014.
[3] Z. Liu *et al*., "A three-domain fuzzy support vector regression for image denoising and experimental studies," *IEEE Trans. Cybern.*, vol. 44, no. 4, pp. 516–525, May 2014.
[4] T. Zhang, "Statistical behaviour and consistency of classification methods based on convex risk minimization," *Ann. Statist.*, vol. 32, no. 1, pp. 56–134, Mar. 2004.
[5] I. Steinwart, "Consistency of support vector machines and other regularized kernel classifiers," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 128–142, Jan. 2005.
[6] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer, 2008.
[7] I. Steinwart and A. Christmann, "Fast learning from non-i.i.d. observations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 22. Vancouver, BC, Canada, Dec. 2009, pp. 1768–1776.
[8] D. R. Chen, Q. Wu, Y. M. Ying, and D. X. Zhou, "Support vector machine soft margin classifiers: Error analysis," *J. Mach. Learn. Res.*, vol. 5, pp. 1143–1175, Sep. 2004.
[9] Q. Wu, Y. M. Ying, and D. X. Zhou, "Learning rates of least-square regularized regression," *Found. Comput. Math.*, vol. 6, no. 2, pp. 171–192, Apr. 2006.
[10] T. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations," *J. Multivariate Anal.*, vol. 100, no. 1, pp. 175–194, Jan. 2009.
[11] M. Vidyasagar, *Learning and Generalization with Applications to Neural Networks*, 2nd ed. London, U.K.: Springer, 2003.
[12] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *Ann. Probab.*, vol. 22, no. 1, pp. 94–116, Jan. 1994.
[13] M. Mohri and A. Rostamizadeh, "Stability bounds for stationary phi-mixing and beta-mixing processes," *J. Mach. Learn. Res.*, vol. 11, pp. 798–814, Feb. 2010.
[14] S. Smale and D. X. Zhou, "Online learning with Markov sampling," *Anal. Appl.*, vol. 7, pp. 87–113, Jan. 2009.

[15] B. Zou, L. Q. Li, and Z. B. Xu, "The generalization performance of ERM algorithm with strongly mixing observations," *Mach. Learn.*, vol. 75, no. 3, pp. 275–295, Jun. 2009.
[16] C. Zhang and D. Tao, "Generalization bounds of ERM-based learning processes for continuous-time Markov chains," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1872–1883, Dec. 2012.
[17] P. M. Laarhouen and E. L. Aarts, *Simulated Annealing: Theory and Application*. Norwell, MA, USA: Kluwer Acdemic, 1987.
[18] P. M. Samson, "Concentration of measure inequalities for Markov chains and Φ-mixing processes," *Ann. Probab.*, vol. 28, no. 1, pp. 416–461, Apr. 2000.
[19] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.
[20] K. Marton, "A measure concentration inequality for contracting Markov chains," *Geom. Funct. Anal.*, vol. 6, no. 3, pp. 556–571, May 1996.
[21] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, May 1950.
[22] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, Apr. 2000.
[23] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. New York, NY, USA: Springer-Verlag, 1993.
[24] P. Doukhan, *Mixing: Properties and Examples* (Lecture Notes in Statistics). Berlin, Germany: Springer, 1995.
[25] A. N. Kolmogorov, "On certain asymptotic characteristics of some completely bounded metric spaces," *Dokl. Akad. Nauk. SSSR*, vol. 108, no. 3, pp. 585–589, Mar. 1956.
[26] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. New York, NY, USA: Springer-Verlag, 1996.
[27] T. Zhang, "Covering number bounds of certain regularized linear function classes," *J. Mach. Learn. Res.*, vol. 2, pp. 527–550, Mar. 2002.
[28] D. X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1743–1752, Jul. 2003.
[29] 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. Inf. Theory*, vol. 44, no. 2, pp. 525–536, Mar. 1998.
[30] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, no. 4, pp. 1–49, Jan. 2001.
[31] I. Steinwart and C. Scovel, "Fast rates for support vector machines," in *Proc. 18th Annu. Conf. Learn. Theory*, Bertinoro, Italy, 2005, pp. 279–294.
[32] H. Z. Tong, D. R. Chen, and L. Z. Peng, "Analysis of support vector machine regression," *Found. Comput. Math.*, vol. 9, no. 2, pp. 243–257, Apr. 2009.
[33] I. Steinwart and C. Scovel, "Fast rates for support vector machines using Gaussian kernels," *Ann. Statist.*, vol. 35, no. 2, pp. 575–607, Jul. 2007.
[34] B. Zou, L. Q. Li, and Z. B. Xu, "Generalization performance of least-square regularized regression algorithm with Markov chain samples," *J. Math. Anal. Appl.*, vol. 388, no. 1, pp. 333–343, Apr. 2012.
[35] Y. L. Xu and D. R. Chen, "Learning rates of regularized regression for exponentially strongly mixing sequence," *J. Stat. Plan. Inference*, vol. 138, no. 7, pp. 2180–2189, Jul. 2008.
[36] F. Cucker and S. Smale, "Best choices for regularization parameters in learning theory: On the bias-variance problem," *Found. Comput. Math.*, vol. 2, no. 4, pp. 413–428, Oct. 2002.
[37] M. P. Qian and G. L. Gong, *Applied Random Processes*. Beijing, China: Peking Univ. Press, 1998.
[38] X. W. Liu *et al.*, "An adaptive approach to learning optimal neighborhood kernels," *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 371–384, Feb. 2013.
[39] X. W. Liu, L. Wang, J. P. Yin, E. Zhu, and J. Zhang, "An efficient approach to integrating radius information into multiple kernel learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 557–569, Apr. 2013.

Jie Xu received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2007.

She is an Associate Professor with the Faculty of Computer Science and Information Engineering, Hubei University, Wuhan.

Yuan Yan Tang (F'04) received the Degree in electrical and computer engineering from Chongqing University, Chongqing, China, the M.Eng. degree in electrical engineering from the Beijing Institute of Posts and Telecommunications, Beijing, China, and the Ph.D. degree in computer science from Concordia University, Montreal, QC, Canada.

He is a Professor with the Department of Computer Science, Chongqing University, Chongqing, China, a Chair Professor with the Faculty of Science and Technology, University of Macau, Macau, China, and an Adjunct Professor of Computer Science with Concordia University.

Bin Zou received the Ph.D. degree in mathematics from Hubei University, Wuhan, China, in 2007.

He was a Post-Doctoral Research Fellow with the Institute for Information and System Science, Xi'an Jiaotong University, Xi'an, China, from 2008 to 2009. He is a Professor with the Faculty of Mathematics and Statistics, Hubei University.

Zongben Xu received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 1987.

He is a Professor with the Institute for Information and System Sciences, Xi'an Jiaotong University. He is currently a Vice President with Xi'an Jiaotong University, and also the Chief Scientist of the National Basic Research Program of China (973 Project).

Prof. Xu was a member of the Chinese Academy of Science in 2011.

Luoqing Li received the B.Sc. degree from Hubei University, Wuhan, China, the M.Sc. degree from Wuhan University, Wuhan, and the Ph.D. degree from Beijing Normal University, Beijing, China.

He is a Full Professor with the Key Laboratory of Applied Mathematics of Hubei Province and the Faculty of Mathematics and Statistics, Hubei University, Wuhan.

Prof. Li has been the Managing Editor of the *International Journal on Wavelets, Multiresolution, and Information Processing*.

Yang Lu received the B.Sc. degree in software engineering from the University of Macau, Macau, China, in 2012, where he is currently pursuing the master's degree.

Baochang Zhang received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, 2001, and 2006, respectively.

From 2006 to 2008, he was a Research Fellow with the Chinese University of Hong Kong, Hong Kong, and Griffith University, Nathan, QLD, Australia. He is an Associate Professor with the Science and Technology on Aircraft Control Laboratory, School of Automation Science and Electrical Engineering, Beihang University, Beijing. He was selected by the Program for New Century Excellent Talents at the University of Ministry of Education of China in 2013.