The Generalization Ability of Online SVM Classification Based on Markov Sampling

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

***Abstract*— In this paper, we consider online support vector machine (SVM) classification learning algorithms with uniformly ergodic Markov chain (u.e.M.c.) samples. We establish the bound on the misclassification error of an online SVM classification algorithm with u.e.M.c. samples based on reproducing kernel Hilbert spaces and obtain a satisfactory convergence rate. We also introduce a novel online SVM classification algorithm based on Markov sampling, and present the numerical studies on the learning ability of online SVM classification based on Markov sampling for benchmark repository. The numerical studies show that the learning performance of the online SVM classification algorithm based on Markov sampling is better than that of classical online SVM classification based on random sampling as the size of training samples is larger.**

***Index Terms*— Generalization ability, Markov sampling, online support vector machine (SVM) classification, uniformly ergodic Markov chain (u.e.M.c.).**

I. Introduction

SUPPORT vector machine (SVM) is one of the most widely used machine learning algorithms for classification problems, in particular for classifying high-dimensional data [1]. Though besides their good learning performance in many practical applications, they also enjoy a good theoretical justification in terms of both universal consistency and learning rates [2]–[4], SVM algorithm might be practically challenging when the size T of training sample is very large. For example, when we consider the SVM algorithm with hinge loss function, solving it is a quadratic optimization problem. Its standard complexity is about $O(T^3)$. In particular, when $T \geq 10\,000$, the SVM algorithm is hard to implement [5]. While when the sample size is large, online learning algorithms with linear complexity $O(T)$ can be applied and provide efficient classifiers. In addition, the previously known works on the generalization ability of online SVM classification algorithm are usually based on the assumption that the training samples are independent identically distributed (i.i.d.) [5], [6]. Independence is a very restrictive concept in several ways [7]–[12]. 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 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 [7]–[9]. Therefore, relaxations of such i.i.d. assumption have been considered for quite a while in both machine learning and statistics literatures. For example, Yu [13] researched the convergence rates of empirical processes for stationary mixing sequences. Modha and Masry [14] considered the minimum complexity regression estimation with m-dependent observations and strongly mixing observations, respectively. Lozano *et al.* [15] showed that regularized boosting algorithms based on β-mixing processes are consistent. Kontorovich and Ramanan [16] established the concentration inequalities for dependent random variables via martingale method. Mohri and Rostamizdeh [17] studied the Rademarcher complexity bounds for non-i.i.d. processes. Smale and Zhou [18] considered the online regression learning algorithm based on Markov sampling. Hu and Zhou [19] studied the learning rates of online learning algorithm with samples drawn from nonidentical distributions. Agarwal and Duchi [20] extended the results on the generalization ability of online algorithms with i.i.d. samples established in [21] to the cases of β- and ϕ-mixing. For these reasons, we have to study the generalization ability of online SVM classification based on non-i.i.d. samples.

There are many non-i.i.d. sampling mechanisms (e.g., α-, β-, and φ-mixing) studied in machine learning literatures [7], [10], in this paper, we focus on the online SVM classification algorithm with uniformly ergodic Markov chain (u.e.M.c.) samples, the reasons are as follows. First, in real-world problems, Markov chain samples appear so often and

Manuscript received May 16, 2013; revised June 22, 2014; accepted September 21, 2014. Date of publication October 16, 2014; date of current version February 16, 2015. This work was supported in part by the National Science Foundation (NSF) of China under Project 11371007, Project 61370002, and Project 61403132, in part by the Multiyear Research, University of Macau, Macau, China, under Grant MYRG205(Y1-L4)-FST11-TYY and Grant MYRG187(Y1-L3)-FST11-TYY, in part by the Start-Up Research, University of Macau, under Grant SRG010-FST11-TYY, in part by the National Basic Research Program of China under Grant 2013CB329404, and in part by the NSF of Hubei Province under Grant 2011CDA003. *(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).

Y. Y. Tang is with the Faculty of Science and Technology, University of Macau, Macau, 999078, China (e-mail: yytang@umac.mo).

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. Lu is with the Faculty of Science, Hong Kong Baptist University, Hong Kong, 999077, China (e-mail: lylylytc@gmail.com).

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/TNNLS.2014.2361026

naturally in applications, such as biological (DNA or protein) sequence analysis, content-based web search, and market prediction. Zou *et al.* [22] presented two examples of learning from Markov chain input samples. Second, the generalization ability of online SVM classification based on Markov chain samples is unknown (particularly, it is unknown how well it performs in terms of consistency and generalization). In addition, inspired by the idea from Markov chain Monte Carlo methods [23], [24], in this paper, we introduce a new online SVM classification based on Markov sampling. Through the numerical studies on the learning performance of online SVM classification for benchmark repository, we find that the online SVM classification based on Markov sampling can provide smaller misclassification rates compared with random sampling.

The rest of this paper is organized as follows. In Section II, we introduce some useful definitions and notations. In Section III, we present the bounds on the generalization ability of online SVM classification based on u.e.M.c. samples. In Section IV, we introduce a new online SVM classification algorithm and present the numerical studies on the learning performance of online SVM classification based on Markov sampling. We conclude this paper in Section V.

II. Preliminaries

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

A. Online SVM Classification Algorithm

In this paper, we consider online SVM classification algorithm generated from Tiknonov regularization schemes associated with hinge loss function and reproducing kernel Hilbert spaces. Let $\mathcal{X}$ be a compact metric space and $\mathcal{Y} = \{-1, +1\}$. A binary classifier is a function $h : \mathcal{X} \to \mathcal{Y}$, which labels every point $x \in \mathcal{X}$ with some $y \in \mathcal{Y}$. A real-valued function $f : \mathcal{X} \to \mathbb{R}$ can be used to generate a classifier $h = sgn(f(x))$, where the sign function is defined as

$$\operatorname{sgn}(f(x)) = \begin{cases} +1, & \text{if } f(x) \geq 0 \\ -1, & \text{if } f(x) < 0. \end{cases}$$

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be continuous, symmetric, and positive semidefinite, i.e., for any finite set of distinct points $\{x_1, x_2, \dots, x_l\} \subset \mathcal{X}$, the matrix $(K(x_i, x_j))_{i,j=1}^{l}$ is positive semidefinite. Such a function is called a Mercer kernel. The reproducing kernel Hilbert space $\mathcal{H}_K$ 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 $\langle K_{x_i}, K_{x_j} \rangle_K = K(x_i, x_j)$

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

The reproducing property takes the form

$$\langle K_x, f \rangle_K = f(x) \quad \forall x \in \mathcal{X} \quad \forall f \in \mathcal{H}_K. \tag{1}$$

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 $\kappa = \sup_{x \in \mathcal{X}} (K(x, x))^{1/2}$, then the above reproducing property tells us that $||f||_\infty \leq \kappa ||f||_K$, $\forall f \in \mathcal{H}_K$.

The SVM classifier associated with the Mercer kernel K is defined as $\operatorname{sgn}(f_{\mathbf{z}})$, where $f_{\mathbf{z}}$ is a minimizer of the following optimization problem involving a set of random samples $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^{T} \in \mathcal{Z}^T$:

$$f_{\mathbf{z}} = \arg \min_{f \in \mathcal{H}_K} \frac{1}{2} ||f||_K^2 + \frac{C}{T} \sum_{i=1}^{T} \xi_i$$
$$\text{s.t. } y_i f(x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad 1 \leq i \leq T \tag{2}$$

where C is a constant that depends on T: $C = C(T)$ and often $\lim_{T \to \infty} C(T) = \infty$ [2], [3].

Algorithm (2) can be rewritten as a regularization scheme [3], [25]. Define the loss function $\ell(f, z)$ as

$$\ell(f, z) = \begin{cases} 0, & \text{if } f(x)y > 1 \\ 1 - f(x)y, & \text{if } f(x)y \leq 1 \end{cases} \tag{3}$$

which is called the hinge loss function [1]. The corresponding generalization error (or risk) is $\mathcal{E}(f) = E[\ell(f, z)]$. If we define the empirical error associated with the sample set $\mathbf{z}$ as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{T} \sum_{i=1}^{T} \ell(f, z_i) = \frac{1}{T} \sum_{i=1}^{T} (1 - y_i f(x_i))_+$$

then algorithm (2) can be written as [25]

$$f_{\mathbf{z}} = \arg \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \frac{1}{2C} ||f||_K^2 \right\}. \tag{4}$$

Algorithm (4) is an offline algorithm, which has been extensively studied in statistical machine learning literatures. In particular, the error analysis is well done [2], [3], [26]–[28]. However, since scheme (4) is a quadratic optimization problem, and its standard complexity is about $O(T^3)$, when $T \geq 10\,000$, algorithm (4) is hard to implement. This implies that algorithm (4) might be practically challenging when the sample size T is very large [5]. To overcome this problem, online learning algorithms are frequently adopted [5], [21], [29].

Definition 1: The online learning algorithm is defined by $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t \{\partial \ell(f_t, z_t) K_{x_t} + \lambda_t f_t\}, \quad t = 1, 2, \dots, T \tag{5}$$

where $\lambda_t > 0$ is the regularization parameter, η_t is called the step size, and $\partial \ell(f, z)$ is the left derivative of $\ell(f, z)$: $\partial \ell(f, z) := \lim_{\delta \to 0_-} (\ell(f + \delta, z) - \ell(f, z))/\delta$ [5].

We call the sequence $\{f_{t+1}\}$ the learning sequence for the online algorithm (5). The classifier is given by the sign function $\operatorname{sgn}(f_{T+1})$. In fully online algorithm, the regularization parameter λ_t changes with the learning step t. Throughout this paper, we assume that $\lambda_{t+1} \leq \lambda_t$ for each $t \in \mathbb{N}$.

For the classical SVM classification algorithm with loss function (3), algorithm (5) can be expressed as $f_1 = 0$ and

$$f_{t+1} = \begin{cases} (1 - \eta_t \lambda_t) f_t, & \text{if } y_t f_t(x_t) > 1 \\ (1 - \eta_t \lambda_t) f_t + \eta_t y_t K_{x_t}, & \text{if } y_t f_t(x_t) \leq 1. \end{cases} \tag{6}$$

Different from the previously known works on the generalization ability of online SVM classification algorithm in [5], [6], and [19], in this paper, we have to analyze the generalization ability of online SVM classification algorithm (6) based on u.e.M.c. samples.

B. Uniformly Ergodic Markov Chains

u.e.M.c. are the discrete cases of uniformly ergodic Markov processes. To present the definition of uniformly ergodic Markov processes, we first give the definition of total variation distance [20].

Definition 2: The total variation distance between distributions P and Q defined on the probability space $(S, \mathcal{F})$, where $\mathcal{F}$ is a σ-field, each with densities p and q with respect to an underlying measure ω, is given by

$$d_{\mathrm{TV}}(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)| = \frac{1}{2} \int_S |p(s) - q(s)| d\omega(s).$$

Define the σ-field $\mathcal{F}_t = \sigma(Z_1, Z_2, \ldots, Z_t)$. Let $P^t(\cdot|\mathcal{F}_s)$ is the conditional probability of Z_t given the signa field $\mathcal{F}_s = \sigma(Z_1, Z_2, \ldots, Z_s)$. A stochastic process $\{Z_t\}$ is said to possess the Markov property with respect to $\mathcal{F}_s$ if for each $A \in \mathcal{F}$, and each $s, t \in \mathbb{N}$ with $s < t$

$$P^t(\cdot|\mathcal{F}_s) = P^t(\cdot|Z_s).$$

In this paper, our main assumption is that there is a stationary distribution Π to which the distribution of Z_t converges as t grows, and the distributions $P^t(\cdot|\mathcal{F}_s)$ and Π are absolutely continuous with respect to an underlying measure ω throughout [30]–[32].

Definition 3: The Markov process $\{Z_t\}$ is said to be uniformly ergodic if there exist constants $\gamma_0 < \infty$ and $\alpha_0 < 1$ such that for any $Z \in \mathcal{Z}$, and for any $1 \leq t, t \in \mathbb{N}$

$$d_{TV}(P^t(\cdot|Z), \Pi) \leq \gamma_0 \alpha_0^t.$$

In the case, where S is a discrete set with the discrete sigma algebra, the discrete-time Markov chains is defined as follows: a Markov chain is a sequence of random variables $\{Z_t\}_{t \geq 1}$ together with a set of transition probability measures $P^n(z_{n+i}|z_i)$, $z_{n+i}, z_i \in \mathcal{Z}$. It is assumed that

$$P^n(z_{n+i}|z_i) := \mathrm{Prob}\{Z_{n+i} = z_{n+i}|Z_j, j < i, Z_i = z_i\}.$$

The fact that the transition probability does not depend on the values of Z_j prior to time i is the Markov property

$$P^n(z_{n+i}|z_i) = \mathrm{Prob}\{Z_{n+i} = z_{n+i}|Z_i = z_i\}.$$

This is commonly expressed in words as given the present, the future, and past states are independent [10], [31]. For u.e.M.c., we have the following remarks.

Remark 1: A weaker condition than uniformly ergodic is V-geometrically ergodic [10]. The difference between uniformly ergodic and V-geometrically ergodic is that here the total variation distance between the t-step transition probability $P^t(\cdot|Z)$ and the stationary distribution Π approaches zero at a geometric rate multiplied by $V(Z)$ [10], [30]. 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 [10].

By [33, in Th. 3.8], we have that if the size of a given data set is finite, and the transition probabilities of Markov chain $\{Z_t\}_{t \geq 1}$ generated from the data set are always positive, then $\{Z_t\}_{t \geq 1}$ is a u.e.M.c..

III. Estimating Misclassification Error

Let Π be a probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and (X, Y) be the corresponding random variable. The prediction ability of classification algorithms are often measured by the misclassification error, which is defined for a classifier $h : \mathcal{X} \to \mathcal{Y}$ to be the probability $\mathcal{R}(h)$ of $\{h(x) \neq y\}$

$$\mathcal{R}(h) = \mathrm{Prob}\{h(x) \neq y\} = \int_{\mathcal{X}} P(Y \neq h(x)|x) d\Pi_X(x).$$

Here, Π_X is the marginal distribution of Π on $\mathcal{X}$ and $P(\cdot|x)$ is the conditional probability measure given $X = x$. The best classifier minimizing the misclassification error is the Bayes rule [34], which can be expressed as $f_c = \mathrm{sgn}(f_\Pi)$, where f_Π is the regression function

$$f_\Pi(x) = \int_{\mathcal{Y}} y \Pi(y|x).$$

Recall that for the online learning algorithm (5) with loss function (3), we are interested in the classifier $\mathrm{sgn}(f_{T+1})$ generated by the real-valued function f_{T+1} from $\mathbf{z}$. The error analysis for online learning algorithm (5) is aimed at the excess misclassification error

$$\mathcal{R}(\mathrm{sgn}(f_{T+1})) - \mathcal{R}(f_c). \tag{7}$$

By the comparison theorem established in [27], we have that for loss function (3), the excess misclassification error (7) can often be done by bounding the excess generalization error

$$\mathcal{E}(f_{T+1}) - \mathcal{E}(f_c). \tag{8}$$

That is, an important relation between the excess misclassification error (7) and the excess generalization error (8) was given in [27] as:

$$\mathcal{R}(\mathrm{sgn}(f_{T+1})) - \mathcal{R}(f_c) \leq \mathcal{E}(f_{T+1}) - \mathcal{E}(f_c). \tag{9}$$

Therefore, to bound the generalization ability of online classification algorithm (5) with u.e.M.c. samples, it is sufficient for us to estimate the excess generalization error (8). To do so, we need the regularization or approximation error between regularizing function f_λ and f_c [5], [35].

Definition 4: The regularizing function f_λ is defined as

$$f_\lambda = \arg\min_{f \in \mathcal{H}_K} \left\{\mathcal{E}(f) + \lambda ||f||_K^2\right\} \tag{10}$$

where

$$\mathcal{E}(f) = E[\ell(f, z)] = \int_{\mathcal{Z}} \ell(f, z) d\Pi(z) = \int_{\mathcal{Z}} \ell(f, z) \pi(z) d\omega(z).$$

The regularization error $\mathcal{D}(\lambda)$ associated with the triple (K, ℓ, Π) is defined as

$$\mathcal{D}(\lambda) = \inf_{f \in \mathcal{H}_K} \left\{\mathcal{E}(f) - \mathcal{E}(f_c) + \frac{\lambda}{2} \|f\|_K^2\right\}.$$

By Definition 4, we have the following error decomposition [5] for the excess generalization error (8) as

$$\begin{aligned}\mathcal{E}(f_{T+1})-\mathcal{E}(f_c)&=\mathcal{E}(f_{T+1})-\mathcal{E}(f_\lambda)+\mathcal{E}(f_\lambda)-\mathcal{E}(f_c)\\&\le\{\mathcal{E}(f_{T+1})-\mathcal{E}(f_\lambda)\}+\mathcal{D}(\lambda).\end{aligned}\qquad(11)$$

The regularization error $\mathcal{D}(\lambda)$ is independent of the sample set $\mathbf{z}$. It can be estimated by the knowledge of approximation theory. For more details, see the discussions in [3] and [35]. The regularization error measures the approximation ability of the space $\mathcal{H}_K$ with respect to the learning process involving $\ell(f,\cdot)$ and Π. The denseness of $\mathcal{H}_K$ implies $\lim_{\lambda\to 0}\mathcal{D}(\lambda)=0$ [19]. A natural assumption is [5], [19]

$$\mathcal{D}(\lambda)\le\mathcal{D}_0\lambda^\beta,\quad\text{for some } 0\le\beta\le 1 \text{ and } \mathcal{D}_0>0.\qquad(12)$$

Hu and Zhou [19] have presented some examples for the assumption (12).

Throughout this paper, we assume that for the loss function (3) and any $z\in\mathcal{Z}$

$$M_0=\sup_{z\in\mathcal{Z}}\ell(0,z)+\sup\left\{\frac{|\ell(f,z)-\ell(0,z)|}{|f|},|f|\le 1\right\}<\infty.$$

Then, the assumption (12) holds true with $\beta=0$ and $\mathcal{D}_0=M_0$ since $\mathcal{D}(\lambda)\le\mathcal{E}(f)+(\lambda/2)||f||_K^2$, and then we have $\mathcal{D}(\lambda)\le\mathcal{E}(0)+0=M_0$ for any $\lambda>0$ by taking $f=0$ in the above inequality [5], [19].

The first term of right-hand side in (11) is called the sample error, which can be bounded by the error $\|f_{T+1}-f_\lambda\|_K$ for the loss function (3) [3], [5], [19]

$$\mathcal{E}(f_{T+1})-\mathcal{E}(f_\lambda)\le\kappa||f_{T+1}-f_\lambda||_K.\qquad(13)$$

By inequalities (9) and (13), we can find that to estimate the excess misclassification error (7), it is sufficient for us to estimate $||f_{T+1}-f_\lambda||_K$. Therefore, we first establish the following bound on the expectation of $||f_{t+1}-f_{\lambda_t}||_K^2$ for a fixed regularization parameter λ_t.

Proposition 1: Define $\{f_t\}$ by (5). Then, we have

$$\begin{aligned}E_{z_t}\big[||f_{t+1}-f_{\lambda_t}||_K^2\big]\le&(1-\lambda_t\eta_t)||f_t-f_{\lambda_t}||_K^2\\&+2\eta_t||\ell(f_{\lambda_t},z_t)\\&-\ell(f_t,z_t)||_\infty\cdot d_{TV}(P^t(\cdot|Z),\Pi)\\&+\eta_t^2E_{z_t}\big[||\partial\ell(f_t,z_t))K_{x_t}+\lambda_tf_t||_K^2\big].\end{aligned}\qquad(14)$$

For the proof of Proposition 1, refer to Appendix A. By Proposition 1, we establish the following bound on the excess misclassification error of online SVM classification-based u.e.M.c. samples.

Theorem 1: Let $\{z_t\}_{t=1}^T$ be u.e.M.c. samples, $\{f_t\}$ by (5) and with some $\lambda_1>0$, $\eta_1>0$, $0<\gamma,\alpha<1$, we take

$$\lambda_t=\lambda_1t^{-\gamma},\ \ \eta_t=\eta_1t^{-\alpha}\quad\forall t\in\mathbb{N}.$$

If $\eta_1\le 1/4\kappa^2+2+\lambda_1$, $\gamma<2/5$, and $\gamma<\alpha<1-3\gamma/2$, then

$$E_{z_1,\ldots,z_T}[\mathcal{R}(\mathrm{sgn}(f_{T+1}))-\mathcal{R}(f_c)]\le CT^{-\hat{\theta}}$$

where $\hat{\theta}=\min\{2-3\gamma-2\alpha,\ \alpha-\gamma,\ 2-2\gamma\}$, and C is a constant depending on $\eta_1,\lambda_1,\kappa,\alpha,\gamma,M_0,\gamma_0$, and α_0, which is given explicitly in the proof of Theorem 1.

For the proof of Theorem 1, refer to Appendix B. By Theorem 1, we also obtain the following bound on the learning rate of online SVM classification with u.e.M.c. samples.

Proposition 2: Let $\{z_t\}_{t=1}^T$ be u.e.M.c. samples, $\{f_t\}$ by (5) and for some $\lambda_1>0$, $0<\eta_1\le 1/(4\kappa^2+2+\lambda_1)$, $0<\epsilon<1/4$, we take

$$\lambda_t=\lambda_1t^{-\frac{1}{4}},\ \ \eta_t=\eta_1t^{\epsilon-\frac{1}{2}}\quad\forall t\in\mathbb{N}.$$

Then, we have

$$E_{z_1,\ldots,z_T}[\mathcal{R}(\mathrm{sgn}(f_{T+1}))-\mathcal{R}(f_c)]\le C_\epsilon T^{-\min\left\{\frac{1}{4}+2\epsilon,\frac{1}{4}-\epsilon\right\}}$$

where C_ϵ is a constant depending on $\eta_1,\lambda_1,\kappa,M_0,\gamma_0$, and α_0.

Ye and Zhou [6] considered the learning rate of fully online SVM classification algorithm based on random sampling (Proposition 1 in [6]). Comparing Proposition 2 with Proposition 1 in [6], we can find that the learning rate stated in Proposition 2 is same with that stated in Proposition 1 of [6]. This implies that we extended the online SVM classification algorithm with i.i.d. samples to the case of u.e.M.c. samples. Although [18] and this paper consider online learning algorithm based on Markov sampling and our proof techniques have many steps similar to that of [18], the difference is obvious: In this paper, we consider the online SVM classification algorithm while Smale and Zhou [18] researched the online regression algorithm. To our knowledge, these studies here are the first works on this topic.

IV. New Online SVM Classification Algorithm and Numerical Studies

In this section, we first introduce a novel online SVM classification algorithm based on Markov sampling, and then we present the numerical studies on the learning performance of online SVM classification based on benchmark data sets and linear prediction models

$$f=W^Tx+w^0=\sum_{i=1}^{d}w^ix^i+w^0\qquad(15)$$

where d is the (input) dimension of x, $w^i\in\mathbb{R}$ for any $0\le i\le d$. The linear prediction model (15) can be written as $f=\hat{W}^T\hat{X}$ with $\hat{W}=(w^0,w^1,\ldots,w^d)$, $\hat{x}=(1,x^1,\ldots,x^d)$. Thus, the corresponding classifier is defined as $\mathrm{sgn}(f)$ [1].

A. New Online SVM Classification Algorithm

In this section, we introduce a novel online SVM classification algorithm based on Markov sampling, the algorithm is defined as follows (Algorithm 1).

Remark 2: Since the training samples of classical online SVM classification algorithm based on random sampling [5], [21] are drawn randomly from a given data set D, comparing the classical online SVM classification algorithm based on random sampling with Algorithm 1, we can find that the classical online SVM classification algorithm based on random sampling can be regarded as the special case of Algorithm 1, that is, all the transition probabilities $\varphi_i=1$, $i=1,2,3$ defined in Step 3 of Algorithm 1.

Algorithm 1 Online SVM Classification Based on Markov Sampling

```
Step 1: Set $\hat{W} = 0$. Draw randomly a sample from $D$ and denote it the current sample $z_{t-1}$.
Step 2: Draw randomly another sample from $D$ and denote it the candidate sample $z_t$.
Step 3: For the case of $y_{t-1} f(x_{t-1}) < 1$, if $y_t f(x_t) < 1$, accept the candidate sample $z_t$ with the probability $\varphi_1 = \min\{1, e^{-\ell(f,z_t)}/e^{-\ell(f,z_{t-1})}\}$, and go to Step 4. Otherwise, accept the candidate sample $z_t$ with the probability $\varphi_2 = \min\{1, e^{-\ell(f,z_{t-1})}/e^{-\ell(f,z_t)}\}$, and go to Step 5.
        For the case of $y_{t-1} f(x_{t-1}) \geq 1$, if $y_t f(x_t) < 1$, accept the candidate sample $z_t$ with the probability $\varphi_3 = \max\{1, e^{-\ell(f,z_{t-1})}/e^{-\ell(f,z_t)}\}$, and go to Step 4. Otherwise, accept the sample $z_t$ with the probability $\varphi_2$, and go to Step 5.
Step 4: Set $\hat{W}_t = \hat{W}_{t-1} + \eta_t y_{t-1} x_{t-1}$. If $\hat{W}_t^T \hat{W}_t > 1/\lambda$, then $\hat{W}_t := \hat{W}_t / (\sqrt{\hat{W}_t^T \hat{W}_t \lambda})$ and go to Step 6.
Step 5: Set $\hat{W}_t = \hat{W}_{t-1}$, and go to Step 6.
Step 6: If $t < T$ then return to Step 2, else stop it. (Here $\eta_t = 1/\sqrt{t}$, $f$ is defined as (15), and $\ell(f, z)$ is defined as (3).)
```

In addition, since the transition probabilities $\varphi_i, i = 1, 2, 3$ defined in Algorithm 1 are always positive, by Remark 1, we can conclude that the training sample sequence $\{z_t\}_{t=1}^T$ generated by Algorithm 1 is a u.e.M.c..

B. Experimental Results

We present the numerical studies on the learning performance of online SVM classification algorithm with Markov sampling based on linear prediction models (15) for benchmark repository. The benchmark repository consists of 11 real-world data sets from abalone, shuttle, magic, mushrooms, isolet, letter, miniBooNE, gisette (http://archive.ics.uci.edu/ml/datasets.html), waveform, splice, and image (see http://www.fml.tuebingen.mpg.de/Members/raetsch/benchmark). We present the information of these data sets in Table I. All these data sets in Table I are two classes real-world data set except abalone, and abalone is redefined as two classes as follows: the sample whose label is equal or greater than 10 in abalone data set is viewed as a group and other samples are categorized as another group.

For online SVM classification based on random sampling, we draw randomly a sample from the given training set D_{train}, and use the first sample and the default hypothesis f_0 to generate the first hypothesis f_1. Next, we draw randomly a sample from the training set D_{train}, and use the second sample and the first hypothesis f_1 to generate the second hypothesis f_2, and so on [21]. At the end of this process, we obtain the hypothesis f_T. We use it to define the classifier $\text{sgn}(f_T)$, and we test it on a given test set D_{test}. After the experiment had been repeated 50 times, the average misclassification rates of online SVM classification algorithm based on random sampling were presented in Tables II–IV, where MR (Misclassification Rates)(i.i.d.) denotes the average misclassification rates of online SVM classification based on random sampling.

TABLE I

ELEVEN REAL-WORLD DATA SETS

Dataset	♯Training set	♯Test set	♯Input Features
abalone	2089	2088	8
shuttle	43500	14500	9
magic	12680	6340	10
letter	13333	6667	16
image	26000	20200	18
waveform	4600	400	21
miniBooNE	97548	32517	50
splice	12000	43500	60
mushrooms	8124	8124	112
isolet	6238	1559	617
gisette	6000	6000	5000

TABLE II

MISCLASSIFICATION RATES FOR 1000 TRAINING SAMPLES

Dataset	MR(i.i.d.)	MR(Markov)
abalone	0.2932 ± 0.0252	0.2804 ± 0.0119
shuttle	0.0641 ± 0.0077	0.0618 ± 0.0063
magic	0.2270 ± 0.0151	0.2191 ± 0.0079
letter	0.2949 ± 0.0124	0.2800 ± 0.0118
image	0.1872 ± 0.0150	0.1675 ± 0.0097
waveform	0.2267 ± 0.0079	0.2165 ± 0.0104
miniBooNE	0.3885 ± 0.2317	0.1707 ± 0.0768
splice	0.3728 ± 0.0348	0.3379 ± 0.0322
mushrooms	0.2042 ± 0.0435	0.0811 ± 0.0736
isolet	0.3032 ± 0.0247	0.2790 ± 0.0227
gisette	0.0667 ± 0.0058	0.0298 ± 0.0027

TABLE III

MISCLASSIFICATION RATES FOR 2000 TRAINING SAMPLES

Dataset	MR(i.i.d.)	MR(Markov)
abalone	0.2844 ± 0.0135	0.2714 ± 0.0029
miniBooNE	0.2960 ± 0.1674	0.1586 ± 0.0045
mushrooms	0.1467 ± 0.0328	0.0268 ± 0.0146
isolet	0.2779 ± 0.0182	0.2417 ± 0.0149
gisette	0.0497 ± 0.0033	0.0275 ± 0.0019

TABLE IV

MISCLASSIFICATION RATES FOR 8000 TRAINING SAMPLES

Dataset	MR(i.i.d.)	MR(Markov)
shuttle	0.0661 ± 0.0084	0.0577 ± 0.0022
magic	0.2169 ± 0.0059	0.2111 ± 0.0020
letter	0.2804 ± 0.0073	0.2582 ± 0.0031
image	0.1678 ± 0.0075	0.1441 ± 0.0046
waveform	0.2229 ± 0.0037	0.2007 ± 0.0065
splice	0.2557 ± 0.0180	0.2221 ± 0.0126

Different from the case of random sampling, for online SVM classification based on Markov sampling, the classifiers are generated by Algorithm 1 from the same training set D_{train}, and then we test it on the same test set D_{test}. After the experiment had been repeated 50 times, the average misclassification rates of online SVM classification based on Markov sampling were presented in Tables II–IV, where MR (Markov) denotes the average misclassification rates of online SVM classification based on Markov sampling.

By Tables II–IV, we can find that the means of misclassification rates based on Markov sampling are smaller than that of random sampling, and the standard deviations of misclassi-

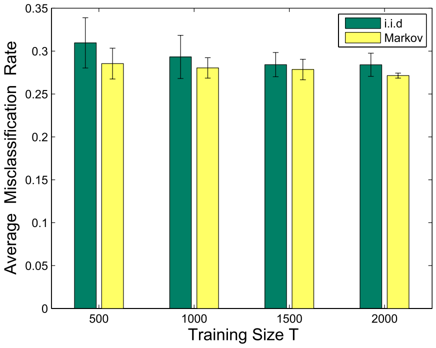

Fig. 1. Average misclassification rates for abalone and $T = 500$, 1000, 1500, and 2000.

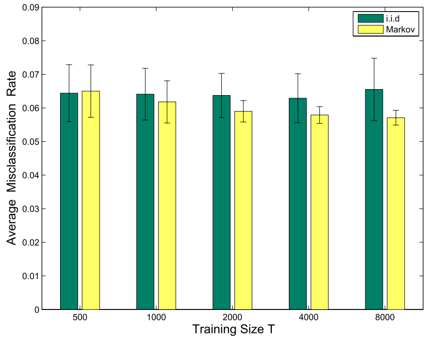

Fig. 2. Average misclassification rates for shuttle and $T = 500$, 1000, 2000, 4000, and 8000.

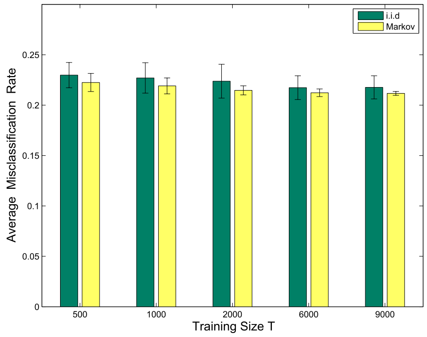

Fig. 3. Average misclassification rates for magic and $T = 500$, 1000, 2000, 6000, and 9000.

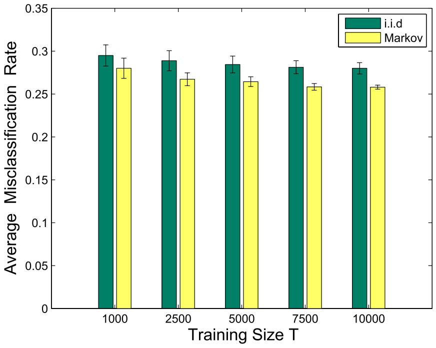

Fig. 4. Average misclassification rates for letter and $T = 1000$, 2500, 5000, 7500, and 10 000.

fication rates based on Markov sampling are also smaller than that of random sampling except mushrooms for 1000 training samples and waveform for 8000 training samples.

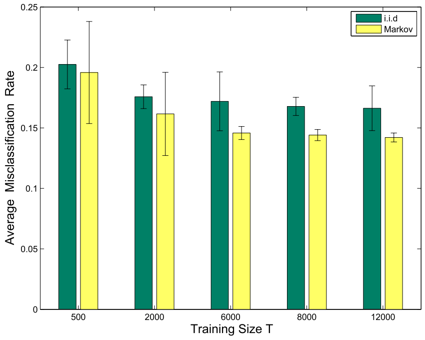

Fig. 5. Average misclassification rates for image and $T = 500$, 2000, 6000, 8000, and 12 000.

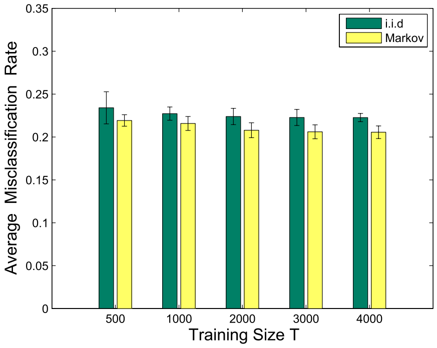

Fig. 6. Average misclassification rates for waveform and $T = 500$, 1000, 2000, 3000, and 4000.

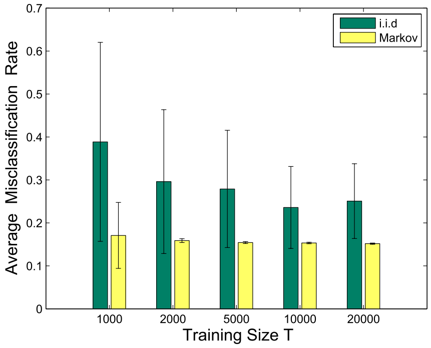

Fig. 7. Average misclassification rates for miniBooNE and $T = 1000$, 2000, 5000, 10 000, and 20 000.

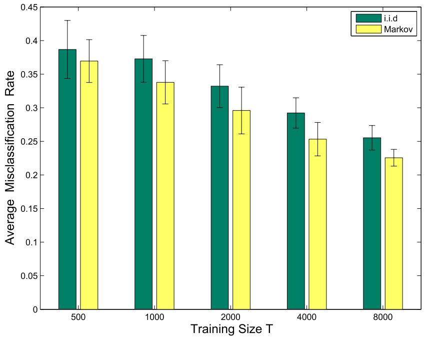

Fig. 8. Average misclassification rates for splice and $T = 500$, 1000, 2000, 4000, and 8000.

To show the learning performance of online SVM classification algorithm based on Markov sampling, we present the average misclassification rates of online SVM classification

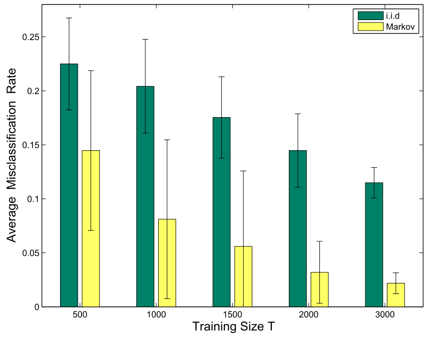

Fig. 9. Average misclassification rates for mushrooms and $T = 500$, 1000, 1500, 2000, and 3000.

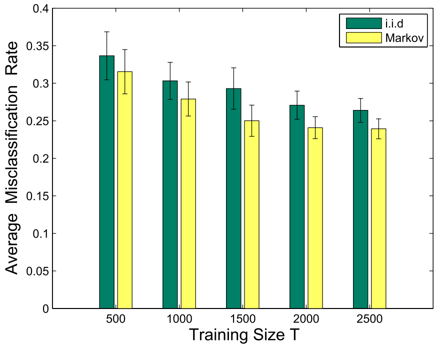

Fig. 10. Average misclassification rates for isolet and $T = 500$, 1000, 1500, 2000, and 2500.

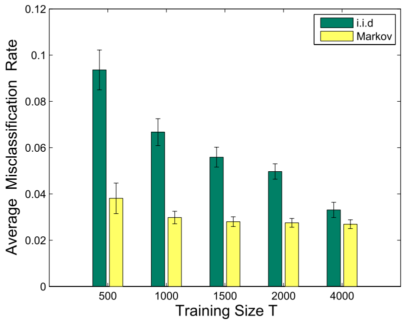

Fig. 11. Average misclassification rates for gisette and $T = 500$, 1000, 1500, 2000, and 4000.

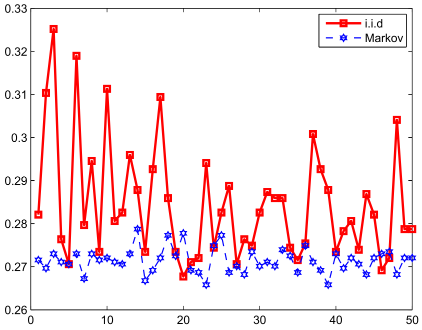

Fig. 12. Fifty times experimental misclassification rates for abalone and $T = 2000$.

algorithm based on Markov sampling (Markov) and random sampling (i.i.d.) for different training sizes in Figs. 1–11. These average misclassification rates in Figs. 1–11 are based on 50 times experimental results.

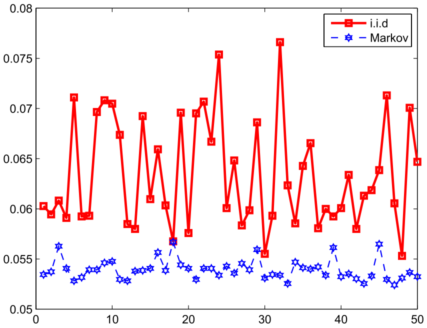

Fig. 13. Fifty times experimental misclassification rates for shuttle and $T = 18\,000$.

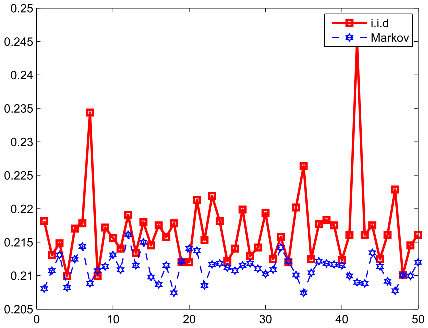

Fig. 14. Fifty times experimental misclassification rates for magic and $T = 8000$.

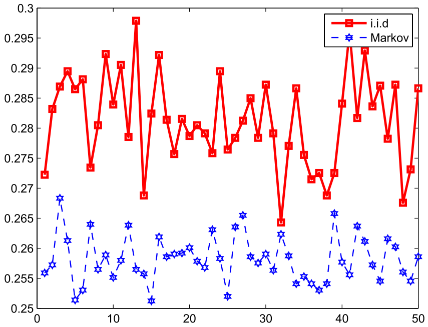

Fig. 15. Fifty times experimental misclassification rates for letter and $T = 7500$.

By Figs. 1–11, we can find that the average misclassification rates of online SVM classification based on Markov sampling are obviously smaller than that of random sampling as the training size is bigger.

To have a better understanding of the learning performance of online SVM classification based on Markov sampling, we also present the following figures to compare the 50 times misclassification rates of online SVM classification based on Markov sampling with that of random sampling. Here, red square denotes the results of random sampling, blue hexagram denotes the results of Markov sampling, and T is the size of training samples. 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.

By Fig. 12, we can find that for abalone and 2000 training samples, the 50 times experimental results of online SVM classification based on Markov sampling are better than that of random sampling expect two times experimental results.

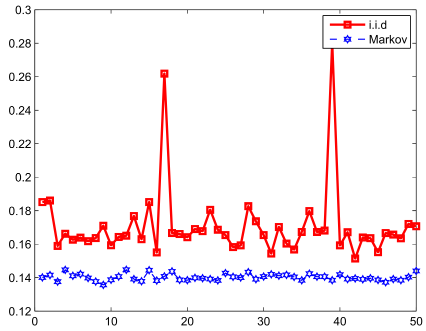

Fig. 16. Fifty times experimental misclassification rates for image and $T = 15\,000$.

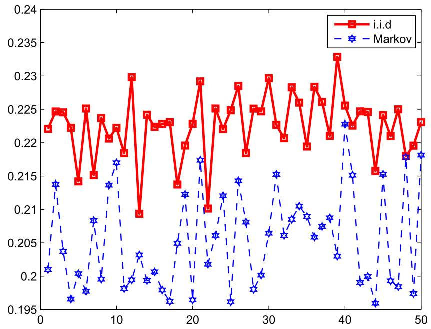

Fig. 17. Fifty times experimental misclassification rates for waveform and $T = 4000$.

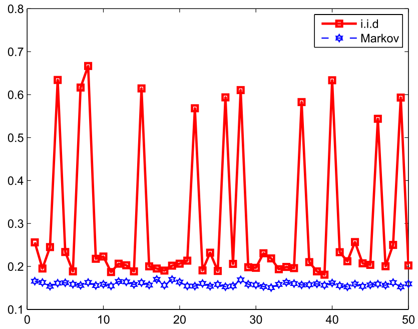

Fig. 18. Fifty times experimental misclassification rates for miniBooNE and $T = 2000$.

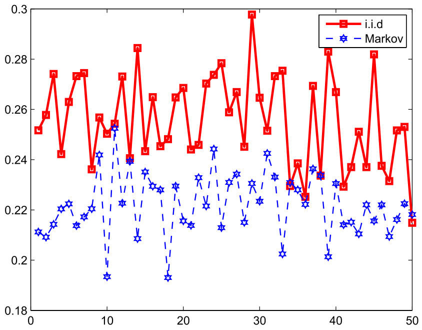

Fig. 19. Fifty times experimental misclassification rates for splice and $T = 8000$.

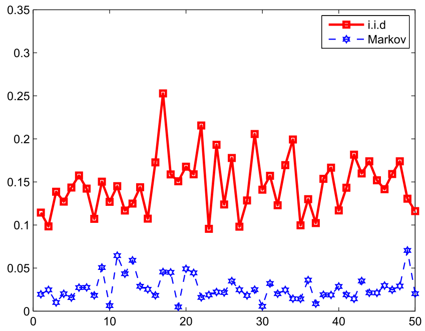

Fig. 20. Fifty times experimental misclassification rates for mushrooms and $T = 2000$.

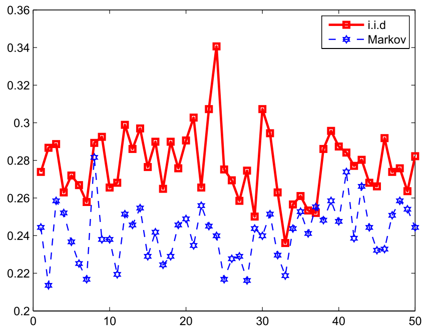

Fig. 21. Fifty times experimental misclassification rates for isolet and $T = 2000$.

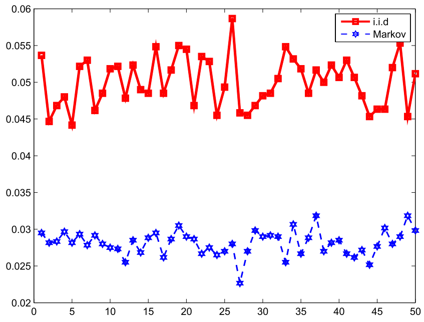

Fig. 22. Fifty times experimental misclassification rates for gisette and $T = 2000$.

By Figs. 13–22, we can find that for shuttle (or magic, letter, image, waveform, miniBooNE, splice, mushrooms, isolet, and gisette) for 18 000 (8000, 7500, 15 000, 4000, 2000, 8000, 2000, 2000, and 2000) training samples, all the 50 times experimental results of online SVM classification based on Markov sampling are better than that of random sampling, respectively.

V. Conclusion

Online classification algorithm is one of the most widely used machine learning algorithms for classification problems, in particular for large size of training samples [5]. Different from the previously known works on the generalization ability of online classification algorithms in [5], [6], and [18]–[20], in this paper, we considered the online SVM classification algorithm based on u.e.M.c. samples. We first established the bounds on the misclassification rates of online SVM classification based on u.e.M.c. samples and obtain a satisfactory convergence rate. We then introduced a novel online SVM classification algorithm based on Markov sampling. Through the numerical studies on the learning performance of online SVM classification based on 11 benchmark data sets and linear prediction models, we found that online SVM classification based on Markov sampling have better learning performance than that of random sampling as the size of training samples is larger. To our knowledge, these studies here are the first works of online classification algorithms on this topic.

Along the line of this paper, several open problems deserve further research. For example, studying the learning performance of online SVM classification based on Markov sampling for nonlinear prediction models and nonstationary data, respectively. Studying the generalization ability of online regression algorithms based on Markov sampling are under our current investigation.

Appendix A

In this section, we prove Proposition 1. For this purpose, we first present the following useful lemma and definition.

Definition 5: We say that $\ell(f,z)$ is a convex loss function if for any $z \in \mathcal{Z}$, the univariate function $\ell(\cdot, z)$ is convex [5].

We say that $\ell(f,z)$ is an admissible loss function if it is convex and differentiable at 0 with $\partial\ell(f,z) < 0$, where $\partial\ell(f,z)$ is defined as that in Definition 1.

Lemma 1: Let $\ell(f,z)$ be an admissible loss function and $\lambda > 0$ [5]. For any $f \in \mathcal{H}_K$, there holds

$$\frac{\lambda}{2}||f - f_\lambda||_K^2 \le \left\{\mathcal{E}(f) + \frac{\lambda}{2}||f||_K^2\right\} - \left\{\mathcal{E}(f_\lambda) + \frac{\lambda}{2}||f_\lambda||_K^2\right\}$$

where f_λ is defined by (10).

Proof of Proposition 1: Let $f_t^{\lambda_t} = \partial\ell(f_t, z_t)K_{x_t} + \lambda_t f_t$. By Definition 1, we have

$$\begin{aligned}||f_{t+1} - f_{\lambda_t}||_K^2 = ||f_t - f_{\lambda_t}||_K^2 + 2\eta_t\langle f_t^{\lambda_t}, f_{\lambda_t} - f_t\rangle_K \\ +\eta_t^2||f_t^{\lambda_t}||_K^2. \end{aligned} \quad (16)$$

By the reproducing property (1), we have

$$\begin{aligned}\langle f_t^{\lambda_t}, f_{\lambda_t} - f_t\rangle_K &= \langle \partial\ell(f_t, z_t)K_{x_t}, f_{\lambda_t} - f_t\rangle_K \\ &\quad +\lambda_t\langle f_t, f_{\lambda_t} - f_t\rangle_K \\ &\le \ell(f_{\lambda_t}, z_t) - \ell(f_t, z_t) \\ &\quad +\lambda_t\langle f_t, f_{\lambda_t} - f_t\rangle_K \\ &\le \ell(f_{\lambda_t}, z_t) - \ell(f_t, z_t) \\ &\quad +\frac{\lambda_t}{2}\left(||f_{\lambda_t}||_K^2 - ||f_t||_K^2\right).\end{aligned}$$

In the first inequality above, we use the fact that if ψ is a convex function on $\mathbb{R}$, $\psi'_-(a)(b-a) \le \psi(b) - \psi(a)$ for any $a, b \in \mathbb{R}$, where $\psi'_-(\cdot)$ is the left derivative of ψ. In the second inequality above, we use the Schwarz inequality

$$\langle f_t, f_{\lambda_t}\rangle_K \le ||f_t||_K||f_{\lambda_t}||_K \le \frac{1}{2}\left(||f_{\lambda_t}||_K^2 + ||f_t||_K^2\right).$$

In addition, by Lemma 1, we have

$$\frac{\lambda_t}{2}\left(||f_{\lambda_t}||_K^2 - ||f_t||_K^2\right) \le \mathcal{E}(f_t) - \mathcal{E}(f_{\lambda_t}) - \frac{\lambda_t}{2}||f_t - f_{\lambda_t}||_K^2.$$

By the above inequalities and inequality (16), we have

$$\begin{aligned}||f_{t+1} - f_{\lambda_t}||_K^2 \le\ & (1 - \lambda_t\eta_t)||f_t - f_{\lambda_t}||_K^2 \\ & +2\eta_t(\mathcal{E}(f_t) - \mathcal{E}(f_{\lambda_t})) \\ & +2\eta_t(\ell(f_{\lambda_t}, z_t) - \ell(f_t, z_t)) + \eta_t^2||f_t^{\lambda_t}||_K^2.\end{aligned}$$

Take the expectation with respect to z_t and notice that f_t depends on $\{z_1, z_2, \ldots, z_{t-1}\}$ but not on z_t, we have

$$\begin{aligned}E_{z_t}\left[||f_{t+1} - f_{\lambda_t}||_K^2\right] \le\ & 2\eta_t\big\{\mathcal{E}(f_t) - \mathcal{E}(f_{\lambda_t}) \\ & + E_{z_t}[\ell(f_{\lambda_t}, z_t)] - E_{z_t}[\ell(f_t, z_t)]\big\} \\ & + (1 - \lambda_t\eta_t)||f_t - f_{\lambda_t}||_K^2 \\ & + \eta_t^2 E_{z_t}\left[||f_t^{\lambda_t}||_K^2\right].\end{aligned} \quad (17)$$

We now bound the first term on the right-hand side of inequality (17), which is denoted by ϕ. By Definitions 2 and 4, we have

$$\begin{aligned}\phi :=\ & 2\eta_t\left\{\int_{\mathcal{Z}} \ell(f_{\lambda_t}, z)(p^t(\cdot|z) - \pi)d\omega(z) \right.\\ & \left. - \int_{\mathcal{Z}} \ell(f_t, z)(p^t(\cdot|z) - \pi)d\omega(z)\right\} \\ \le\ & 2\eta_t\int_{\mathcal{Z}} |\ell(f_{\lambda_t}, z) - \ell(f_t, z)| \cdot |p^t(\cdot|z) - \pi|d\omega(z) \\ \le\ & 2\eta_t||\ell(f_{\lambda_t}, z) - \ell(f_t, z)||_\infty \cdot d_{TV}(P^t(\cdot|Z), \Pi).\end{aligned}$$

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

Appendix B

In this section, we give the proof of Theorem 1. To prove Theorem 1, we first present the following useful lemmas and definitions.

Definition 6: The drift error is defined as

$$d_t = ||f_{\lambda_t} - f_{\lambda_{t-1}}||_K.$$

The drift error have been estimated by the regularization error in [6] and [18].

Lemma 2: Let ℓ be a convex loss function, f_λ by (10) and $\mu > \lambda > 0$ [19]. We have

$$\begin{aligned}||f_\lambda - f_\mu||_K &\le \frac{\mu}{2}\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)(||f_\lambda||_K + ||f_\mu||_K) \\ &\le \frac{\mu}{2}\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)\left(\sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}} - \sqrt{\frac{2\mathcal{D}(\mu)}{\mu}}\right).\end{aligned}$$

In particular, if for some $0 < \gamma \le 1$, we take $\lambda_t = \lambda_1 t^{-\gamma}$ for $t \ge 1$, then

$$d_t \le 2t^{\frac{\gamma}{2}-1}\sqrt{\mathcal{D}(\lambda_1 t^{-\gamma})/\lambda_1}.$$

Definition 7: Denote

$$N(\lambda) = \sup\{|\partial\ell(f,z)| : z \in \mathcal{Z}, |f| \le A_\lambda\}$$

where $A_\lambda = \max\left\{\kappa^2 M_0/\lambda, \kappa(2M_0/\lambda)^{1/2}\right\}$ [19].

We say that ℓ has incremental exponent $p > 0$ if for some $N_1 > 0$ and $\lambda_1 > 0$, we have

$$N(\lambda) \le N_1\left(\frac{1}{\lambda}\right)^p \quad \forall 0 < \lambda \le \lambda_1. \quad (18)$$

We say that $\partial\ell$ is locally Lipchitz at the origin if

$$M = \sup\left\{\frac{|\partial\ell(f,z) - \partial\ell(0,z)|}{|f|} : z \in \mathcal{Z}, |f| \le 1\right\} < \infty.$$

Lemma 3: Assume that $\partial\ell$ is locally Lipchitz at the origin [5]. Define $\{f_t\}$ by (5). If

$$\eta_t(\kappa^2(M + 2N(\lambda_t)) + \lambda_t) \le 1$$

for $t = 1, 2, \ldots, T$, then we have

$$||f_t||_K \le \frac{\kappa M_0}{\lambda_t}, \quad t = 1, 2, \ldots, T+1.$$

Lemma 4: 1) For any $c > 0, q_2 \ge 0$, and $0 < q_1 < 1$ [18]

$$\sum_{i=1}^{t-1} i^{-q_2} e^{-c\sum_{j=i+1}^{t} j^{-q_1}} \le \left(\frac{2^{q_1+q_2}}{c} + \left(\frac{1+q_2}{ec(1-2^{q_1-1})}\right)^{\frac{1+q_1}{1-q_1}}\right).$$

2) For any $c, a, \xi > 0$ [5]

$$\exp\{-c\xi\} \le \left(\frac{a}{ec}\right)^a \xi^{-a}.$$

To prove Theorem 1, we make use of the same procedure as that in [5] and [19]. A crucial estimate that differ from that [5] and [19] is the estimate on the second term of inequality (14).

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

Step 1: For the loss function (3), we have that $M \le 4$ for any $\lambda > 0$ (Corollary 5 in [5]). It follows that:

$$\begin{aligned}\eta_t[\kappa^2(M + 2N(\lambda_t)) + \lambda_t] &\le \eta_1\big[4\kappa^2 t^{-\alpha} + 2N_1\lambda_1^{-p} t^{\lambda p-\alpha} \\ &\qquad + \lambda_1 t^{-(\lambda+\alpha)}\big] \\ &\le \eta_1\big(4\kappa^2 + 2N_1\lambda_1^{-p} + \lambda_1\big) \le 1.\end{aligned}$$

In the last inequality above, we use the assumption $\eta_1 \le 1/(4\kappa^2 + 2N_1\lambda_1^{-p} + \lambda_1)$. Thus by Lemma 3, we have that for any $t = 1, 2, \ldots, T+1$

$$||f_t||_K \le \frac{\kappa M_0}{\lambda_t}.$$

Taking $f = 0$ in (10), we have that for any $t = 1, 2, \ldots, T$

$$||f_{\lambda_t}||_K^2 \le \frac{M_0}{\lambda_t}.$$

By Lemma 4 2) with $c = \ln(1/\alpha_0)$, $\xi = t$ and $a = 1$, we have

$$\alpha_0^t \le \left(\frac{2}{e\ln(1/\alpha_0)}\right)^2 \cdot t^{-2} \tag{19}$$

where α_0 is defined as that in Definition 3.

We denote the second and the third terms on the right-hand side of inequality (14) in Proposition 1 to be S_1 and S_2, respectively, by Definition 3, we have

$$\begin{aligned}S_1 &:= 2\eta_t||\ell(f_{\lambda_t}, z_t) - \ell(f_t, z_t)||_\infty \cdot d_{TV}(P^t(\cdot|Z), \Pi) \\ &\le 2\kappa\eta_t\left(\sqrt{\frac{M_0}{\lambda_t}} + \frac{\kappa M_0}{\lambda_t}\right)\gamma_0\alpha_0^t \\ &\le 2\kappa\eta_1\gamma_0\left(\frac{2}{e\ln(1/\alpha_0)}\right)^2\left(\sqrt{\frac{M_0}{\lambda_1}} + \frac{\kappa M_0}{\lambda_1}\right)t^{\gamma-\alpha-2} \\ &:= A_0 t^{\gamma-\alpha-2}\end{aligned}$$

where $A_0 = 2\kappa\eta_1\gamma_0(2/e\ln(1/\alpha_0))^2(\sqrt{M_0/\lambda_1} + \kappa M_0/\lambda_1)$, in the second inequality above, we use inequality (19).

Step 2: Recall that $d_t = ||f_{\lambda_t} - f_{\lambda_{t-1}}||_K$, we have

$$||f_t - f_{\lambda_t}||_K^2 \le ||f_t - f_{\lambda_{t-1}}||_K^2 + 2||f_t - f_{\lambda_{t-1}}||_K d_t + d_t^2.$$

Take $\tau = \gamma + \alpha/(1 - \gamma(1-\beta)/2)$, by the assumption of α in Theorem 1, we have $0 < \tau < 1$.

Taking $A_1 = \eta_1\lambda_1^{1+\tau(1-\beta)/2}/(2^{1+\tau}\mathcal{D}_0^{\tau/2})$ and using the following elementary inequality [19]:

$$2ab = 2\big[\sqrt{A_1}ab^{\tau/2}\big]\big[b^{1-\tau/2}/\sqrt{A_1}\big] \le A_1a^2b^\tau + b^{2-\tau}/A_1$$

to $a = ||f_t - f_{\lambda_{t-1}}||_K$ and $b = d_t$, we have

$$\begin{aligned}E_{z_1,\ldots,z_{t-1}}\big[||f_t - f_{\lambda_t}||_K^2\big] &\le \big(1 + A_1d_t^\tau\big)E_{z_1,\ldots,z_{t-1}}\big[||f_t - f_{\lambda_{t-1}}||_K^2\big] \\ &\quad + d_t^{2-\tau}/A_1 + d_t^2.\end{aligned}$$

By Lemma 2 and the assumption (12), we have

$$d_t \le 2\sqrt{\mathcal{D}_0}\lambda_1^{\frac{\beta-1}{2}} t^{\frac{\gamma(1-\beta)}{2}-1} := A_2 t^{\frac{\gamma(1-\beta)}{2}-1}$$

where $A_2 = 2\sqrt{\mathcal{D}_0}\lambda_1^{\beta-1/2}$.

In addition, by Definition 7, we have

$$\begin{aligned}S_2 &:= \eta_t^2 E_{z_t}\big[||\partial\ell(f_t, z_t))K_{x_t} + \lambda_t f_t||_K^2\big] \le \kappa^2\eta_t^2(N(\lambda_t) + M_0)^2 \\ &\le \kappa^2\eta_1^2\big(N_1\lambda_1^{-p} + M_0\big)^2 t^{-2\alpha+\gamma p}.\end{aligned}$$

Thus, by inequality (14), we have

$$\begin{aligned}&E_{z_1,\ldots,z_t}\big[||f_{t+1} - f_{\lambda_t}||_K^2\big] \\ &\quad\le \big(1 + A_1d_t^\tau - \eta_t\lambda_t\big)E_{z_1,\ldots,z_{t-1}}\big[||f_t - f_{\lambda_{t-1}}||_K^2\big] + A_3t^{-\theta}\end{aligned} \tag{20}$$

where $\theta = \min\{2 - \gamma(2-\beta) - \alpha,\ 2\alpha - \gamma p,\ 2 + \alpha - \gamma\}$ and

$$A_3 = A_2^{2-\tau}/A_1 + A_2^2 + A_0 + \kappa^2\eta_1^2\big(N_1\lambda_1^{-p} + M_0\big)^2.$$

In inequality (20), we use the inequality

$$(1 - \eta_t\lambda_t)\big(1 + A_1d_t^\tau\big) \le 1 + A_1d_t^\tau - \eta_t\lambda_t.$$

By the definitions of A_1 and A_2, we have

$$1 + A_1d_t^\tau - \eta_t\lambda_t = 1 - \frac{\eta_1\lambda_1}{2}t^{-\gamma-\alpha}.$$

From inequality (20), we have

$$\begin{aligned}&E_{z_1,\ldots,z_t}\big[||f_{t+1} - f_{\lambda_t}||_K^2\big] \\ &\quad\le \left(1 - \frac{\eta_1\lambda_1}{2}t^{-\gamma-\alpha}\right)E_{z_1,\ldots,z_{t-1}}\big[||f_t - f_{\lambda_{t-1}}||_K^2\big] + A_3t^{-\theta}.\end{aligned}$$

Step 3: Applying this above bound iteratively for $t = 1, 2, \ldots, T$ implies

$$\begin{aligned}&E_{z_1,\ldots,z_t}\big[||f_{t+1} - f_{\lambda_t}||_K^2\big] \\ &\quad\le A_3\sum_{t=1}^{T}\Pi_{j=t+1}^{T}\left(1 - \frac{\eta_1\lambda_1}{2}j^{-\gamma-\alpha}\right)t^{-\theta} \\ &\qquad + \left\{\Pi_{t=1}^{T}\left(1 - \frac{\eta_1\lambda_1}{2}t^{-\gamma-\alpha}\right)\right\}||f_1 - f_{\lambda_1}||_K^2 \\ &\quad := S_3 + S_4.\end{aligned} \tag{21}$$

Now, we bound the above two terms by two elementary inequalities in Lemma 4, respectively. Applying Lemma 4 (1) to $c = \eta_1\lambda_1/2$, $q_1 = \gamma + \alpha$, and $q_2 = \theta$, since $1 - u \le e^{-u}$ for any $u \ge 0$

$$\begin{aligned} S_3 &\le A_3 \sum_{t=1}^{T} \exp\left\{ -\frac{\eta_1\lambda_1}{2} \sum_{j=t+1}^{T} j^{-\gamma-\alpha} \right\} \\ &\le A_3 A_4 \cdot T^{\gamma+\alpha-\theta} \end{aligned}$$

where

$$A_4 = \frac{2^{\gamma+\alpha+\theta+1}}{\eta_1\lambda_1} + 1 + \left(\frac{2+2\theta}{e\eta_1\lambda_1(1-2^{\gamma+\alpha-1})} \right)^{\frac{1+\theta}{1-\gamma-\alpha}}.$$

Applying Lemma 4 2) to $c = \lambda_1\eta_1/2(1-\gamma-\alpha)$, $a = 2/(1-\gamma-\alpha)$, and $\xi = (T+1)^{1-\gamma-\alpha}$ and using Lemma 2, we have

$$\begin{aligned} S_4 &\le \exp\left\{ -\frac{\eta_1\lambda_1}{2} \sum_{j=1}^{T} t^{-\gamma-\alpha} \right\} \frac{2M_0}{\lambda_1} \\ &\le \exp\left\{ \frac{\lambda_1\eta_1}{2(1-\gamma-\alpha)} \right\} \left(\frac{4}{e\lambda_1\eta_1} \right)^{\frac{2}{1-\gamma-\alpha}} \frac{2M_0}{\lambda_1} \cdot T^{-2}. \end{aligned}$$

In addition, for the loss function (3), we have $N(\lambda) \equiv 1$ for any $\lambda > 0$ and $M = 0$ [6], hence $p = 0$. Moreover, for the loss function (3), we have that inequality (18) holds true with $N_1 = 1$, and $\mathcal{D}(\lambda) = 1$ for any $\lambda > 0$ [5]. This implies that $\mathcal{D}_0 = 1$ and $\beta = 1$.

Thus, by inequality (21), we have

$$E_{z_1,\ldots,z_t}\left[||f_{t+1} - f_{\lambda_t}||_K^2 \right] \le CT^{-\theta'}$$

where $\theta' = \min\{2 - 3\gamma - 2\alpha,\ \alpha - \gamma,\ 2 - 2\gamma\}$, and

$$C = \hat{A}_3 A_4 + \exp\left\{ \frac{\lambda_1\eta_1}{2(1-\gamma-\alpha)} \right\} \left(\frac{4}{e\lambda_1\eta_1} \right)^{\frac{2}{1-\gamma-\alpha}} \frac{2M_0}{\lambda_1}$$

$$\hat{A}_3 = 2^{\frac{6+\gamma}{2-\gamma}} / \eta_1 \lambda_1^{\frac{4}{2-\gamma}} + 2\lambda_1^{\frac{1}{2}} + A_0 + \kappa^2\eta_1^2(1+M_0)^2.$$

Combining inequalities (9), (13), and Proposition 1, we finish the proof of Theorem 1.

Acknowledgment

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

References

[1] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[2] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer-Verlag, 2008.

[3] D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou, "Support vector machine soft margin classifiers: Error analysis," *J. Mach. Learn. Res.*, vol. 5, pp. 1143–1175, Apr. 2004.

[4] B. Zou, Z. Peng, and Z. Xu, "The learning performance of support vector machine classification based on Markov sampling," *Sci. China Inf. Sci.*, vol. 56, no. 3, pp. 1–16, 2013.

[5] Y. Ying and D.-X. Zhou, "Online regularized classification algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4775–4788, Nov. 2006.

[6] G.-B. Ye and D.-X. Zhou, "Fully online classification by regularization," *Appl. Comput. Harmonic Anal.*, vol. 23, no. 2, pp. 198–214, 2007.

[7] I. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations," *J. Multivariate Anal.*, vol. 100, no. 1, pp. 175–194, 2009.

[8] B. Zou, Y. Y. Tang, Z. Xu, L. Li, J. Xu, and Y. Lu, "The generalization performance of regularized regression algorithms based on Markov sampling," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1497–1507, Sep. 2014.

[9] B. Zou, L. Li, Z. Xu, T. Luo, and Y. Y. Tang, "Generalization performance of Fisher linear discriminant based on Markov sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 2, pp. 288–300, Feb. 2013.

[10] M. Vidyasagar, *Learning and Generalization: With Applications to Neural Networks*. London, U.K.: Springer-Verlag, 2003.

[11] B. Zou, L. Li, and Z. Xu. "The generalization performance of ERM algorithm with strongly mixing observations," *Mach. Learn.*, vol. 75, no. 3, pp. 275–295, 2009.

[12] I. Steinwart and A. Christmann, "Fast learning from non-i.i.d. observations," in *Advances in Neural Information Processing Systems*, vol. 22. Red Hook, NY, USA: Curran & Associates Inc., 2009, pp. 1768–1776.

[13] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *Ann. Probab.*, vol. 22, no. 1, pp. 94–116, 1994.

[14] D. S. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2133–2145, Nov. 1996.

[15] A. C. Lozano, S. R. Kulkarni, and R. E. Schapire, "Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations," in *Advances in Neural Information Processing Systems*, vol. 18, Y. Weiss, B. Schölkopf, and J. C. Platt, Eds. Cambridge, MA, USA: MIT Press, 2006, pp. 819–826.

[16] L. Kontorovich and K. Ramanan, "Concentration inequalities for dependent random variables via the martingale method," *Ann. Probab.*, vol. 36, no. 6, pp. 2126–2158, 2008.

[17] M. Mohri and A. Rostamizadeh, "Rademacher complexity bounds for non-i.i.d. processes," in *Advances in Neural Information Processing Systems*. Vancouver, BC, Canada: MIT Press, 2009.

[18] S. Smale and D.-X. Zhou, "Online learning with Markov sampling," *Anal. Appl.*, vol. 7, no. 1, pp. 87–113, 2009.

[19] T. Hu and D.-X. Zhou, "Online learning with samples drawn from non-identical distributions," *J. Mach. Learn. Res.*, vol. 10, pp. 2873–2898, Sep. 2009.

[20] A. Agarwal and J. C. Duchi, "The generalization ability of online algorithms for dependent data," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 573–587, Jan. 2013.

[21] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "On the generalization ability of on-line learning algorithms," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2050–2057, Sep. 2004.

[22] B. Zou, H. Zhang, and Z. Xu, "Learning from uniformly ergodic Markov chains," *J. Complexity*, vol. 25, no. 2, pp. 188–200, 2009.

[23] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[24] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[25] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, 2000.

[26] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Statist. Assoc.*, vol. 101, no. 473, pp. 138–156, 2006.

[27] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Ann. Statist.*, vol. 32, no. 1, pp. 56–85, 2004.

[28] Q. Wu and D.-X. Zhou, "SVM soft margin classifiers: Linear programming versus quadratic programming," *Neural Comput.*, vol. 17, no. 5, pp. 1160–1187, 2005.

[29] S. Smale and Y. Yao, "Online learning algorithms," *Found. Comput. Math.*, vol. 6, no. 2, pp. 145–170, 2006.

[30] G. O. Roberts and J. S. Rosenthal, "General state space Markov chains and MCMC algorithms," *Probab. Surv.*, vol. 1, pp. 20–71, Mar. 2004.

[31] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. New York, NY, USA: Springer-Verlag, 1993.

[32] D. Aldous, L. Lovász, and P. Winkler, "Mixing times for uniformly ergodic Markov chains," *Stochastic Process. Appl.*, vol. 71, no. 2, pp. 165–185, 1997.

[33] M. P. Qian and G. L. Gong, *Applied Random Processes*. Beijing, China: Peking Univ., 1998.

[34] L. Devroye, L. Gyöorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York, NY, USA: Springer-Verlag, 1997.

[35] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, 2001.

Jie Xu received the B.Sc. degree from Hubei University, Wuhan, China, and the M.Sc. and Ph.D. degrees from the Huazhong University of Science and Technology, Wuhan.

She is currently an Associate Professor with the Faculty of Computer Science and Information Engineering, Hubei University. Her current research interests include machine learning and pattern recognition.

Yuan Yan Tang (S'88–M'88–SM'96–F'04) received the B.Sc. degree in electrical and computer engineering from Chongqing University, Chongqing, China, the M.Eng. degree in electrical engineering from the Beijing Institute of Post and Telecommunications, Beijing, China, and the Ph.D. degree in computer science from Concordia University, Montreal, QC, Canada.

He is currently a Chair Professor with the Faculty of Science and Technology, University of Macau, Macau, China, and holds positions as a Professor, Adjunct Professor, and Honorary Professor with several institutes, including Chongqing University, Chongqing, China, Concordia University, and Hong Kong Baptist University, Hong Kong. His current research interests include wavelet theory and applications, pattern recognition, image processing, document processing, artificial intelligence, and Chinese computing.

Prof. Tang is a fellow of the International Association for Pattern Recognition. He is the Founder and Editor-in-Chief of the *International Journal on Wavelets, Multiresolution, and Information Processing*, and the Associate Editor-in-Chief of the *International Journal on Frontiers of Computer Science*. He is the Founder and Chair of Pattern Recognition Committee in the IEEE International Conference on Systems, Man, and Cybernetics. He is the Founder and General Chair of the series International Conferences on Wavelets Analysis and Pattern Recognition.

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 currently with the Key Laboratory of Applied Mathematics and the Faculty of Mathematics and Statistics, Hubei University, where he became a Full Professor in 2014. His current research interests include statistical learning theory, machine learning, and pattern recognition.

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

He currently serves as the Vice President of Xi'an Jiaotong University, the Chief Scientist of the National Basic Research Program of China (973 Project), and the Director of the Institute for Information and System Sciences, Xi'an Jiaotong University. His current research interests include nonlinear functional analysis and intelligent information processing.

Dr. Xu was a recipient of the National Natural Science Award of China in 2007, and the CSIAM Su Buchin Applied Mathematics Prize in 2008. He delivered a 45-min talk on the International Congress of Mathematicians in 2010.

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 currently with the Key Laboratory of Applied Mathematics and the Faculty of Mathematics and Statistics, Hubei University, where he became a Full Professor in 1994. His current research interests include approximation theory, wavelet analysis, learning theory, signal processing, and pattern recognition.

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

Yang Lu received the B.S. and M.S. degrees in software engineering from the University of Macau, Macau, China, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree in computer science with Hong Kong Baptist University, Hong Kong.