Neurocomputing 167 (2015) 528–534

Contents lists available at ScienceDirect

Neurocomputing

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

Generalization ability of extreme learning machine with uniformly ergodic Markov chains

Peipei Yuan [a], Hong Chen [b,*], Yicong Zhou [c], Xiaoyan Deng [b,**], Bin Zou [d]

[a] College of Engineering, Huazhong Agricultural University, Wuhan 430070, China
[b] College of Science, Huazhong Agricultural University, Wuhan 430070, China
[c] Department of Computer and Information Sciences, University of Macau, Macau 999078, China
[d] Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China

ARTICLE INFO

Article history:
Received 3 February 2015
Received in revised form
7 April 2015
Accepted 15 April 2015
Communicated by G.-B. Huang
Available online 27 April 2015

Keywords:
Generalization ability
Extreme learning machine
Uniformly ergodic Markov chain

ABSTRACT

Extreme learning machine (ELM) has gained increasing attention for its computation feasibility on various applications. However, the previous generalization analysis of ELM relies on the independent and identically distributed (i.i.d) samples. In this paper, we go far beyond this restriction by investigating the generalization bound of the ELM classification associated with the uniform ergodic Markov chains (u.e.M.c) samples. The upper bound of the misclassification error is estimated for the ELM classification showing that the satisfactory learning rate can be achieved even for the dependent samples. Empirical evaluations on real-word datasets are provided to compare the predictive performance of ELM with independent and Markov sampling.

1. Introduction

Extreme learning machine (ELM) can be considered as a single-hidden layer feedforward neural networks (FNNs), where the output weights can be adjusted while the input weights and the threshold of hidden layer are fixed randomly [6,9]. This idea of training FNNs is different from the traditional neural network theories and is related with the discussions in [13,14]. Because only the Moore–Penrose generalized inverse is necessary to be calculated, the original ELM and its variations have shown the computation feasibility in the various applications, see, e.g., [2,4,5,11,23]. With the rapid development of the ELM-based applications, there are some theoretical works for its universal consistency in [25] and generalization ability in [10,19,2]. In particular, the generalization bounds of ELM are established in [10], which demonstrate that ELM can achieve the same learning rates as FNNs under mild conditions. Moreover, analysis of the generalization ability is extended to the magnitude-preserving regularization ranking in [2]. Although these works enrich our understanding of ELM, they just consider the setting where the samples are drawn independently from an unknown distribution. In the real-world applications, the independence of samples is difficult to be verified and does not hold true usually [20,16,26,28]. Therefore, it is important to further investigate the generalization ability of ELM with dependent samples.

Recently, the Markov chain samples have attracted increasing attention in statistical learning theory. In [17], the learning rate is estimated for the online algorithm with the Markov chains. For the uniformly ergodic Markov chains (u.e.M.c), the generalization bounds are established for the regularized regression in [27] and support vector machines classification in [21,22]. Despite the rapid theoretical progresses, there is no any generalization analysis for the regularized ELM with dependent samples. To fill the theoretical gap, in this paper, we investigate the generalization ability of the ELM classification with the Markov samples. The derived results on theory and experiments demonstrate that the satisfying generalization performance can be reached by the ELM with Markov sampling.

The rest of this paper is organized as follows. ELM and some necessary definitions are introduced in Section 2. The main result on generalization analysis is presented for the ELM-based classification in Section 3. Some empirical examples are reported in Section 4. Finally, we conclude this paper in Section 5.

2. Preliminaries

Let $X \in R^d$ be the input space and $Y = \{-1, 1\}$. The training samples $\mathbf{z} = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ are drawn from a probability distribution ρ on $Z = X \times Y$. Given $\mathbf{z}$, the main goal of the classification algorithm is searching a predictor $f_{\mathbf{z}} : X \to Y$ such that

* Corresponding author.
** Principal Corresponding author.
E-mail addresses: chenh@mail.hzau.edu.cn (H. Chen), dxygh@mail.hzau.edu.cn (X. Deng).

http://dx.doi.org/10.1016/j.neucom.2015.04.041
0925-2312/

the misclassification rate is as low as possible. In learning theory, the misclassification risk is defined as

$$\mathcal{R}(f)=\int_Z I\{y\neq f(x)\}\,d\rho$$

and the Bayes risk is denoted by

$$\mathcal{R}^*=\min\int_Z I\{y\neq f(x)\}\,d\rho.$$

For the regression function $f_\rho=\int_Y y\,d\rho(y|x)$, we know that $\mathcal{R}^*=\mathcal{R}(f_c)$, where $f_c=\mathrm{sign}(f_\rho)$, and $\mathrm{sign}\{t\}=1$ if $t\geq 0$ and $\mathrm{sign}\{t\}=-1$ otherwise. The performance of a classifier is measured by the excess risk $\mathcal{R}(f)-\mathcal{R}(f_c)$. Since the indictor loss I is nonconvex and noncontinuous, we usually use the convex loss to replace it. In original ELM, the least square loss $\ell(f,z)=(y-f(x))^2$ is used.

Denote $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_n)^T\in\mathbb{R}^{n\times l}$ in which α_i is generated independently and identically according to a uniform distribution μ on $[0,1]^l$. In ELM, the hypothesis space is defined as

$$\mathcal{M}_n=\left\{f_n(x,\alpha,\beta)=\sum_{i=1}^n\beta_i\phi(\alpha_i,x):x\in X,\beta=(\beta_1,\ldots,\beta_n)^T\in\mathbb{R}^n\right\},\tag{1}$$

where $\phi:\mathbb{R}^l\times\mathbb{R}^d\to\mathbb{R}$ is an activation function. The activation functions include the sigmoid function, Gaussian function, hyperbolic tangent function, multiquadric function and Fourier series [7,8,5].

For $f\in\mathcal{M}_n$, define

$$\|f\|_{\ell_2}^2=\inf\left\{\sum_{i=1}^n\beta_i^2:f=\sum_{i=1}^n\beta_i\phi(\alpha_i,\cdot)\right\}$$

Under the Tikhonov regularization scheme, the regularized ELM (see [5,6]) can be formulated as

$$f_{\mathbf{z},\lambda}=\arg\min_{f\in\mathcal{M}_n}\{\mathcal{E}_{\mathbf{z}}(f)+\lambda\|f\|_{\ell_2}^2\},\tag{2}$$

where

$$\mathcal{E}_{\mathbf{z}}(f)=\frac{1}{m}\sum_{i=1}^m(y_i-f(x_i))^2$$

is the empirical risk and $\lambda>0$ is the regularization parameter.

The regularized ELM can be rewritten as the optimization scheme

$$\beta^*=\arg\min_\beta\left\{\frac{1}{m}\|H\beta-Y\|_2^2+\lambda\|\beta\|_2^2\right\},$$

where $Y=(y_1,y_2,\ldots,y_m)^T$ and

$$H=\begin{pmatrix}\phi(\alpha_1,\mathbf{x}_1) & \cdots & \phi(\alpha_n,\mathbf{x}_1)\\ \vdots & \cdots & \vdots\\ \phi(\alpha_1,\mathbf{x}_m) & \cdots & \phi(\alpha_n,\mathbf{x}_m)\end{pmatrix}_{m\times n}.$$

It is easy to verify that

$$\beta^*=(H^TH+\lambda mI)^{-1}H^TY.$$

The expected convex risk, associated with the least square loss, is defined as

$$\mathcal{E}(f)=\int_Z(y-f(x))^2\,d\rho(x,y).$$

Let $L^2_{\rho_X}$ be the Hilbert space consisted all square integrable functions on X, with norm $\|\cdot\|_\rho$. For every $f\in L^2_{\rho_X}$, we have $\mathcal{E}(f)-\mathcal{E}(f_\rho)=\|f-f_\rho\|_\rho^2$. From [24], we know

$$\mathcal{R}(f)-\mathcal{R}(f_c)\leq\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\rho)}=\|f-f_\rho\|_\rho.$$

This paper focuses on bounding the excess risk $\mathcal{E}(f)-\mathcal{E}(f_\rho)$ to measure the generalization ability of ELM. The current analysis is based on the u.e.M.c samples different from the previous works in [10,19].

Now we recall some preliminary definition and properties of the u.e.M.c [12,18,22]. Let $(Z,\mathcal{S})$ be a measurable space. We call $\{Z_t\}_{t\geq 1}$ is a Markov chain, if the sequence $\{Z_t\}_{t\geq 1}$ is randomly generated and its transition probability measure satisfies

$$P^k(A|Z_i)=Prob\{Z_{k+i}\in A|Z_j,j<i,Z_i=z_i\}.\tag{3}$$

Starting from the initial state z_i at time i, the probability, that the state z_{k+i} will belong to set A after k-steps, is denoted by $P^k(A|Z_i)$. Hence, if $k=1$, we have $P^1(A|Z_i)=Prob\{Z_{i+1}\in A|Z_j,j<i,Z_i=z_i\}$, which is independent of the values of $Z_j(j<i)$. For the given probabilities p_1 and p_2, the total variance distance is defined as $\|p_1-p_2\|_{TV}=\sup_{A\in\mathcal{S}}|p_1(A)-p_2(A)|$. The definition of u.e.M.c can be described as below (see [20]).

Definition 1. A Markov chain $\{Z_t\}_{t\geq 1}$ is said to be uniformly ergodic if

$$\|P^k(\cdot|z)-\pi(\cdot)\|_{TV}\leq\gamma\tau^k,\tag{4}$$

for some $0<\gamma<\infty$ and $0<\tau<1$. Here $k\geq 1$, $k\in\mathbb{N}$ and $\pi(\cdot)$ is the stationary distribution of $\{Z_t\}_{t\geq 1}$.

From [12], we know that the transition probability $P^k(A|Z_i)$ of the u.e.M.c satisfies the Doeblin condition as below.

Proposition 1. *Let $\{Z_t\}_{t\geq 1}$ be a Markov chain with the transition probability measure $P^k(\cdot|\cdot)$ and let μ be a specific nonnegative measure with nonzero mass μ_0. Assume that, for some integer t and all measurable sets A, $P^t(A|z)\leq\mu(A)$, $\forall z\in Z$. Then, we have*

$$\|P^k(\cdot|z)-P^k(\cdot|z')\|_{TV}\leq 2(1-\mu_0)^{k/t},\quad\forall k\in\mathbb{N},\ z,z'\in Z.\tag{5}$$

3. Generalization bound

To evaluate the generalization ability of ELM, we should estimate the approximation between $f_{\mathbf{z},\lambda}$ and f_ρ. That is to say, we should estimate the excess convex risk $\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)$.

Proposition 2. *For any $\mathbf{z}\in Z^m$ and $f_{\mathbf{z},\lambda}$ defined in (2), there holds*

$$\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)\leq S_1+S_2,\tag{6}$$

where

$$S_1=\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})-\mathcal{E}_{\mathbf{z}}(f_\rho))$$

and

$$S_2=\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})-\mathcal{E}_{\mathbf{z}}(f_\rho)+\lambda\|f_{\mathbf{z},\lambda}\|_{\ell_2}^2.$$

Definition 2. For a subset $\mathcal{G}$ of a metric space and any $\epsilon>0$, the covering number $\mathcal{N}(\mathcal{G},\epsilon)$ is defined to be the smallest integer $l\in\mathbb{N}$ such that there exist l disks with radius ϵ and centers in $\mathcal{G}$ covering $\mathcal{G}$.

For any given $R>0$, we define a class of functions:

$$B_R=\{f\in\mathcal{M}_n:\|f\|_{\ell_2}^2\leq R^2\}.$$

The covering number of B_R is estimated in [3].

Lemma 1. *For any $R>0$, $\epsilon>0$, there holds*

$$\log\mathcal{N}(B_R,\epsilon)\leq n\cdot\log\left(\frac{4R}{\epsilon}\right).\tag{7}$$

Denote $\|\Gamma\|=\sqrt{2}/(1-(1-\mu_0)^{1/2t})$, where μ_0 and t are defined in Proposition 1. In fact, $\|\Gamma\|$ measures the "L^2-dependence" of the

random samples (see [18]). In order to estimate the generalization bound, we introduce the following lemma established in [22].

Lemma 2. *Let $\mathcal{G}$ be a countable class of bounded measurable functions and let $\mathbf{z}=\{z_i\}_{i=1}^m$ be a set of u.e.M.c samples. For some $C>0$, there exists $0\le g(z)\le C$ for all $g\in\mathcal{G}, z\in Z$. Then, for any $\epsilon>0$, we have*

$$Prob_{\mathbf{z}\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^m g(z_i)-E(g)\right|\ge\epsilon\right\}\le 2\exp\left\{\frac{-m\epsilon^2}{56C\|\Gamma\|^2E(g)}\right\}.$$

As shown in [22], the following lemma can be deduced by Lemma 2. For completeness, we presented its proof in Appendix.

Lemma 3. *Let $\mathcal{G}$ be a countable class of bounded measurable functions and let $\mathbf{z}=\{z_i\}_{i=1}^m$ be a set of u.e.M.c samples. For all $g\in\mathcal{G}, z\in Z$, assume that $0\le g(z)\le C$ for some $C>0$. Then, for any $\epsilon>0$, there holds that*

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m g(z_i)-E(g)}{\sqrt{E(g)+\epsilon}}\ge 4\sqrt{\epsilon}\right\}\le\mathcal{N}(\mathcal{G},\epsilon)\exp\left\{\frac{-m\epsilon}{56C\|\Gamma\|^2}\right\}.$$

Now, we present the main result on the excess risk $\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)$.

Theorem 1. *Assume that $\mathbf{z}=\{z_i\}_{i=1}^m$ is a set of u.e.M.c samples and $\|\phi\|_\infty\le\kappa$. For any $0<\delta<1$, there holds*

$$E_{\rho^m}E_{\mu^n}(\|f_{\mathbf{z},\lambda}-f_\rho\|_\rho^2)\le\frac{Cn\log(m/\lambda)}{m}+2E_{\mu^n}\inf_{f\in\mathcal{M}_n}\left(\int_X(f(x)-f_\rho(x))^2\,d\rho+\lambda\|f\|_{\ell_2}^2\right)$$

with confidence at least $1-\delta$, where $C=224(\kappa+\sqrt{\lambda})(\kappa+3\sqrt{\lambda})\|\Gamma\|^2/\lambda$.

Proof. From the Proposition 2, we conclude that

$$\begin{aligned}E_\rho^mE_\mu^n(\|f_{\mathbf{z},\lambda}-f_\rho\|_\rho^2)&\le E_\rho^mE_\mu^n(\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})-\mathcal{E}_{\mathbf{z}}(f_\rho)))\\&\quad+E_\rho^mE_\mu^n(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})-\mathcal{E}_{\mathbf{z}}(f_\rho)+\lambda\|f_{\mathbf{z},\lambda}\|_{\ell_2}^2)\\&=E_\rho^mE_\mu^n(S_1)+E_\rho^mE_\mu^n(S_2),\end{aligned}\tag{8}$$

Firsty, we estimate S_1. Set

$$\mathcal{G}_R=\{(y-f(x))^2-(y-f_\rho(x))^2:f\in B_R\},$$

for any $g\in\mathcal{G}_R$, there exists $f\in B_R$ such that

$$g(z)=(y-f(x))^2-(y-f_\rho(x))^2.$$

We can observe that

$$E_{\rho^m}(g)=\mathcal{E}(f)-\mathcal{E}(f_\rho)\ge 0$$

and

$$\frac{1}{m}\sum_{i=1}^m g(z_i)=\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho).$$

Since $\|\phi\|_\infty\le\kappa$, from Cauchy–Schwarz inequality, we have

$$|f(x)|=\left|\sum_{i=1}^n\beta_i\phi(\alpha_i,x_i)\right|\le\sqrt{\sum_{i=1}^n\beta_i^2}\sqrt{\sum_{i=1}^n\phi^2}\le\kappa R.$$

Then we deduce that

$$|g(z)|=|(y-f(x))^2-(y-f_\rho(x))^2|\le(\kappa R+1)(\kappa R+3).$$

By Lemma 3, we have that

$$\begin{aligned}&Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_R}\frac{\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)-(\mathcal{E}(f_\rho)-\mathcal{E}_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\rho)+\epsilon}}\ge 4\sqrt{\epsilon}\right\}\\&=Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_R}\frac{E(g)-\frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{E(g)+\epsilon}}\ge 4\sqrt{\epsilon}\right\}\\&\le\mathcal{N}(\mathcal{G}_R,\epsilon)\exp\left\{\frac{-m\epsilon}{56(\kappa R+1)(\kappa R+3)\|\Gamma\|^2}\right\}.\end{aligned}\tag{9}$$

For any $g_1,g_2\in\mathcal{G}_R, z\in Z$, there exists that

$$|g_1(z)-g_2(z)|\le 2(\kappa R+1)\|f_1-f_2\|_\infty.$$

Therefore, for any $\epsilon>0$, an $\epsilon/2(\kappa R+1)$-covering of B_R can provide ϵ-covering of $\mathcal{G}_R$. Accordingly,

$$\mathcal{N}(\mathcal{G}_R,\epsilon)\le\mathcal{N}\left(B_R,\frac{8R(\kappa R+1)}{\epsilon}\right).$$

From Lemma 1, we have

$$\log\mathcal{N}(\mathcal{G}_R,\epsilon)\le n\cdot\log\left(\frac{8R(\kappa R+1)}{\epsilon}\right).$$

Then, (9) tells us that

$$\begin{aligned}&Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_R}\frac{\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)-(\mathcal{E}(f_\rho)-\mathcal{E}_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\rho)+\epsilon}}\ge\sqrt{\epsilon}\right\}\\&\le\mathcal{N}(\mathcal{G}_R,\epsilon)\exp\left\{\frac{-m\epsilon}{56(\kappa R+1)(\kappa R+3)\|\Gamma\|^2}\right\}\\&\le\exp\left\{n\log\frac{8R(\kappa R+1)}{\epsilon}-\frac{m\epsilon}{56(\kappa R+1)(\kappa R+3)\|\Gamma\|^2}\right\}.\end{aligned}\tag{10}$$

Since

$$\sqrt{\epsilon}\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\rho)+\epsilon}\le\tfrac{1}{2}(\mathcal{E}(f)-\mathcal{E}(f_\rho))+\epsilon,$$

there exists that

$$\sup_{f\in B_R}\left\{(\mathcal{E}(f)-\mathcal{E}(f_\rho))-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho))\right\}\le\tfrac{1}{2}(\mathcal{E}(f)-\mathcal{E}(f_\rho))+\epsilon$$

with confidence at least

$$1-\exp\left\{n\log\frac{8R(\kappa R+1)}{\epsilon}-\frac{m\epsilon}{56(\kappa R+1)(\kappa R+3)\|\Gamma\|^2}\right\}.$$

Hence,

$$\begin{aligned}&Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_R}\left\{(\mathcal{E}(f)-\mathcal{E}(f_\rho))-2(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho))\right\}\le\epsilon\right\}\\&\ge 1-\exp\left\{n\log\frac{16R(\kappa R+1)}{\epsilon}-\frac{m\epsilon}{112(\kappa R+1)(\kappa R+3)\|\Gamma\|^2}\right\}.\end{aligned}$$

From the definition of $f_{\mathbf{z},\lambda}$, we can deduce that

$$\|f_{\mathbf{z},\lambda}\|_{\ell_2}^2=\sum_{i=1}^n|\alpha_i|^2\le\frac{1}{\lambda}.$$

Hence, $f_{\mathbf{z},\lambda}\in B_R$ with $R=1/\sqrt{\lambda}$.

Setting

$$\mathcal{K}=\{\mathcal{E}(f)-\mathcal{E}(f_\rho)-2(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho))\},$$

then we have

$$\mathcal{E}(f)-\mathcal{E}(f_\rho)\le\mathcal{K}+2S_2.\tag{11}$$

For any $t \geq 16(\kappa+\sqrt{\lambda})/m$, we conclude that

$$
\begin{aligned}
E_\rho^m(\mathcal{K}) &= \int_0^{+\infty} Prob_{\mathbf{z}\in Z^m}\left\{\mathcal{E}(f)-\mathcal{E}(f_\rho)-2(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho)) \leq \epsilon\right\} d\epsilon \\
&\leq t+\int_t^{+\infty} \exp\left\{ n \log \frac{16(\kappa+\sqrt{\lambda})}{\lambda}\epsilon - \frac{\lambda m\epsilon}{112(\kappa+\sqrt{\lambda})(\kappa+3\sqrt{\lambda})\|\Gamma\|^2}\right\} d\epsilon \\
&\leq t+\exp\left\{-\frac{\lambda m t}{112(\kappa+\sqrt{\lambda})(\kappa+3\sqrt{\lambda})\|\Gamma\|^2}\right\}\int_t^{+\infty}\left(\frac{16(\kappa+\sqrt{\lambda})}{\lambda\epsilon}\right)^n d\epsilon \\
&\leq t+\exp\left\{-\frac{\lambda m t}{112(\kappa+\sqrt{\lambda})(\kappa+3\sqrt{\lambda})\|\Gamma\|^2}\right\}\left(\frac{16(\kappa+\sqrt{\lambda})}{\lambda t}\right)^n t \\
&\leq t+\lambda^{-n}\exp\left\{-\frac{\lambda m t}{112(\kappa+\sqrt{\lambda})(\kappa+3\sqrt{\lambda})\|\Gamma\|^2}\right\} m^n t.
\end{aligned}
$$

Setting $t = 112n(\kappa+\sqrt{\lambda})(\kappa+3\sqrt{\lambda})\|\Gamma\|^2 \log m/\lambda m$,we have

$$
E_{\rho^m}(\mathcal{K}) \leq 2t = \frac{224n(\kappa+\sqrt{\lambda})(\kappa+3\sqrt{\lambda})\|\Gamma\|^2 \log(m/\lambda)}{\lambda m}.
$$

Now, we give the upper bound of $E_\rho^m(S_2)$:

$$
\begin{aligned}
E_\rho^m(S_2) &= E_\rho^m\left(\frac{1}{m}\sum_{i=1}^m (y_i - f_{\mathbf{z},\lambda}(x_i))^2 - \frac{1}{m}\sum_{i=1}^m (y_i - f_\rho(x_i))^2 + \lambda\|f_{\mathbf{z},\lambda}\|_{\ell_2}^2\right) \\
&= E_\rho^m\left\{\inf_{f\in\mathcal{M}_n}\left(\frac{1}{m}\sum_{i=1}^m (y_i - f_{\mathbf{z},\lambda}(x_i))^2 - \frac{1}{m}\sum_{i=1}^m (y_i - f_\rho(x_i))^2 + \lambda\|f\|_{\ell_2}^2\right)\right\} \\
&\leq \inf_{f\in\mathcal{M}_n}\left\{E_\rho^m(y-f(x))^2 - E_\rho^m(y-f_\rho(x))^2 + \lambda\|f\|_{\ell_2}^2\right\} \\
&= \inf_{f\in\mathcal{M}_n}\left(\int_X (f(x)-f_\rho(x))^2\, d\rho + \lambda\|f\|_{\ell_2}^2\right). \qquad (12)
\end{aligned}
$$

The desired result follows by combining the inequations (11) and (12).

From Theorem 1, we know that the excess convex risk $E_{\rho^m}E_{\mu^n}(\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho))$ depends on the sample number m, the net number n, the regularized parameter λ, and the hypothesis space $\mathcal{M}_n$. The generalization bound for ELM with the u.e.M.c samples is consistent with the result in [10] for the i.i.d samples.

Corollary 1. *Under the condition and notations in Theorem 1, we have*

$$
\begin{aligned}
&E_\rho^m E_\mu^n(\mathcal{R}(f_{\mathbf{z}},\lambda)-\mathcal{R}(f_c)) \\
&\quad\leq \sqrt{\frac{Cn\log(m/\lambda)}{m}} + 2E_\mu^n \inf_{f\in\mathcal{M}_n}\sqrt{\left(\int_X (f(x)-f_\rho(x))^2\, d\rho + \lambda\|f\|_{\ell_2}^2\right)}.
\end{aligned}
$$

The learning rate $O(\sqrt{n\log m/m})$ can be achieved when the optional λ is selected and $\inf_{f\in\mathcal{M}_n}\sqrt{(\int_X (f(x)-f_\rho(x))^2\, d\rho + \lambda\|f\|_{\ell_2}^2)} \leq Cn\log m/m$. This learning rate is the same with the regularized ELM based on the i.i.d samples. To the best of our knowledge, this is the first touch on the convergence rate for the ELM-based classification with dependent samples.

4. Empirical evaluations

To better verify the theoretical analysis of ELM with the Markov chains, we evaluate its performance on some datasets. Here, we generate the u.e.M.c samples by the sampling algorithm in Table 1. In fact, this sampling algorithm has been used for learning algorithms in [27,21,22].

The UCI datasets are used to evaluate ELM and their characteristics are summarized in Table 2. The experiment can be divided into three steps: firstly, the training set Z' with m samples is

Table 2
Specifications of datasets.

Datasets	Attributes	Training size	Testing size
Waveform	21	2500	2500
Abalone	7	3133	1044
Magic	10	9510	9510
Letter	16	15,000	5000
Shuttle	9	10,000	4500

Table 3
Misclassification rate (MR) for 1000 training samples.

Datasets	MR(i.i.d.)	MR(Markov)
Waveform	0.1410 ± 0.0076	0.1364 ± 0.0062
Abalone	0.2068 ± 0.0041	0.2087 ± 0.0030
Magic	0.1956 ± 0.0069	0.1946 ± 0.0053
Letter	0.1791 ± 0.0078	0.1779 ± 0.0072
Shuttle	0.0141 ± 0.0193	0.0114 ± 0.0028

Table 4
Misclassification rate (MR) for 1500 training samples.

Datasets	MR(i.i.d.)	MR(Markov)
Waveform	0.1184 ± 0.0042	0.1164 ± 0.0029
Abalone	0.2016 ± 0.0038	0.2057 ± 0.0029
Magic	0.1842 ± 0.0282	0.1782 ± 0.0048
Letter	0.1658 ± 0.0057	0.1587 ± 0.0034
Shuttle	0.0096 ± 0.0037	0.0082 ± 0.0018

Table 1
Markov sampling algorithm.

Step 1:	Draw the training samples $\mathbf{z}=\{(x_i,y_i), i=1,2,\ldots,N_1\}$ randomly from dataset $D:=\{(x_i,y_i)\|x_i\in\mathbb{R}^d, y_i\in\{-1,1\}\}$. Use ELM to train the N_1 samples and obtain a predictor $\widehat{f}$. Set $m_+=0$, $m_-=0$ and $k=0$.
Step 2:	Draw a sample from D randomly, and denote it as the current sample z_t. Let $m_+=m_++1$ if the label of z_t is 1 and $m_-=m_-+1$ otherwise.
Step 3:	Draw another sample from D randomly, and denote it as the candidate sample z_*.
Step 4:	Calculate the ratio γ of $e^{-\ell(\widehat{f},z)}$ at the candidate sample z_* and the current sample z_t, $\gamma=\dfrac{e^{-\ell(\widehat{f},z_*)}}{e^{-\ell(\widehat{f},z_t)}}$.
Step5:	If $\gamma\geq P$, accept the candidate sample z_* with probability γ. If there are k candidate samples z_* cannot be accepted, accept the kth candidate sample with probability γ. Then set $z_{t+1}=z_*$, $m_+=m_++1$ if the label of z_{t+1} is 1 and $m_-=m_-+1$ otherwise.
Step 6:	If $m_+<m/2$ or $m_-<m/2$, return to Step 3, else stop it.

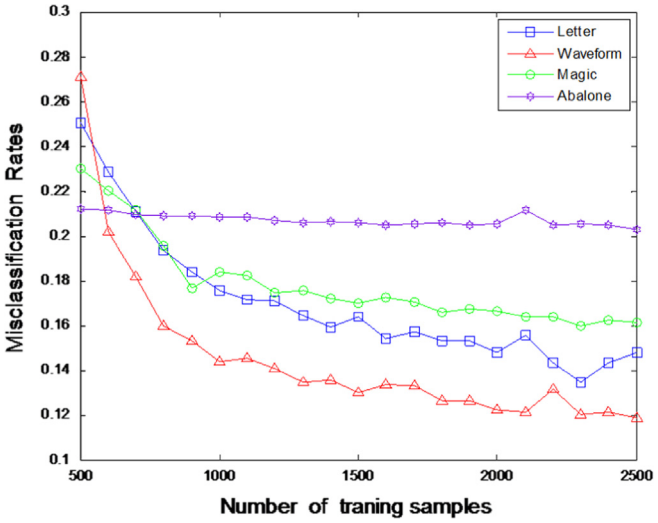

Fig. 1. Misclassification rates for Waveform, Abalone, Magic, and Letter with different training samples.

generated by the Markov sampling algorithm in [21,22]; secondly, we consider the combination of the square function and Gaussian function as the activation function of ELM [15]; finally, we train ELM on Z' and evaluate its performance on the test set.

We conduct the experiment for 50 times and the average misclassification rates are presented in Tables 3 and 4. The results tell us that ELM with Markov sampling can provide the competitive prediction according to the misclassification rates and standard deviations. We also evaluate the ELM with Markov sampling for different numbers of training samples in Fig. 1, which shown that the misclassification rate will decrease with the increasing training samples. This empirical result is consistent with the theoretical analysis in Theorem 1.

In order to better understand the efficiency of ELM, we also present several experiments to compare the standard deviations with the independent and Markov sampling methods. Figs. 2, 3, 4, and 5 report these experimental results for different training numbers on Waveform, Abalone, Magic, and Shuttle datasets. From these figures, we can find that the standard deviations of ELM with Markov sampling are usually smaller than ELM with i.i.d samples. That is to say, the Markov sampling usually can improve the stability of ELM with i.i.d samples.

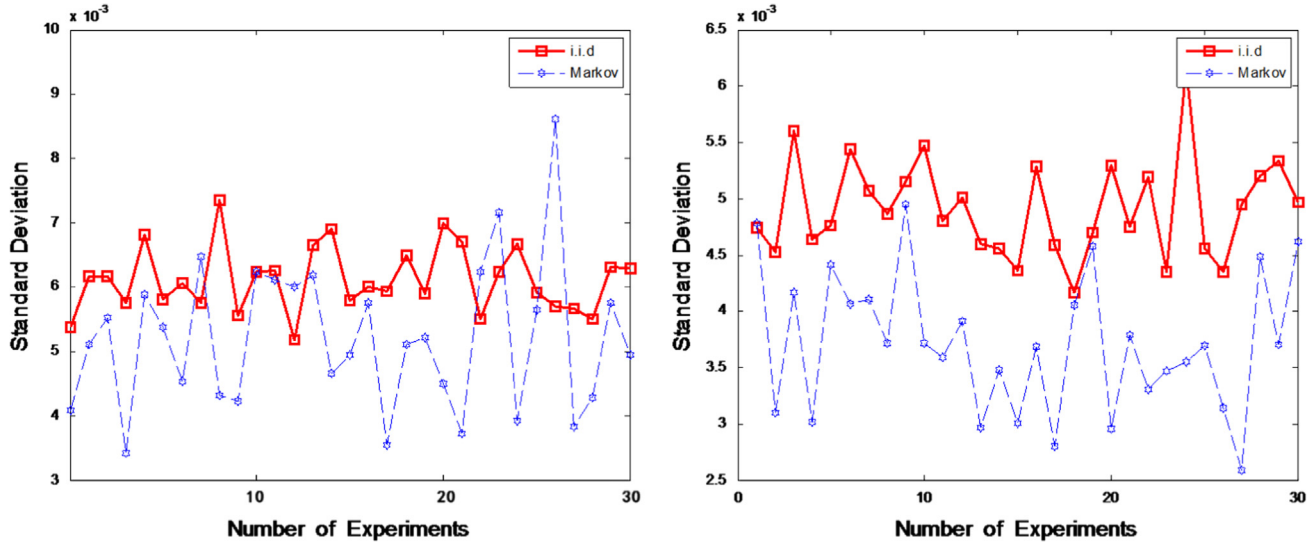

Fig. 2. Standard deviations for Waveform with $m = 1000$ (left) and $m = 1500$ (right).

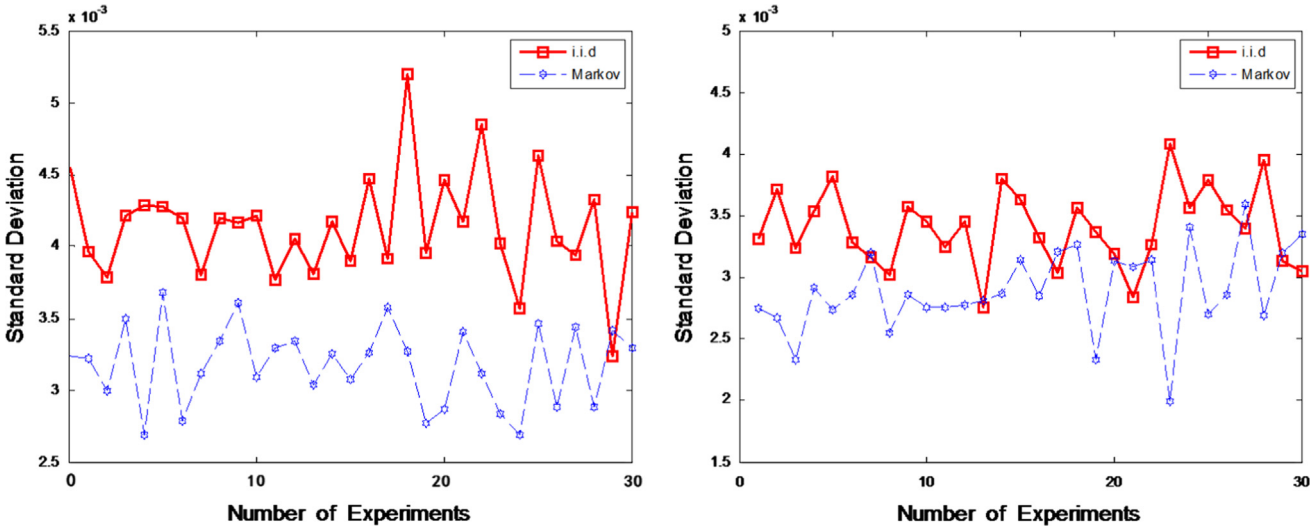

Fig. 3. Standard deviations for Abalone with $m = 1000$ (left) and $m = 1500$ (right).

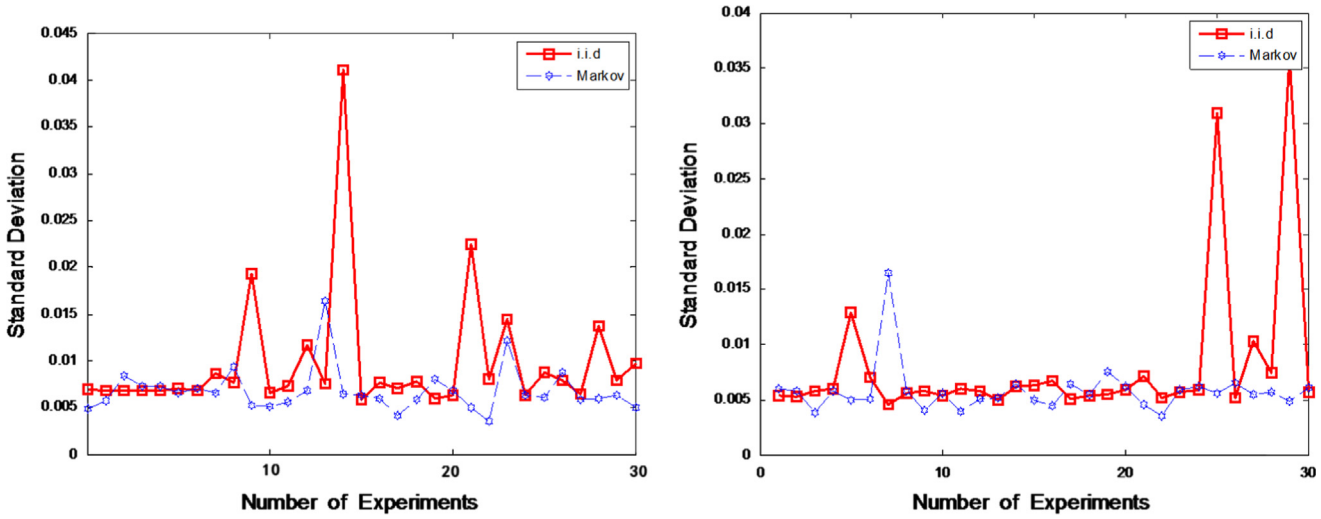

Fig. 4. Standard deviations for Magic with $m=1000$ (left) and $m=1500$ (right).

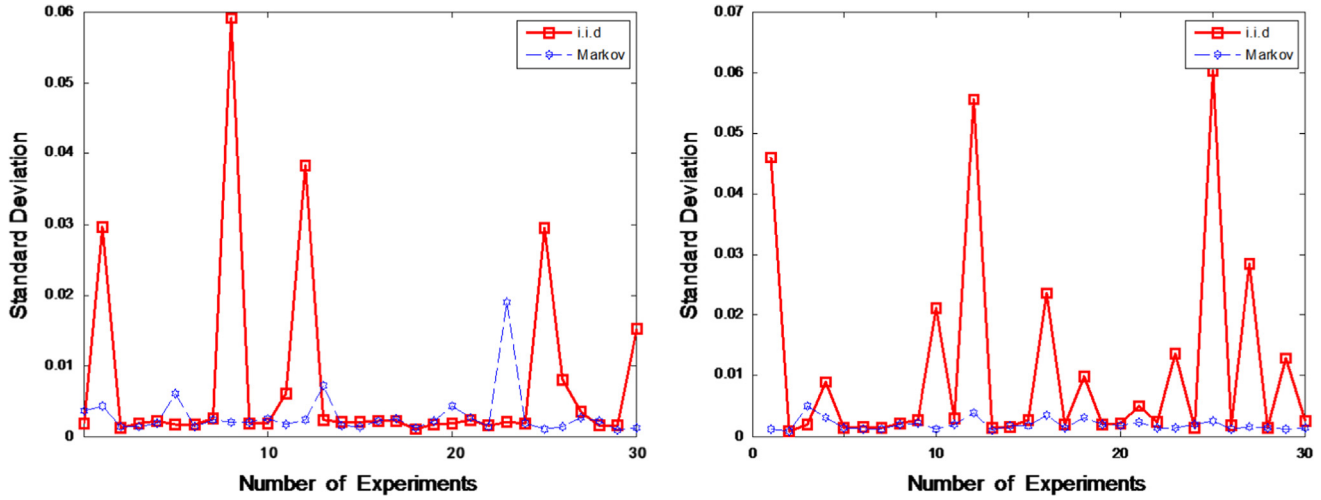

Fig. 5. Standard deviations for Shuttle with $m=1000$ (left) and $m=1500$ (right).

5. Conclusion

In this paper, we have investigated the generalization ability of ELM with the Markov chain samples. The generalization bound of ELM has been established and some empirical evaluations have been provided. In particular, the learning rate derived here is the same as the previous work for the i.i.d samples. Along the line of the present work, some subjects deserve to further study, e.g., the generalization bounds of semi-supervised and unsupervised ELMs in [4] and the generalization analysis for the sparse ELMs in [1].

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities (Program no. 2014PY025, 2662015PY046), the National Natural Science Foundation of China (61370002, 61403132), the Macau Science and Technology Development Fund (FDCT) under Grant 106/2013/A3, and by the Research Committee at University of Macau under Grants MYRG2014-00003-FST, MRG017/ZYC/2014/FST, MYRG113(Y1-L3)-FST12-ZYC and MRG001/ZYC/2013/FST.

Appendix

Proof of Lemma 3. Firstly, denote $J=\mathcal{N}(\mathcal{G},\epsilon)$ and consider $\{D_j\}_{j=1}^{J}$ as a cover of $\mathcal{G}$. Here, the balls $D_j=\{g\in\mathcal{G}:\|g-g_j\|_\infty\le\epsilon\}$. Then, for any $g\in\mathcal{G}$, there is g_j such that $\|g-g_j\|\le\epsilon$. Therefore, we have

$$|E(g)-E(g_j)|\le\|g-g_j\|\le\epsilon$$

and

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

Then, there exists

$$|E(g)-E(g_j)|/\sqrt{E(g)+\epsilon}\le\epsilon$$

and

$$\left|\frac{1}{m}\sum_{i=1}^{m}g(z_i)-\frac{1}{m}\sum_{i=1}^{m}g_j(z_i)\right|\Big/\sqrt{E(g)+\epsilon}\le\epsilon.$$

Secondly, from Lemma 2, for any $\epsilon > 0$, we have

$$Prob\left\{\frac{E(g)-\frac{1}{m}\sum_{i=1}^{m} g(z_i)}{\sqrt{E(g)+\epsilon}} \geq \sqrt{\epsilon}\right\} \leq \exp\left\{\frac{-m\epsilon}{56C\|\Gamma\|^2}\right\}. \quad (13)$$

Then for any $g_j \in \mathcal{G}$, we have

$$Prob\left\{\frac{E(g_j)-\frac{1}{m}\sum_{i=1}^{m} g_j(z_i)}{\sqrt{E(g_j)+\epsilon}} \geq \sqrt{\epsilon}\right\} \leq \exp\left\{\frac{-m\epsilon}{56C\|\Gamma\|^2}\right\}.$$

Lastly, for $|E(g)-E(g_j)| \leq \epsilon$, we have $\sqrt{E(g_j)+\epsilon} \leq 2\sqrt{E(g)+\epsilon}$. Therefore, there holds

$$\begin{aligned} &Prob\left\{\sup_{g\in\mathcal{G}}\frac{E(g)-\frac{1}{m}\sum_{i=1}^{m} g(z_i)}{\sqrt{E(g)+\epsilon}} \geq 4\sqrt{\epsilon}\right\} \\ &\leq \sum_{j=1}^{J} Prob\left\{\frac{E(g_j)-\frac{1}{m}\sum_{i=1}^{m} g_j(z_i)}{\sqrt{E(g_j)+\epsilon}} \geq \sqrt{\epsilon}\right\} \\ &\leq \mathcal{N}(\mathcal{G},\epsilon)\exp\left\{\frac{-m\epsilon}{56C\|\Gamma\|^2}\right\}. \end{aligned}$$

References

[1] Z. Bai, G.-B. Huang, D. Wang, H. Wang, M.B. Westover, Sparse extreme learning machine for classification, IEEE Trans. Cybern. (2015), in press.
[2] H. Chen, J.T. Peng, Y. Zhou, L.Q. Li, Z.B. Pan, Extreme learning machine for ranking: generalization analysis and applications, Neural Netw. 53 (2014) 119–126.
[3] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Am. Math. Soc. 39 (2000) 1–49.
[4] G. Huang, S. Song, J.N.D. Gupta, C. Wu, Semi-supervised and unsupervised extreme learning machines, IEEE Trans. Cybern. (2015), in press.
[5] G. Huang, G.-B. Huang, S. Song, K. You, Trends in extreme learning machines: a review, Neural Netw. 61 (2015) 32–48.
[6] G.-B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1) (2006) 489–501.
[7] G.-B. Huang, L. Chen, C.-K. Siew, Universal approximation using incremental constructive feedforward networks with random hidden nodes, IEEE Trans. Neural Netw. 17 (4) (2006) 879–892.
[8] G.-B. Huang, L. Chen, Convex incremental extreme learning machine, Neurocomputing 70 (2007) 3056–3062.
[9] G.-B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man Cybern. Part B Cybern. 42 (2) (2012) 513–529.
[10] X. Liu, S.B. Lin, Z.B. Xu, Is extreme learning machine feasible? A theoretical assessment (part I), IEEE Trans. Neural Netw. Learn. Syst. 26 (1) (2015) 7–20.
[11] M. Luo, K. Zhang, A hybrid approach combining extreme learning machine and sparse representation for image classification, Eng. Appl. Artif. Intell. 27 (1) (2014) 228–235.
[12] S.P. Meyn, R.L. Tweedie, Markov Chains and Stochastic Stability, Cambridge University Press, New York, 2009.
[13] C.L.P. Chen, A rapid supervised learning neural network for function interpolation and approximation, IEEE Trans. Neural Netw. 7 (5) (1993) 1220–1230.
[14] C.L.P. Chen, J.Z. Wan, A rapid learning and dynamic stepwise updating algorithm for flat neural networks and the application to time-series prediction, IEEE Trans. Syst. Man Cybern. Part B Cybern. 29 (1) (1999) 62–72.
[15] J.T. Peng, L.Q. Li, Y.Y. Tang, Combination of activation functions in extreme learning machines for multivariate calibration, Chemom. Intell. Lab. Syst. 120 (2013) 53–58.
[16] I. Steinwart, D. Hush, C. Scovel, Learning from dependent observations, J. Multivar. Anal. 100 (1) (2009) 175–194.
[17] S. Smale, D.X. Zhou, Online learning with Markov sampling, Anal. Appl. 7 (1) (2009) 87–113.
[18] P.M. Samson, Concentration of measure inequalities for Markov chains and Φ mixing processes, Ann. Probab. 28 (1) (2000) 416–461.
[19] S.B. Lin, X. Liu, J. Fang, Z.B. Xu, Is extreme learning machine feasible? A theoretical assessment (Part II), IEEE Trans. Neural Netw. Learn. Syst. 26 (1) (2015) 21–34.
[20] M. Vidyasagar, Learning and Generalization: With Applications to Neural Networks, Springer, London, 2003.
[21] J. Xu, Y.Y. Tang, B. Zou, Z.B. Xu, L.Q. Li, Y. Lu, Generalization performance of Gaussian kernels SVMC based on Markov sampling, Neural Netw. 53 (2014) 41–51.
[22] J. Xu, Y.Y. Tang, B. Zou, Z.B. Xu, L.Q. Li, Y. Lu, B.C. Zhang, The generalization ability of SVM classification based on Markov sampling, IEEE Trans. Cybern. (2014), http://dx.doi.org/10.1109/TCYB.2014.2346536.
[23] Kai Zhang, Minxia Luo, Outlier-robust extreme learning machine for regression problems, Neurocomputing 151 (2015) 1519–1527.
[24] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, Ann. Stat. 32 (2004) 56–85.
[25] R. Zhang, Y. Lan, G.-B. Huang, Z.B. Xu, Universal approximation of extreme learning machine with adaptive growth of hidden nodes, IEEE Trans. Neural Netw. Learn. Syst. 23 (2) (2012) 365–371.
[26] B. Zou, L.Q. Li, Z.B. Xu, The generalization performance of ERM algorithm with strongly mixing observations, Mach. Learn. 75 (2009) 275–295.
[27] B. Zou, Y.Y. Tang, Z.B. Xu, L.Q. Li, J. Xu, Y. Lu, The generalization performance of regularized regression algorithms based on Markov sampling, IEEE Trans. Cybern. 44 (9) (2014) 1497–1507.
[28] B. Zou, L.Q. Li, Z.B. Xu, T. Luo, Y.Y. Tang, Generalization performance of Fisher linear discriminant based on Markov sampling, IEEE Trans. Neural Netw. Learn. Syst. 24 (2) (2013) 288–300.

Peipei Yuan received the B.Sc. degree in information and computation science from Huazhong Agricultural University in 2014. She is currently pursuing the master's degree in College of Engineering, Huazhong Agricultural University, Wuhan, China. Her research interests are machine learning and data mining.

Hong Chen received the B.Sc. degree and the Ph.D. degree from Hubei University, Wuhan, China, in 2003 and 2009, respectively. He is an Associate Professor with Department of Mathematics and Statistical Sciences, College of Science, Huazhong Agricultural University, China. His research interests include statistical learning theory, approximation theory, and machine learning.

Yicong Zhou received his B.S. degree from Hunan University, Changsha, China, and his M.S. and Ph.D. degrees from Tufts University, Massachusetts, USA, all degrees in electrical engineering. He is currently an Assistant Professor in the Department of Computer and Information Science at University of Macau, Macau, China. His ongoing research interests focus on multimedia security, image/signal processing, medical imaging and object recognition.

Xiaoyan Deng received the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2006. He is a Professor with Department of Mathematics and Statistical Sciences, College of Science, Huazhong Agricultural University, Wuhan, China. His research interests include statistical learning, signal processing, and computation intelligence.

Bin Zou received the Ph.D. degree from Hubei University, China, in 2007. From 2008 to 2009, he was a Postdoctoral Research Fellow at the Institute for Information and System Science, Xi'an Jiaotong University, Xi'an, China. He is currently with the Key Laboratory of Applied Mathematics, Hubei Province, and the Faculty of Mathematics and Computer Science, Hubei University, where he became a Professor, in 2014. His research interests include statistical learning theory, machine learning, and pattern recognition.