Generalization Performance of Fisher Linear Discriminant Based on Markov Sampling

Bin Zou, Luoqing Li, Zongben Xu, Tao Luo, and Yuan Yan Tang, *Fellow, IEEE*

***Abstract*— Fisher linear discriminant (FLD) is a well-known method for dimensionality reduction and classification that projects high-dimensional data onto a low-dimensional space where the data achieves maximum class separability. The previous works describing the generalization ability of FLD have usually been based on the assumption of independent and identically distributed (i.i.d.) samples. In this paper, we go far beyond this classical framework by studying the generalization ability of FLD based on Markov sampling. We first establish the bounds on the generalization performance of FLD based on uniformly ergodic Markov chain (u.e.M.c.) samples, and prove that FLD based on u.e.M.c. samples is consistent. By following the enlightening idea from Markov chain Monto Carlo methods, we also introduce a Markov sampling algorithm for FLD to generate u.e.M.c. samples from a given data of finite size. Through simulation studies and numerical studies on benchmark repository using FLD, we find that FLD based on u.e.M.c. samples generated by Markov sampling can provide smaller misclassification rates compared to i.i.d. samples.**

***Index Terms*— Fisher linear discriminant (FLD), generalization performance, Markov sampling, uniformly ergodic Markov chain.**

I. Introduction

In THE past, most works describing the generalization ability of learning algorithms have been based on the assumption of independent and identically distributed (i.i.d.) samples. However, independence is a very restrictive concept in several ways (see [1] and [2]). First, it is often an assumption, rather than a deduction, based on 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 [2]. 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 [1]. Therefore, relaxations of such i.i.d. assumption have been considered for quite a while in both the machine learning and statistics literature. For example, Yu [3] established the rates of convergence for empirical processes of stationary mixing sequences. Modha and Masry [4] studied the minimum complexity regression estimation with m dependent observations and strongly mixing observations, respectively. Lozano *et al.* [5] proved that regularized boosting based on β-mixing processes are consistent. Vidyasagar [2] considered the notions of mixing and proved that most of the desirable properties (e.g., probably approximately correct, uniform convergence of empirical means uniformly in probability) of i.i.d. sequences are preserved when the underlying sequence is a mixing sequence. Kontorovich and Ramanan [6] established the concentration inequalities for dependent random variables via the martingale method. Mohri and Rostamizdeh [7] studied Rademarcher complexity bounds for non-i.i.d. processes. Steinwart and Christmann [8] considered the fast learning rates of regularized empirical risk minimizing algorithm for α-mixing process. Smale and Zhou [9] considered an online learning algorithm based on Markov sampling. Steinwart *et al.* [1] proved that the support vector machine for both classification and regression are consistent only if the data-generating process satisfies a certain type of law of large numbers (e.g., weak law of large numbers for events, strong law of large numbers for events). Sun and Wu [10] studied the regularized least-squares regression with dependent samples. Zou *et al.* [11] established the bounds on the generalization performance of the empirical risk minimization (ERM) algorithm with strongly mixing observations.

Fisher linear discriminant (FLD) is a well-known method for dimensionality reduction and linear classification, which has been studied for a long time under different cases (see [12], [16]–[19]). For example, Friedman [13] proposed regularized discriminant analysis. Hastie *et al.* [14], [15] proposed mixture discriminant analysis (MDA) and flexible discriminant analysis (FDA). More recently, Hou *et al.* [20] studied the complexity-reduced scheme for feature extraction with linear discriminant analysis. Unlike these works, in this paper, we study the generalization ability of the FLD based on dependent samples. There have been many dependent (non-i.i.d.) sampling mechanisms (e.g., α-mixing, β-mixing) studied in machine learning literature (see [1]–[4], [11]–[22]).

Manuscript received November 9, 2011; revised November 23, 2012; accepted November 23, 2012. Date of publication December 24, 2012; date of current version January 11, 2012. This work was supported in part by the NSFC under Grant 11071058 and Grant 61070225, multi-year research of University of Macau under Grant MYRG205(Y1-L4)-FST11-TYY and Grant MYRG187(Y1-L3)-FST11-TYY, and start-up research of the University of Macau under Grant SRG010-FST11-TYY.

B. Zou and L. Li are with the Faculty of Mathematics and Computer Science, Hubei University, Wuhan 430062, China (e-mail: zoubin0502, lilq@hubu.edu.cn).

Z. Xu and T. Luo are with the Institute for Information and System Science, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zbxu@mail.xjtu.edu.cn; luotaotao@stu.xjtu.edu.cn).

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

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.2012.2230406

In this paper, we focus only on an analysis in the case when the training samples of the FLD are Markov chains, the reasons for which are as follows. First, in real-world problems, Markov chain samples appear so often and naturally in applications such as biological (DNA or protein) sequence analysis, content-based web search, market prediction, and so on. See [22] for examples of learning from Markov chain samples. In addition, when the size of the dataset is very large, learning is very time consuming, and we usually sample randomly a part of samples from the dataset of large size and learn from the part samples. Then a problem is posed: how to sample a part of samples from the large dataset such that FLD has good generalization performance. For these purposes, in this paper we study the generalization ability of FLD based on uniformly ergodic Markov chain (u.e.M.c.) samples and introduce a Markov sampling algorithm for FLD (see Algorithm 1) to generate u.e.M.c. samples from a given dataset of finite size by following the enlightening idea from Markov chain Monto Carlo (MCMC) methods [23], [24]. Through simulation studies and numerical studies on benchmark repository using the FLD method, we find that the FLD algorithm based on Markov sampling introduced in this paper can provide smaller misclassification rates compared to i.i.d. sampling from the same dataset, in particular for a large dataset. This implies that Markov sampling from a given dataset of finite large size can be considered to be a new method of manipulating the training samples [25] such that the learning performance of FLD method is improved.

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 results on the generalization performance of FLD based on u.e.M.c. samples. In Section IV, we introduce a Markov sampling algorithm to generate Markov chain samples and present the simulation and numerical studies on benchmark repository of the FLD method. Finally, we conclude this paper in Section V.

II. Preliminaries

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

A. Uniformly Ergodic Markov Chains

Suppose $(\mathcal{Z}, \mathcal{S})$ is a measurable space; a Markov chain is a sequence of random variables $\{Z_t\}_{t\geq 1}$ together with a set of transition probability measures $P^n(z_{n+i}|z_i)$, $z_{n+i}, z_i \in \mathcal{Z}$. It is assumed that

$$P^n(z_{n+i}|z_i) \doteq P\left\{Z_{n+i} = z_{n+i}|Z_j, j < i, Z_i = z_i\right\}.$$

Then $P^n(z_{n+i}|z_i)$ denotes the probability that the state z_{n+i}, after n time steps, starting from the initial state z_i at time i. It is common to denote the one-step transition probability by

$$P^1(z_{i+1}|z_i) \doteq P\left\{Z_{i+1} = z_{i+1}|Z_j, j < i, Z_i = z_i\right\}$$

so that $P^1(z_{i+1}|z_i) = P(z_{i+1}|z_i)$. The fact that the transition probability does not depend on the values of Z_j prior to time i is the Markov property

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

This is commonly expressed in words as "given the present state, the future and past states are independent."

Given two probabilities ν_1, ν_2 in the space $(\mathcal{Z}, \mathcal{S})$, we define the total variation distance between the two measures ν_1, ν_2 as $||\nu_1 - \nu_2||_{TV} \doteq \sup_{A\in\mathcal{S}} |\nu_1(A) - \nu_2(A)|$. Thus we have the following definition of u.e.M.c. [21], [26], [27].

Definition 1: A Markov chain $\{Z_t\}_{t\geq 1}$ is said to be uniformly ergodic if there exist constants $\gamma < \infty$ and $0 < \rho_1 < 1$ such that for any $z \in \mathcal{Z}$, and any $n \geq 1$

$$||P^n(\cdot|z) - \pi(\cdot)||_{TV} \leq \gamma\rho_1^n$$

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

Remark 1: A weaker condition than uniformly ergodic is V-geometrically ergodic (see Definition 2 of Appendix A). The difference between V-geometrically ergodic and uniformly ergodic is that here the total variation distance between the n-step transition probability $P^n(\cdot|z)$ and the invariant measure $\pi(\cdot)$ approaches zero at a geometric rate multiplied by $V(z)$ (see [2], [21]). Thus the rate of geometric convergence is independent of z, but the multiplicative constant is allowed to depend on z. Especially, if the space $\mathcal{Z}$ is finite, then all irreducible and aperiodic Markov chains are V-geometrically (in fact, uniformly) ergodic. And a Markov chain is V-geometrically ergodic if the condition that $V(\cdot)$ has finite expectation with respect to the invariant measure π holds.

B. Fisher Linear Discriminant (FLD)

FLD is a well-known method for dimensionality reduction and classification that projects high-dimensional data onto a low-dimensional space where the data achieves maximum class separability (see [28]–[31]). FLD gives a projection matrix $\mathbf{w}$ that reshapes the scatter of a dataset D to maximize class separability, which is defined as the ratio of the between-class scatter matrix to the within-class scatter matrix. That is, let $\{\mathbf{x}_i\}_{i=1}^N$ be a set of N column vectors of dimension h. The mean of the data set D is defined as $\mu = 1/N\sum_{i=1}^N \mathbf{x}_i$. There are k classes $\{C_1, C_2, \ldots, C_k\}$. The mean of the ith class $(1 \leq i \leq k)$ containing N_i members is $\mu_i = 1/N_i \sum_{\mathbf{x}\in C_i} \mathbf{x}$. The between-class scatter matrix is $S_b = \sum_{i=1}^k N_i(\mu_i - \mu)(\mu_i - \mu)^T$. The within-class scatter matrix is defined as $S_w = \sum_{i=1}^k \sum_{\mathbf{x}\in C_i}(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T$. The mixture scatter matrix is the covariance matrix of all samples, regardless of their class assignments, and it is given by $S_m = \sum_{i=1}^k (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T = S_w + S_b$. The purpose of FLD method is to consider maximizing the quantity

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}. \tag{1}$$

However, in practice, the small sample size (SSS) problem is often encountered if S_w in (1) is singular [32]. Therefore, the maximization problem of (1) can be difficult to solve. In order to overcome this problem, the term ϵI is added, where ϵ is a small positive number and I the identity matrix of proper size. That is, for the case of SSS problem, one is to maximize the following quantity:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T (\epsilon I + S_w)\mathbf{w}} \tag{2}$$

which can be solved without any numerical problems [33].

III. Bounds of Generalization Ability

In this section, we estimate the bounds on the generalization performance of FLD based on u.e.M.c. samples by following the enlightening idea of [32]. In [32], Zhang and Riedel established the connection between the solution of FLD and the solution of empirical risk minimization with the least-squares loss function. Namely, for given a set $\mathbf{z} = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^m$ of m training examples drawn from the probability space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Here, the probability measure ρ is defined and unknown, and $\mathbf{x}_i$ are the h-dimensional inputs. In this paper, we mainly consider two-class problems, that is, $k = 2$; then we have $\mathcal{Y} = \{-1, 1\}$. For simplicity, in this paper we assume $\mathbf{x}$ has zero mean, i.e., for any $\mathbf{x} \in \mathcal{X}$, $\mathrm{E}(\mathbf{x}) = 0$.

We define f_ρ as the best function that minimizes the least-squares expected error over all possible measure functions

$$f_\rho = \arg\min_f \mathcal{E}(f) = \arg\min_f \int_{\mathcal{Z}} (y - f(\mathbf{x}))^2 d\rho. \quad (3)$$

Since the probability measure ρ is unknown and so is f_ρ, the minimizer of the expected error (3) cannot be computed directly. According to the ERM principle [38], we then consider function $f_\mathbf{z}$ as an approximation of the target function f_ρ

$$f_\mathbf{z} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_m(f) = \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m (y_i - f(\mathbf{x}_i))^2 \quad (4)$$

where $\mathcal{H}$ is a hypothesis space. In this paper, we assume that $\mathcal{H}$ is linear functions set

$$\mathcal{H} = \left\{ f \mid f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}, ||\mathbf{w}||_2 \le a, ||\mathbf{x}||_2 \le b \right\} \quad (5)$$

where $\mathbf{x} \in \mathcal{X}^h$. Then we can rewrite (4) as

$$\mathbf{w}_\mathbf{z} = \arg\min_\mathbf{w} \frac{1}{m}\sum_{i=1}^m \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2. \quad (6)$$

Zhang and Riedel [32] proved that the solution of (6) is the same as the solution of (1).

Lemma 1: The linear system derived by the least-squares criterion (6) is equivalent to the one derived by the Fisher's criterion (1), up to a constant, in two-class problems.

Remark 2: The relationship between least-squares regression and FLD has been well known for a long time. There are good reviews in [33]–[35]. Fisher already pointed out its connection to the regression solution in [36].

Therefore, the central question of estimating the generalization ability of FLD based on u.e.M.c. samples is how well $f_\mathbf{z}$ really approximate f_ρ. In other words, one tries to learn the function $f_\mathbf{z}$ that is as close as possible to the optimal function f_ρ for the sample set $\mathbf{z}$. For this reason, we are to estimate the excess error $\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)$. If $f_\rho \in \mathcal{H}$, simplifications will occur. But in general, we will not even assume that $f_\rho \in \mathcal{C}(\mathcal{X})$ [37], $\mathcal{C}(\mathcal{X})$ is the Banach space of continuous functions on $\mathcal{X}$ with the norm $||f||_\infty = \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|$. Then we will have to consider another target function $f_\mathcal{H}$ in $\mathcal{H}$: $f_\mathcal{H}$ is a function minimizing the error $\mathcal{E}(f)$ over $f \in \mathcal{H}$. Thus we have

$$\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho) = \left\{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\mathcal{H})\right\} + \left\{\mathcal{E}(f_\mathcal{H}) - \mathcal{E}(f_\rho)\right\}. \quad (7)$$

The first term of the right-hand side of (7) depends on the choice of $\mathcal{H}$ and the sample set $\mathbf{z}$. We call it the sample error. The second term depends on $\mathcal{H}$ and ρ but is independent of sampling. We call it the approximation error [37], which measures how well the functions in $\mathcal{H}$ can approach the target function f_ρ. Since ρ is not known, in this paper we focus only on the sample error. The approximation error for the least-squares loss function is well understood in [37]. Thus by Lemma 1 and (7), we establish the bound on the excess error of FLD (1) based on u.e.M.c. samples.

Theorem 1: Let $\{z_i\}_{i=1}^m$ be u.e.M.c. samples and $A_{\mathcal{H},\rho}(f) = \mathcal{E}(f_\mathcal{H}) - \mathcal{E}(f_\rho)$. Set

$$m^{(\beta)} = \left\lfloor m \left\lceil \left\{8m/\ln(1/\rho_1)\right\}^{\frac{1}{2}} \right\rceil^{-1} \right\rfloor$$

where $\lfloor u \rfloor (\lceil u \rceil)$ denotes the greatest (least) integer less (greater) than or equal to u. Then for any $0 < \eta < 1$, the inequality

$$\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho) \le \varepsilon(m,\eta) + \frac{(ab+1)}{2}\sqrt{\frac{\ln(C_1/\eta)}{\ln(1/\rho_1)^{\frac{1}{2}} m^{\frac{1}{2}}}} + A_{\mathcal{H},\rho}(f) \quad (8)$$

holds true with probability at least $1 - 2\eta$ provided that $m \ge \max\{m_1, m_2, m_3\}$, where C_0 is a constant independent of m or η, $m_1 = \left\{\ln(1/\rho_1)/8, 128/\ln(1/\rho_1)\right\}$, $C_1 = 1 + \gamma e^{-2}$

$$m_2 = \max\left\{ \frac{1}{6(ab+1)}\sqrt{\frac{\ln(C_1/\eta)}{\ln(1/\rho_1)^{\frac{1}{2}}}}, \frac{1}{6}\sqrt{\frac{2^{\frac{3}{2}}\ln(C_1/\eta)}{\ln(1/\rho_1)^{\frac{1}{2}}}} \right\}$$

$$m_3 = \frac{ab}{9(ab+1)}\sqrt{\frac{2^{\frac{3}{2}}C_0}{\ln(1/\rho_1)^{\frac{1}{2}}}}$$

$$m_4 = 2(ab+1)\left[\frac{8C_0a^2b^2(ab+1)^2}{m^{(\beta)}}\right]^{\frac{1}{4}}$$

and $\varepsilon(m,\eta) \le \max\left\{2(ab+1)^2\left[\frac{2\ln\left(C_1/\eta\right)}{m^{(\beta)}}\right]^{\frac{1}{2}}, m_4\right\}$.

Remark 3: To estimate the excess error of FLD (1) based on u.e.M.c. samples, in Theorem 1 we introduce the quantity $m^{(\beta)}$, which is called the "effective number of observations" for u.e.M.c. samples. By Theorem 1, we can find that $m^{(\beta)}$ plays the same role in our analysis as that played by the number m of observations in the i.i.d. case (see [38]–[40]). To our knowledge, this result here is the first work of FLD for u.e.M.c. samples in this topic.

For the proof of Theorem 1, refer to Appendix B. Since $A_{\mathcal{H},\rho}(f) = 0$, if $f_\rho \in \mathcal{H}$, by Theorem 1, and using the fact that $\lceil t \rceil \le 2t$ for all $t \ge 1$ and $\lfloor t \rfloor \ge t/2$ for all $t \ge 2$, we have the following bound on the learning rate of FLD (1) based on u.e.M.c. samples.

Corollary 1: Let $\{z_i\}_{i=1}^m$ be u.e.M.c. samples. If $f_\rho \in \mathcal{H}$; then for any $0 < \eta < 1$, the inequality

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le \frac{(ab+1)}{2}\sqrt{\frac{\ln(C_1/\eta)}{\ln(1/\rho_1)^{\frac{1}{2}}m^{\frac{1}{2}}}} + 4(ab+1)\left[\frac{2\sqrt{2}C_0a^2b^2(ab+1)^2}{m^{\frac{1}{2}}(\ln(1/\rho_1))^{\frac{1}{2}}}\right]^{\frac{1}{4}}$$

holds true with probability at least $1-2\eta$ provided that $m \ge \max\{m_1, m_2, m_3, m_4'\}$, where C_0, C_1, m_1, m_2, and m_3 are defined as in Theorem 1, and

$$m_4' = \frac{(\ln(C_1/\eta))^4}{132\ln(1/\rho_1)C_0^2[ab(ab+1)]^4}.$$

By Corollary 1, we can conclude that $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \to 0$ as $m \to \infty$ and $f_\rho \in \mathcal{H}$. This implies that, in this case, FLD (1) based on u.e.M.c. samples is consistent.

However, by the statistical learning theory [38], solving (4) often leads to overfitting data if the complexity of the hypothesis space is high, and, when the sample size is smaller than the dimensionality, it is an ill-posed problem and the solution is not unique. For these reasons, another purpose of this paper is to estimate the excess error of FLD (2) based on u.e.M.c. samples as follows.

Zhang and Riedel [32] proved that the solution of (2) is the same as the solution of the following least-squares regularization regression (see [41], [42]):

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m (y_i - f(\mathbf{x}_i))^2 + \lambda||f||_{\mathcal{H}}^2 \qquad (9)$$

where λ is a regularization parameter. $||f||_{\mathcal{H}}$ is a norm of the space $\mathcal{H}$. Since in this paper we assume that the set $\mathcal{H}$ is a linear functions space, then we can rewrite (9) as

$$\mathbf{w}_{\mathbf{z},\lambda} = \arg\min_{\mathbf{w}} \frac{1}{m}\sum_{i=1}^m (y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda\mathbf{w}^T\mathbf{w}. \qquad (10)$$

Zhang and Riedel [32] proved that the solution of (2) is the same as the solution $f_{\mathbf{z},\lambda}$ of (9). To establish the bound on the excess error of FLD (2) based on u.e.M.c. samples, we first introduce a regularizing function $\tilde{f}_\lambda \in \mathcal{H}$. This is arbitrarily chosen and depends on λ. A special and standard choice is

$$f_\lambda = \arg\min_{f\in\mathcal{H}} \left\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda||f||_{\mathcal{H}}^2\right\}.$$

By the definition of $f_{\mathbf{z},\lambda}$, for any $\tilde{f}_\lambda \in \mathcal{H}$, there holds $\mathcal{E}_m(f_{\mathbf{z},\lambda}) + \lambda||f_{\mathbf{z},\lambda}||_{\mathcal{H}}^2 \le \mathcal{E}_m(\tilde{f}_\lambda) + \lambda||\tilde{f}_\lambda||_{\mathcal{H}}^2$. Hence we have that, for any $\tilde{f}_\lambda \in \mathcal{H}$, and $f_{\mathbf{z},\lambda}$ be defined as (9)

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) &\le \mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) + \lambda||f_{\mathbf{z},\lambda}||_{\mathcal{H}} \\ &\le \left\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_m(f_{\mathbf{z},\lambda}) + \mathcal{E}_m(\tilde{f}_\lambda) - \mathcal{E}(\tilde{f}_\lambda)\right\} \\ &\quad + \left\{\mathcal{E}(\tilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda||\tilde{f}_\lambda||_{\mathcal{H}}\right\}. \qquad (11)\end{aligned}$$

In this way, we decompose the excess error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)$ into two parts: the sample error (the first term), and the regularization error (the second term) which is dependent on the choose of the space $\mathcal{H}$. We establish the following bound on excess error of FLD (2) based on u.e.M.c. samples.

Theorem 2: Let $\tilde{D}(\lambda) = \mathcal{E}(\tilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda||\tilde{f}_\lambda||_{\mathcal{H}}^2$ for any $\tilde{f}_\lambda \in \mathcal{H}$. Let $\{z_i\}_{i=1}^m$ be u.e.M.c. samples. Then for any $0 < \delta < 1$, the inequality

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \le \left(a\sqrt{\frac{\tilde{D}(\lambda)}{\lambda}} + 1\right)^2 \sqrt{\frac{2\ln\left[(1+\gamma e^{-2})/\delta\right]}{m^{(\beta)}}} + \frac{4(1+\lambda)^2}{\lambda^2}\cdot\left[\frac{C_0(1+\lambda)^2}{\lambda^2 m^{(\beta)}}\right]^{\frac{1}{4}} + \tilde{D}(\lambda)$$

holds true with probability at least $1-2\delta$ provided that $m^{(\beta)} \ge \max\{\tilde{m}, \hat{m}\}$, where $\hat{m} = 4^4(1+\lambda)^{10}/(3^4\lambda^6)[a\sqrt{\tilde{D}(\lambda)} + \sqrt{\lambda}]^8$, and

$$\tilde{m} = \left\{\frac{\ln\left[\left(1+\gamma e^{-2}\right)/\delta\right]^2\lambda^2}{C_0(1+\lambda)^2}, \frac{16C_0(1+\lambda)^2}{81\lambda^2}\right\}$$

where C_0 and $m^{(\beta)}$ are defined as in Theorem 1.

For the proof of Theorem 2, refer to Appendix C. In Theorem 2, we present the bound on the generalization ability of FLD (2) based on u.e.M.c. samples. In particular, if $f_\rho \in \mathcal{H}$, by Theorem 2, we can conclude that as $m \to \infty$, $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \to 0$. This implies that in this case FLD (2) based on u.e.M.c. samples is consistent. Different from the previously known results in [32], in this paper, we study the bounds on the excess error of FLD based on u.e.M.c. samples. In other words, in this paper we generalize this i.i.d. classical results of FLD to the case of u.e.M.c. samples.

IV. Markov Sampling and Numerical Studies

In this section, we introduce a Markov sampling algorithm such that we can generate u.e.M.c. from a given dataset. Then we give numerical studies on the learning performance of FLD method based on Markov sampling, and give some useful discussions.

A. Markov Sampling Algorithm

In this subsection, we introduce a Markov sampling algorithm for FLD to generate u.e.M.c. samples from a given dataset S_1 of finite size. Here $m\%2$ denotes the remainder of m divided by 2, m is the number of training samples.

We can summarize Algorithm 1 as first computing the acceptance probability (or transition probability) $\alpha = \min\{1, e^{-\mathcal{L}(f_0,z_*)}/e^{-\mathcal{L}(f_0,z_t)}\}$, $\mathcal{L}(f,z) = (f(\mathbf{x}) - y)^2$ and then accepting the candidate sample z_* with probability α. By Algorithm 1, we can generate a sequence $z_1, z_2, \ldots, z_p$, $p \in \mathbf{N}$. Since the size of dataset S_1 is finite, and the acceptance probability α is always positive, by the theory of Markov chains in [21], we can conclude that $\{z_1, \ldots, z_p\}$ is a u.e.M.c. sequence.

Remark 4: To define the transition probability α, we introduce two technical conditions: the preliminary learning model f_0, and the function $\mathcal{L}(f,z) = (f(\mathbf{x}) - y)^2$. This is because under the two technical conditions, we can compute

Algorithm 1 Markov Sampling for FLD

Step 1: Draw randomly N_1 ($N_1 \leq m$) samples $\{z_i, i = 1, \ldots, N_1\}$ from the data S_1. Use FLD to train these samples of size N_1, and obtain a preliminary learning model f_0. Set $m_+ = 0$ and $m_- = 0$.

Step 2: Draw randomly a sample from S_1 and denote it the current sample z_t. If $m\%2 = 0$, then set $m_+ = m_+ + 1$ if the label of z_t is $+1$. Set $m_- = m_- + 1$ if the label of z_t is -1.

Step 3: Draw randomly another sample from S_1 and denote it the candidate sample z_*.

Step 4: Calculate the ratio α of $e^{-\mathcal{L}(f_0,z)}$ at the sample z_* and the current sample z_t, $\alpha = e^{-\mathcal{L}(f_0,z_*)}/e^{-\mathcal{L}(f_0,z_t)}$ where $\mathcal{L}(f,z) = (f(\mathbf{x}) - y)^2$.

Step 5: If $\alpha \geq 1$, accept the sample z_* and set $z_{t+1} = z_*$, $m_+ = m_+ + 1$ if the label of z_t is $+1$, or set $m_- = m_- + 1$ if the label of z_t is -1. If $\alpha < 1$, with the probability α accept the sample z_* and set $z_{t+1} = z_*$, $m_+ = m_+ + 1$ if the label of z_t is $+1$, or set $m_- = m_- + 1$ if the label of z_t is -1.

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

easily the transition probability α and α is always positive. Different from MCMC algorithms, Algorithm 1 is a method of generating u.e.M.c. samples from a given data with finite size, and does not use the information of probability distribution of training samples (since the probability distribution of training samples is unknown). In addition, in order to generate the balance training samples, in Algorithm 1 we introduce the notations m_+ and m_-.

B. Simulation Datasets

We first conduct simulation study on the learning performance of FLD based on u.e.M.c. samples generated from a given data S_2 by Algorithm 1 and random sampling from the same data S_2, respectively.

The data S_2 was generated as follows: the input values $x_i, i = 1, 2, \ldots, 11$ were generated from normal distribution $N(0, 1)$ such that $x_{11} = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 + \xi$, where ξ was generated from normal distribution $N(0, \sigma)$, $\sigma = 1, 2, 3, 4$. The outputs were generated by $\text{sgn}(\xi)$, where $\text{sgn}(\xi)$ is defined as $\text{sgn}(\xi) = 1$ if $\xi \geq 0$ and $\text{sgn}(\xi) = -1$ if $\xi < 0$. Then, the data S_2 of size 10 000 was generated randomly, and a test samples set S_0 of size 300 was also generated separately according to identical input and output distributions.

For the case of random sampling, we decompose the experiment into two steps: First, a set S_T of m training samples was generated randomly from the data S_2. We use FLD to train the training samples in S_T, and then we test it on the test samples set S_0. Second, after all experiments had been repeated for 100 times, the misclassification rates of FLD are presented in Tables I and II, where "MR (i.i.d.)" denotes the misclassification rate of FLD based on random sampling.

TABLE I
MISCLASSIFICATION RATES FOR 300 TRAINING SAMPLES

σ	MR (i.i.d.)	MR (Markov)
1	0.0664 ± 0.0145	0.0416 ± 0.0128
2	0.0474 ± 0.0130	0.0211 ± 0.0092
3	0.0446 ± 0.0125	0.0295 ± 0.0103
4	0.0418 ± 0.0116	0.0266 ± 0.0096

TABLE II
MISCLASSIFICATION RATES FOR 500 TRAINING SAMPLES

σ	MR (i.i.d.)	MR (Markov)
1	0.0670 ± 0.0141	0.0358 ± 0.0115
2	0.0464 ± 0.0100	0.0388 ± 0.0074
3	0.0341 ± 0.0104	0.0225 ± 0.0075
4	0.0307 ± 0.0104	0.0234 ± 0.0082

For the case of Markov sampling, we first generate a set S'_T of m training samples by Algorithm 1. Then we use again FLD to train these Markov chain samples in S'_T, and test it on the same test set S_0. After all experiments had been repeated for 100 Markov chain sample sets, the misclassification rates are presented in Tables I and II, where "MR (Markov)" denotes the misclassification rate of FLD based on Markov sampling.

Remark 5: Tables I and II show that FLD based on Markov sampling can present obviously smaller misclassification rates compared to random sampling for both 300 and 500 training samples. In addition, the input samples $\mathbf{x}_i, i \geq 1$ of data S_2 are generated according to the normal distribution $N(0, 1)$. For other distributions such as uniform distribution, exponential distribution, and other size of the training samples, we can also obtain similar results as those presented in Tables I and II. Since FLD is a well-known method for dimensionality reduction and classification by projecting high-dimensional data onto a low-dimensional space where the data achieves maximum class separability, as the variance or noise of the training samples is larger, we can easily use the FLD method to project these samples onto a low-dimensional space and the samples have maximum class separability. Therefore, for the simulation dataset, the misclassification rates of FLD method become smaller as the variance or noise of the training samples gets bigger.

C. Benchmark Datasets

In this subsection, we give an extensive numerical studies on the learning performance of FLD based on Markov sampling for a benchmark repository. The benchmark repository consists of 13 real-world datasets from UCI-abalone, UCI-magic, UCI-pageblocks, UCI-shuttle, UCI-mushrooms, UCI-adult, UCI-gisette (see http://archive.ics.uci.edu/ml/datasets.html), DoubleUSPS(0,2), DoubleUSPS(3,8), DoubleUSPS(0,9) (see http://www.cs.nyu.edu/roweis/data.html), Waveform, Splice, and Image (see http://www.fml.tuebingen.mpg.de/Members/raetsch/benchmark). We present the information on these datasets in Table III. All these data in the benchmark repository

TABLE III
GENERAL INFORMATION ABOUT BENCHMARK DATASETS

Dataset	Training Size	Test Size	Input Dimension
UCI-abalone	2089	2088	8
UCI-magic	12 680	6340	10
UCI-pageblocks	3649	1824	10
UCI-shuttle	43 500	14 500	9
Waveform	4600	400	21
Splice	20 000	43 500	60
Image	26 000	20 200	18
UCI-mushrooms	8124	8124	112
UCI-adult	802	802	123
DoubleUSPS(0,2)	1100	1100	256
DoubleUSPS(3,8)	1100	1100	256
DoubleUSPS(0,9)	1100	1100	256
UCI-gisette	6000	6000	5000

TABLE IV
MISCLASSIFICATION RATES FOR 300 TRAINING SAMPLES

Dataset	MR (i.i.d.)	MR (Markov)
UCI-abalone	0.2326 ± 0.0050	0.2321 ± 0.0070
UCI-magic	0.2138 ± 0.0064	0.2127 ± 0.0069
UCI-pageblocks	0.0980 ± 0.0077	0.0868 ± 0.0080
UCI-shuttle	0.0692 ± 0.0081	0.0690 ± 0.0109
Waveform	0.1937 ± 0.0112	0.1654 ± 0.0170
Splice	0.1976 ± 0.0112	0.1921 ± 0.0082
Image	0.1759 ± 0.0133	0.1714 ± 0.0108
UCI-mushrooms	0.0060 ± 0.0045	0.0051 ± 0.0035
UCI-adult	0.2257 ± 0.0124	0.2219 ± 0.0115
DoubleUSPS(0,2)	0.1417 ± 0.0247	0.1376 ± 0.0260
DoubleUSPS(3,8)	0.1722 ± 0.0303	0.1694 ± 0.0285
DoubleUSPS(0,9)	0.1230 ± 0.0282	0.1179 ± 0.0249
UCI-gisette	0.0535 ± 0.0051	0.0529 ± 0.0115

are two-class real-world data except UCI-abalone and UCI-pageblocks. UCI-abalone and UCI-pageblocks are redefined as two classes as follows: the sample whose label is equal to or greater than 10 in UCI-abalone is viewed as a group and other samples are categorized as another group; the sample whose label is equal to or greater than 1 in UCI-pageblocks is viewed as a group and other samples are categorized as another group. After all experiments have been repeated 50 times, the misclassification rates of FLD for i.i.d. sampling and Markov sampling are presented in Tables IV and V.

Remark 6: Tables IV and V show that, for the same size of training samples and the same test samples set, the FLD method based on u.e.M.c. samples also has a smaller misclassification rate compared to i.i.d. samples.

In order to have a better understanding of learning performance of FLD based on Markov sampling, we also present the following figures on 50 times misclassification rates of FLD based on Markov sampling and random sampling, respectively. Here, the red curve denotes the results based on i.i.d. samples and the blue curve denotes the results based on Markov chain samples, and m is the number of training samples.

Remark 7: Figures 1–10 show that: 1) FLD based on Markov sampling will have obviously better learning performance than that of random sampling as the number of training samples is large, and 2) the number of samples and dimensionality of the datasets are important for the effective

TABLE V
MISCLASSIFICATION RATES FOR 500 TRAINING SAMPLES

Dataset	MR (i.i.d.)	MR (Markov)
UCI-abalone	0.2310 ± 0.0055	0.2255 ± 0.0053
UCI-magic	0.2112 ± 0.0056	0.2108 ± 0.0045
UCI-pageblocks	0.0953 ± 0.0116	0.0760 ± 0.0130
UCI-shuttle	0.0665 ± 0.0074	0.0654 ± 0.0062
Waveform	0.1900 ± 0.0105	0.1447 ± 0.0098
Splice	0.1805 ± 0.0075	0.1787 ± 0.0067
Image	0.1655 ± 0.0098	0.1613 ± 0.0087
UCI-mushrooms	0.0035 ± 0.0027	0.0029 ± 0.0026
UCI-adult	0.2103 ± 0.0101	0.2079 ± 0.0105
DoubleUSPS(0,2)	0.0477 ± 0.0080	0.0471 ± 0.0085
DoubleUSPS(3,8)	0.0248 ± 0.0060	0.0244 ± 0.0060
DoubleUSPS(0,9)	0.0199 ± 0.0064	0.0196 ± 0.0053
UCI-gisette	0.0459 ± 0.0031	0.0444 ± 0.0030

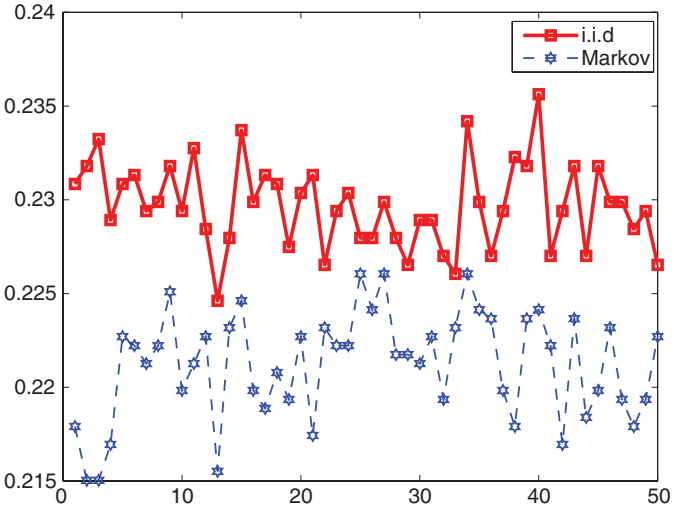

Fig. 1. UCI-abalone, $m = 1200$.

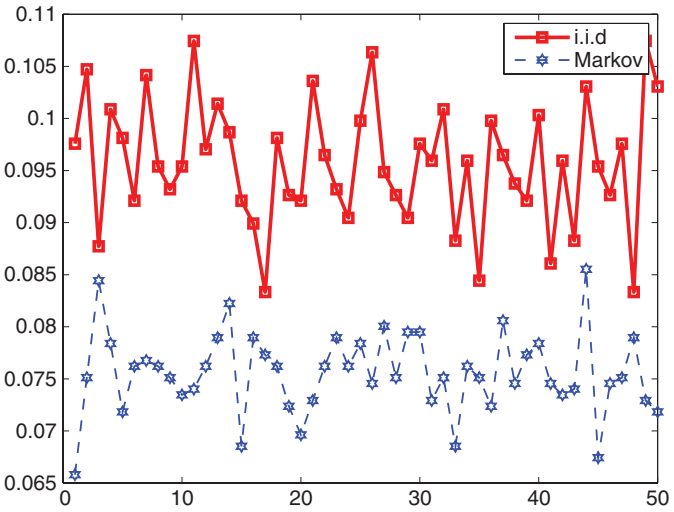

Fig. 2. UCI-pageblocks, $m = 700$.

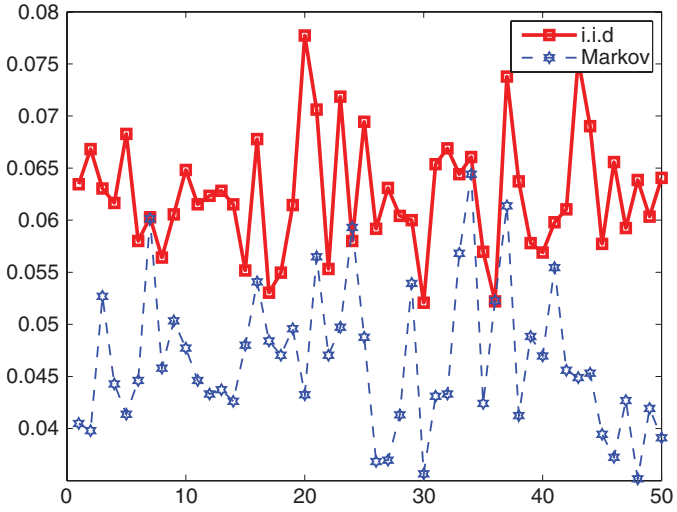

Fig. 3. UCI-shuttle, $m = 2500$.

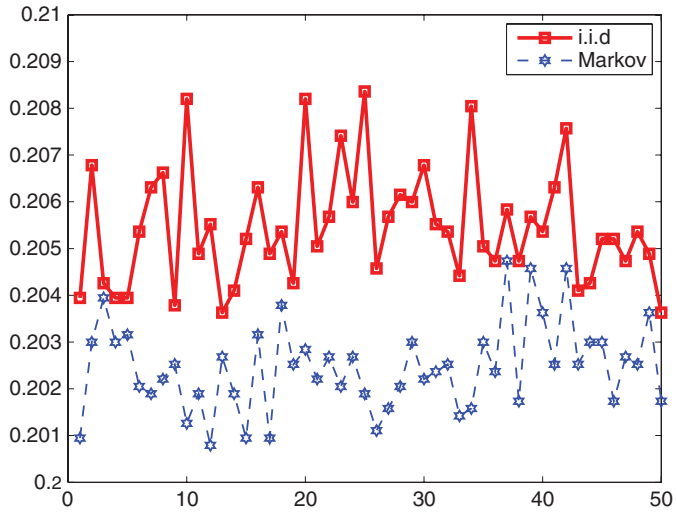

Fig. 4. UCI-magic, $m = 7000$.

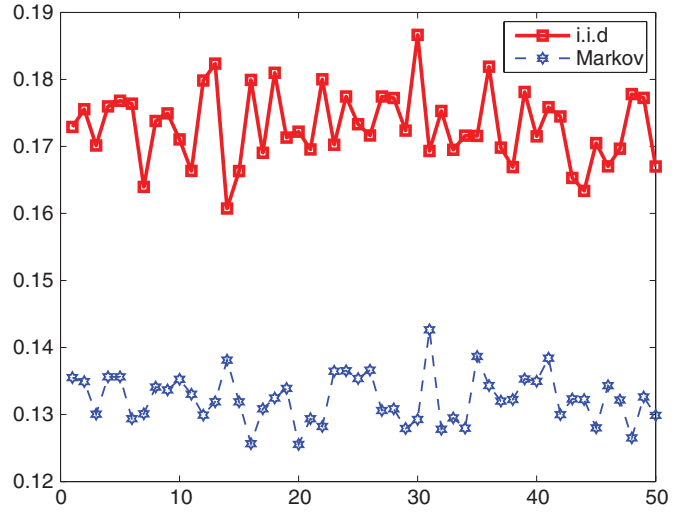

Fig. 5. Waveform, $m = 700$.

learning performance of FLD based on Markov sampling. In the experiments, we find that for most real-world datasets, the experimental results based on $N_1 < m$ are similar to those based on $N_1 = m$. For this reason, in order to simplify the experiments, we take $N_1 = m$ in Algorithm 1 for all of these experiments above. In addition, all the experimental results above are based on the FLD (2) method. Comparing the experimental results based on FLD (1) and FLD (2), we can find that, as the size of training samples is large, the bounds on the misclassification rates of FLD (2) are often tighter than those of FLD (1). By the figures and tables above, we can find that the misclassification rates of FLD based on Markov sampling will become smaller as the size of the training samples gets bigger. This proves the consistency of FLD based on Markov sampling.

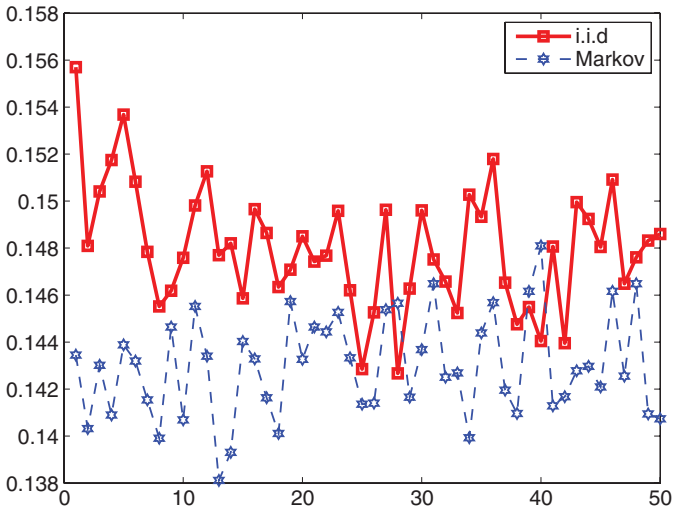

Fig. 6. Splice, $m = 6000$.

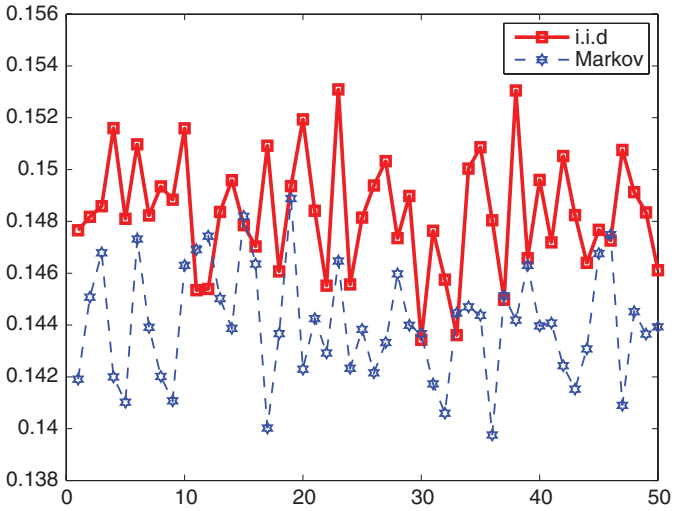

Fig. 7. Image, $m = 7000$.

D. Discussion

In this subsection, we give some useful discussion on the results obtained in the last subsections. First, in Algorithm 1, to generate u.e.M.c. samples, we first introduce the preliminary learning model f_0. Then the u.e.M.c. samples that are generated according to the preliminary learning model f_0 have

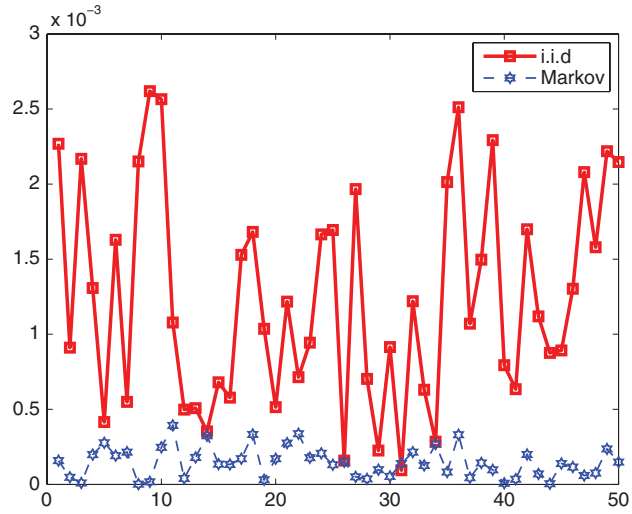

Fig. 8. UCI-mushrooms, $m = 2000$.

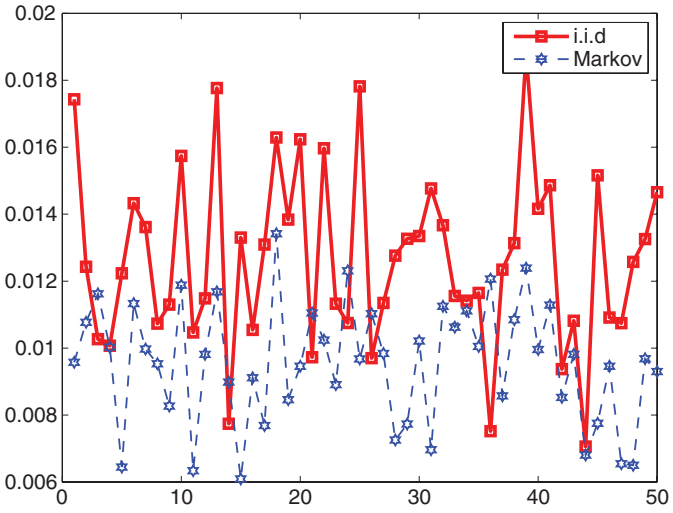

Fig. 9. DoubleUSPS(0,2), $m = 700$.

the structure information of the data. Thus, FLD based on Markov training samples can improve obviously the results of FLD based on i.i.d. samples.

Second, by the definition of the acceptance probability α, we can find that, for different Markov chain samples z_i and z_{i+1}, the loss $\mathcal{L}(f, z_{i+1})$ of Markov chain samples z_{i+1} is almost as large as that of $\mathcal{L}(f, z_i)$, whereas these samples that are drawn randomly from the same data do not possess this property since the i.i.d. samples are sampled randomly. In other words, these Markov chain samples are selective and representative compared to the i.i.d. samples. Therefore, the misclassification rates of FLD based on Markov chain samples can be smaller than those of i.i.d. samples. This implies that generating u.e.M.c. samples from a given data of finite large size by Algorithm 1 can be regarded as a strategy for improving the learning performance of FLD based on i.i.d. training samples. In other words, Algorithm 1 can be considered to be a method of manipulating the training samples [25] such that the learning performance of FLD is improved. Unlike other methods of manipulating the training samples in [25], the method presented in this paper generates u.e.M.c. samples

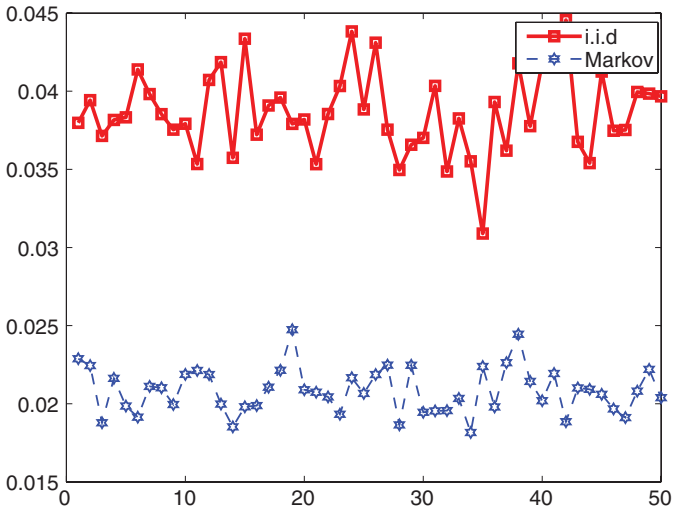

Fig. 10. UCI-gisette, $m = 4500$.

TABLE VI
MISCLASSIFICATION RATES FOR DIFFERENT TRAINING SIZES

Data	i.i.d. (1000)	Markov (600)	Markov (800)	Markov (1000)
A	0.2274 ± 0.0024	0.2262 ± 0.0047	0.2236 ± 0.0037	0.2215 ± 0.0032
M	0.2079 ± 0.0035	0.2110 ± 0.0050	0.2076 ± 0.0037	0.2071 ± 0.0038
Sh	0.0639 ± 0.0049	0.0679 ± 0.0073	0.0641 ± 0.0060	0.0615 ± 0.0058
W	0.1541 ± 0.0089	0.1127 ± 0.0129	0.1092 ± 0.0110	0.1046 ± 0.0086
Sp	0.1861 ± 0.0063	0.1952 ± 0.0078	0.1859 ± 0.0065	0.1812 ± 0.0052

from a given dataset of finite size. In addition, in order to have a better understanding of the Markov sampling algorithm, we also present the following numerical studies results of FLD based on u.e.M.c. samples for different sample sizes in Table VI.

Table VI shows that for the datasets of A(UCI-abalone), M(UCI-magic), Sh(UCI-shuttle), W(Waveform), and Sp(Splice), FLD based on smaller Markov chain samples (600 for UCI-abalone and waveform, 800 for UCI-magic and splice) can present smaller misclassification rates compared to more (1000) i.i.d. samples.

V. CONCLUSION

Previous works on the generalization ability of the FLD method were usually based on the assumption of i.i.d. samples. In this paper, we went far beyond this classical framework by studying the generalization ability of the FLD method based on u.e.M.c. samples. We first established the bounds on the generalization performance of FLD based on u.e.M.c. samples, and proved that FLD with u.e.M.c. samples is consistent. By following the enlightening idea from MCMC methods, we also introduced a Markov sampling algorithm to generate u.e.M.c. samples from a given data of finite size. Through simulation studies and numerical studies on benchmark repository using FLD, we found that FLD based on u.e.M.c. samples generated by Markov sampling could provide smaller misclassification rates compared to the i.i.d. samples. This implies that, for a given data of large size, we can improve the learning

performance of the FLD method by manipulating the training samples. In other words, generating u.e.M.c. samples from the given data of large size by Markov sampling can be regarded as a new method of manipulating the training samples such that the learning performance of the FLD method can be obviously improved. To our knowledge, the studies presented here are the first on this topic.

Along the line of the present work, several open problems deserve further research. For example, how to establish the bounds on the fast convergence rate of FLD with u.e.M.c. samples and how to apply the Markov sampling algorithm introduced in this paper to other learning algorithms (e.g., the online learning algorithm) are under current investigation.

APPENDIX A

In this section, we present the main tools used in this paper.

Definition 2 [2]: A Markov chain $\{Z_t\}_{t\geq 1}$ is said to be V-geometrically ergodic with respect to a measurable function $V : \mathcal{Z} \to [1, \infty)$ if there exist constants $\gamma_1 < \infty$ and $\rho_2 < 1$ such that $||P^n(\cdot|z_i) - \pi||_{TV} \leq \gamma_1 \rho_2^n V(z_i), z_i \in \mathcal{Z}, \forall n \geq 1$, and in addition $\mathrm{E}(V, \pi) = \int_{\mathcal{Z}} V(z)\pi(dz) < B < \infty$, where π is the stationary distribution of Markov chain $\{Z_t\}_{t\geq 1}$ and $\mathrm{E}(V, \pi)$ is the expectation of $V(z)$ with respect to π.

Definition 3: Let $(\mathcal{M}, d)$ be a pseudo-metric space and $\mathcal{S} \subset \mathcal{M}$ a subset. For every $\epsilon > 0$, the covering number $\mathcal{N}(\mathcal{S}, \epsilon, d)$ of $\mathcal{S}$ with respect to ϵ and d is defined as the minimal number $\ell \in \mathbf{N}$ of balls of radius ϵ whose union covers $\mathcal{S}$

$$\mathcal{N}(\mathcal{S}, \epsilon, d) = \min\{\ell : \mathcal{S} \subset \cup_{j=1}^{\ell} B(s_j, \epsilon), \{s_j\}_{j=1}^{\ell} \subset \mathcal{M}\}$$

where $B(s_j, \epsilon) = \{s \in \mathcal{M} : d(s, s_j) \leq \epsilon\}$ is a ball in $\mathcal{M}$.

The ℓ^2-empirical covering number of a function set is defined by means of the normalized ℓ^2-metric d_2 on the Euclidian space $\mathbf{R}$ given by

$$d_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{m}\sum_{i=1}^{m} |a_i - b_i|^2\right)^{\frac{1}{2}}, \mathbf{a} = (a_i)_{i=1}^m, \mathbf{b} = (b_i)_{i=1}^m.$$

Definition 4: Let $\mathcal{F}$ be a set of functions on $\mathcal{X}$, $\mathbf{x} = (\mathbf{x}_i)_{i=1}^m \subset \mathcal{X}^m$, and $\mathcal{F}|_{\mathbf{x}} = \{(f(\mathbf{x}_i))_{i=1}^m : f \in \mathcal{F}\}$. Set

$$\mathcal{N}_2(\mathcal{F}, \epsilon, d_2) = \sup_{\mathbf{x}\in\mathcal{X}^m} \mathcal{N}(\mathcal{F}|_{\mathbf{x}}, \epsilon, d_2), \quad \epsilon > 0.$$

Zhang established the bound (see Corollary 3.1 of [47]) on the covering number of linear function class $\mathcal{H}$ as follows:

$$\ln \mathcal{N}_2(\mathcal{H}, \epsilon, m) \leq \left\lceil \frac{a^2 b^2}{\epsilon^2} \right\rceil \ln(2h + 1), \quad \epsilon > 0.$$

It follows that there exists a constant C_0 such that

$$\mathcal{N}_2(\mathcal{H}, \epsilon, m) \leq \exp\left\{C_0\left(\frac{ab}{\epsilon}\right)^2\right\}, \quad \epsilon > 0 \tag{12}$$

where C_0 is a constant dependent of h.

To test the generalization ability of FLD based on u.e.M.c. samples, we explore the use of the β-mixing property of Markov chains. We present the definition of β-mixing and the others lemmas as follows: let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary process defined on a probability space $(X^\infty, \mathcal{S}^\infty, \tilde{P})$. For $-\infty < i < \infty$, let $\mathcal{A}_{-\infty}^k$ denote the σ-algebra generated by the random variables $X_i, i \leq k$, and similarly let $\mathcal{A}_k^\infty$ denote the σ-algebra generated by the random variables $X_i, i \geq k$. Let $\tilde{P}_{-\infty}^k$ and $\tilde{P}_k^\infty$ denote the corresponding marginal probability measures. Let $\tilde{P}_0$ denote the marginal probability of each of the X_i. Let $\bar{\mathcal{A}}_1^{k-1}$ denote the σ-algebra generated by the random variables $X_i, i \leq 0$ as well as $X_j, j \geq k$.

Definition 5 [2]: The sequence $\{X_t\}$ is called geometrically β-mixing if there exist constants ν and $\lambda_1 < 1$ such that β-mixing coefficient $\beta(k)$ satisfies

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

Lemma 2 [2]: Suppose X_i is a β-mixing process on a probability space $(X^\infty, \mathcal{S}^\infty, \tilde{P})$. Suppose $g : X^\infty \to R$ is essentially bounded and depends only on the variables $x_{ik}, 0 \leq i \leq l$. Let $\tilde{P}_0$ denote the 1-D marginal probability of each of the X_i. Then $|\mathrm{E}(g, \tilde{P}) - \mathrm{E}(g, \tilde{P}_0^\infty)| \leq l\beta(k)||g||_\infty$ where $\mathrm{E}(g, \tilde{P})$ and $\mathrm{E}(g, \tilde{P}_0^\infty)$ are the expectations of g with respect to $\tilde{P}$ and $\tilde{P}_0^\infty$, respectively.

Lemma 3 [48]: Suppose that ζ is a zero-mean random variable assuming values in the interval $[c_1, d_1]$. Then for any $s_1 > 0$, we have $\mathrm{E}[\exp(s_1\zeta)] \leq \exp(s_1^2(d_1 - c_1)^2/8)$.

Lemma 4 [2]: Suppose $\{\xi_t\}$ is a Markov chain V-geometrically ergodic. Then the sequence $\{\xi_t\}$ is geometrically β-mixing, and the β-mixing coefficient $\beta(n)$ is given by

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

Lemma 5 [39]: Let $c_2, c_3 > 0$, and $p_1 > p_2 > 0$. Then the Eq. $x^{p_1} - c_2x^{p_2} - c_3 = 0$ has a unique positive zero x^*. In addition $x^* \leq \max\{(2c_2)^{1/(p_1-p_2)}, (2c_3)^{(1/p_1)}\}$.

In order to prove the main results obtained in Section III, we first establish the following two important lemmas.

Lemma 6: Let ξ be a random variable on a probability space $\mathcal{Z}$ with mean $\mathrm{E}(\xi) = \mu$, and $\{z_i\}_{i=1}^m$ be u.e.M.c. samples. If $|\xi(z) - \mathrm{E}(\xi)| \leq B$ for all $z \in \mathcal{Z}$, then for any $\varepsilon, 0 < \varepsilon \leq 3B$

$$P\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \varepsilon\right\} \leq 2(1 + \gamma e^{-2})\exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2B^2}\right\}$$

where $m^{(\beta)}$ is defined as in Theorem 1.

Proof: We decompose the proof into three steps.

Step 1: Since u.e.M.c. is V-geometrically ergodic and by Lemma 4, we have that u.e.M.c. is geometrically β-mixing. To exploit the β-mixing property, we then decompose the index set $\hat{I} = \{1, 2, \ldots, m\}$ into different parts by following the idea of [2], i.e., given an integer m, choose any integer $k_m \leq m$, and define $l_m = \lfloor m/k_m \rfloor$ to be the integer part of m/k_m. For the time being, k_m and l_m are denoted, Respectively, by k and l, so as to reduce notational clutter. Let $r = m - kl$, and

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

Let $p_i = |I_i|/m$ for $i = 1, 2, \ldots, k$, and define

$$T_i = \xi(z_i) - \mu, a_m(\mathbf{z}) = \frac{1}{m}\sum_{i=1}^{m} T_i, b_i(\mathbf{z}) = \frac{1}{|I_i|}\sum_{j\in I_i} T_j.$$

Then we have $1/m\sum_{i=1}^{m}\xi(z_i)-\mu = a_m(\mathbf{z}) = \sum_{i=1}^{k} p_i b_i(\mathbf{z})$.

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

$$\exp\left[sa_m(\mathbf{z})\right] = \exp\left[\sum_{i=1}^{k} p_i s b_i(\mathbf{z})\right] \le \sum_{i=1}^{m} p_i \exp\left[sb_i(\mathbf{z})\right].$$

It follows that:

$$\mathrm{E}\left(e^{sa_m(\mathbf{z})}, \tilde{P}\right) \le \sum_{i=1}^{k} p_i \mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}\right). \tag{13}$$

Since

$$\exp[sb_i(\mathbf{z})] = \exp\left[\frac{s}{|I_i|}\sum_{j\in I_i} T_j\right] = \prod_{j\in I_i}\exp\left(\frac{sT_j}{|I_i|}\right) \le \left[\exp\left(\frac{sB}{|I_i|}\right)\right]^{|I_i|} \le e^{sB}$$

where in the last step we use the assumption $|T_i| = |\xi(z_1)-\mu| \le B$. Note that for $i = 1,2,\ldots,r$, the quantities $\mathrm{E}(e^{sb_i(\mathbf{z})}, \tilde{P})$ are all the same since the stochastic process is stationary. Moreover, since the components in the index set I_i are separated by at least k, it follows from Lemma 2 that

$$\mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}\right) \le l\beta(k)||e^{sb_i(\mathbf{z})}||_\infty + \mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}_0^\infty\right).$$

Similarly, for $i = r+1,\ldots,k$, $\mathrm{E}(e^{sb_i(\mathbf{z})}, \tilde{P})$ is the same due to the stationarity of the stochastic process. Moreover, it follows from the same lemma as above that

$$\mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}\right) \le (l-1)\beta(k)||e^{sb_i(\mathbf{z})}||_\infty + \mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}_0^\infty\right).$$

Then we have, for any $i = 1,2,\ldots,k$

$$\mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}\right) \le (|I_i|-1)\beta(k)||e^{sb_i(\mathbf{z})}||_\infty + \mathrm{E}\left(e^{sb_i(\mathbf{z})}, \tilde{P}_0^\infty\right).$$

Since under the measure $\tilde{P}_0^\infty$ the various z_i are independent, we have

$$\begin{aligned}\mathrm{E}(e^{sb_i(\mathbf{z})}, \tilde{P}_0^\infty) &= \mathrm{E}\Big[\prod_{j\in I_i}\exp(sT_j/|I_i|), \tilde{P}_0^\infty\Big] \\ &= \left\{\mathrm{E}[\exp(sT_j/|I_i|), \tilde{P}_0^\infty]\right\}^{|I_i|}. \end{aligned}\tag{14}$$

Apply Lemma 3 to the function T_j, we get

$$\mathrm{E}\Big[\exp(sT_j/|I_i|), \tilde{P}_0^\infty\Big] \le \exp\left(s^2B^2/2|I_i|^2\right).$$

Then we have, for any $s>0$

$$\mathrm{E}(e^{sb_i(\mathbf{z})}, \tilde{P}) \le \exp\left(\frac{s^2B^2}{2|I_i|}\right) + (|I_i|-1)\beta(k)e^{sB}.$$

Thus by inequality (13) and the inequality above, we have for any $s>0$

$$\mathrm{E}(e^{sa_m(\mathbf{z})}, \tilde{P}) \le \sum_{i=1}^{k} p_i\Big[\exp\left(\frac{s^2B^2}{2|I_i|}\right) + (|I_i|-1)\beta(k)e^{sB}\Big]. \tag{15}$$

Step 2: We now bound the second term on the right-hand side of inequality (15), which is denoted henceforth by ϕ. By Lemma 4 and Definition 1, we have

$$\beta(k) = \mathrm{E}\{||P^k(\cdot|x)-\pi(\cdot)||_{TV}, \pi\} \le \mathrm{E}[\gamma\rho_1^k, \pi] = \gamma\rho_1^k.$$

Then we have, for any $0 < s \le 3|I_i|/B$

$$\begin{aligned}\phi &= \exp\left(\frac{s^2B^2}{2|I_i|}\right) + (|I_i|-1)\beta(k)e^{sB} \\ &\le \exp\left(\frac{s^2B^2}{2|I_i|}\right) + e^{|I_i|}e^{-2}\gamma\rho_1^k\cdot e^{sB} \\ &\le \exp\left(\frac{s^2B^2}{2|I_i|}\right) + \gamma e^{-2}\exp\left\{k\ln(\rho_1) + 4|I_i|\right\}.\end{aligned}$$

The above inequality follows from the fact that $|I_i-1| \le e^{|I_i|-2}$ for $|I_i| \ge 2$. We require $\exp\{k\ln(\rho_1)+4|I_i|\} \le 1$. But $|I_i| \le (m/k+1)$, thus the bound holds if $4(m/k+1) \le k\ln(1/\rho_1)$ or $4(m+k) \le k^2\ln(1/\rho_1)$. Since $m+k \le 2m$, the bound holds if $\{8m/\ln(1/\rho_1)\}^{\frac{1}{2}} \le k$. Let $k = \lceil\{8m/\ln(1/\rho_1)\}^{\frac{1}{2}}\rceil$. Since for all $i = 1,2,\ldots,k$, $|I_i| \ge l$, and $l = \lfloor m/k\rfloor$, we have

$$\phi \le \exp\left(s^2B^2/(2l)\right) + \gamma e^{-2}. \tag{16}$$

Since inequality (16) is true for all $s, 0 < s \le 3|I_i|/B$. To make the constraint uniform over all i, we require s to satisfy $0 < s < 3l/B \le 3|I_i|/B$. Since $s^2B^2/2l > 0$, we have

$$\phi \le (1+\gamma e^{-2})\exp\left(s^2B^2/(2l)\right).$$

Returning to inequality (15), we have for any $0 < s < 3l/B$

$$\mathrm{E}(e^{sa_m(\mathbf{z})}, \tilde{P}) \le (1+\gamma e^{-2})\exp\left(s^2B^2/(2l)\right). \tag{17}$$

Step 3: By Markov's inequality and inequality (17), we have for any $0 < s \le 3l/B$

$$\begin{aligned}P\left\{\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)-\mu \ge \varepsilon\right\} &= \mathrm{P}\left\{e^{s[\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)-\mu]} \ge e^{s\varepsilon}\right\} \\ &\le \frac{\mathrm{E}\left\{e^{s[\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)-\mu]}\right\}}{e^{s\varepsilon}} \\ &\le C_1\exp\left\{-s\varepsilon + s^2B^2/(2l)\right\}\end{aligned}$$

where $C_1 = 1+\gamma e^{-2}$. Substituting $s = l\varepsilon/B^2$, and noting that for any $\varepsilon \le 3B$, s satisfies $s < 3l/B$, we obtain

$$P\left\{\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)-\mu \ge \varepsilon\right\} \le (1+\gamma e^{-2})\exp\left\{\frac{-l\varepsilon^2}{2B^2}\right\}.$$

By symmetry, we also have

$$P\left\{\mu-\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) \ge \varepsilon\right\} \le (1+\gamma e^{-2})\exp\left\{\frac{-l\varepsilon^2}{2B^2}\right\}.$$

Combining the two inequalities above and replacing l by $m^{(\beta)}$, we complete the proof of Lemma 6.

Lemma 7: Let $\mathcal{L}(f) = \mathcal{E}(f) - \mathcal{E}_m(f)$. If $\{z_i\}_{i=1}^m$ is u.e.M.c., then for any $\varepsilon, 0 < \varepsilon \le 6(ab+1)^2$, and $m \ge \max\{\ln(1/\rho_1)/8, 128/\ln(1/\rho_1)\}$

$$P\left\{\sup_{f\in\mathcal{H}}|\mathcal{L}(f)| \ge \varepsilon\right\} \le 2C_1\mathcal{N}_2(\mathcal{H}, L', m)\exp\left\{\frac{-2^{\frac{1}{2}}m^{\frac{1}{2}}\varepsilon^2}{K}\right\}$$

where $\mathcal{H}$ is defined as (5) $K = [\ln(1/\rho_1)]^{\frac{1}{2}}(ab+1)^4$, $C_1 = 1+\gamma e^{-2}$ and $L' = \varepsilon/8(ab+1)$.

Proof: First we need to slightly modify the inequality in Lemma 6. We observe that $\lceil t \rceil \le 2t$ for any $t \ge 1$ and $\lfloor t \rfloor \ge t/2$ for any $t \ge 2$ (see [8]). Then it is easy to conclude that for m satisfying $m \ge m_0 := \max\{\ln(1/\rho_1)/8, 128/\ln(1/\rho_1)\}$, we have $m^{(\beta)} \ge 8(\ln(1/\rho_1)/2)^{1/2} m^{1/2}$. Then by Lemma 6 and replacing $m^{(\beta)}$ by $8(\ln(1/\rho_1)/2)^{1/2} m^{1/2}$, we have, for any $\varepsilon, 0 < \varepsilon \le 3B$

$$P\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \ge \varepsilon\right\} \le 2C_1 \exp\left\{\frac{-2^{\frac{5}{2}} m^{\frac{1}{2}} \varepsilon^2}{[\ln(1/\rho_1)]^{\frac{1}{2}} B^2}\right\}.$$

Since for any $f \in \mathcal{H}$, $y \in \mathcal{Y}$, we have $\ell(f,z) \doteq (f(x)-y)^2 \le (ab+1)^2$. It follows that $|\ell(f,z) - \mathrm{E}[\ell(f,z)]| \le (ab+1)^2$. By the above inequality, we have that for any $0 < \varepsilon \le 3(ab+1)^2$, and $m \ge \max\{\ln(1/\rho_1)/8128/\ln(1/\rho_1)\}$

$$P\big\{|\mathcal{L}(f)| \ge \varepsilon\big\} \le 2C_1 \exp\left\{\frac{-2^{\frac{5}{2}} m^{\frac{1}{2}} \varepsilon^2}{[\ln(1/\rho_1)]^{\frac{1}{2}} (ab+1)^2}\right\}. \quad (18)$$

In addition, for any $f_1, f_2 \in \mathcal{H}$, we have

$$\begin{aligned}|\ell(f_1,z) - \ell(f_2,z)| &:= |(f_1(\mathbf{x}) - y)^2 - (f_2(\mathbf{x}) - y)^2| \\ &\le 2(ab+1) \cdot |f_1(\mathbf{x}) - f_2(\mathbf{x})|.\end{aligned}$$

The final inequality follows as $|f(\mathbf{x})| \le ab$. Thus by inequality (18) and with a similar argument as Theorem B in [37], we can finish the proof of Lemma 7.

Proof of Theorem 1: By inequality (12) and Lemma 7, we have that for any $\varepsilon, 0 < \varepsilon \le 6(ab+1)^2$

$$P\left\{\sup_{f\in\mathcal{H}} |\mathcal{L}(f)| \ge \varepsilon\right\} \le 2C_1 \exp\left\{C_0\left(\frac{ab}{L'}\right)^2 - \frac{-2^{\frac{1}{2}} m^{\frac{1}{2}} \varepsilon^2}{K}\right\}.$$

Let us rewrite the above inequality in the equivalent form. We equate the right-hand side of the above inequality to a positive value η $(0 < \eta < 1)$

$$C_1 \exp\left\{C_0\left(\frac{8ab(ab+1)}{\varepsilon}\right)^2 - \frac{\ln(1/\rho_1)^{\frac{1}{2}} m^{\frac{1}{2}} \varepsilon^2}{2^{1/2}(ab+1)^4}\right\} = \eta.$$

It follows that:

$$\varepsilon^4 - \frac{2^{1/2}(ab+1)^4 \ln(C_1/\eta)}{m^{\frac{1}{2}} \ln(1/\rho_1)^{\frac{1}{2}}}\varepsilon^2 - \frac{2^{\frac{9}{2}} C_0 a^2 b^2 (ab+1)^6}{m^{\frac{1}{2}} \ln(1/\rho_1)^{\frac{1}{2}}} = 0.$$

By Lemma 5, we can solve this equation with respect to ε. The solution is given by

$$\varepsilon \doteq \varepsilon(m,\eta) \le \max\left\{(ab+1)^2\Big[\frac{2^{\frac{3}{2}}\ln(C_1/\eta)}{m^{\frac{1}{2}}\ln(1/\rho_1)^{\frac{1}{2}}}\Big]^{\frac{1}{2}}, 2(ab+1)\Big[\frac{2^{\frac{3}{2}} C_0 a^2 b^2 (ab+1)^2}{m^{\frac{1}{2}}\ln(1/\rho_1)^{\frac{1}{2}}}\Big]^{\frac{1}{4}}\right\}.$$

Then we can deduce that, with probability at least $1-\eta$ simultaneously for all functions in the function set $\mathcal{H}$, inequality $\mathcal{E}(f) \le \mathcal{E}_m(f) + \varepsilon(m,\eta)$ holds true. Since with probability at least $1-\eta$ this inequality holds for all functions of the function set $\mathcal{H}$, it holds in particular for the function $f_{\mathbf{z}}$ that minimizes the empirical error $\mathcal{E}_m(f)$ over $\mathcal{H}$. For this function with probability at least $1-\eta$, the following inequality then holds true:

$$\mathcal{E}(f_{\mathbf{z}}) \le \mathcal{E}_m(f_{\mathbf{z}}) + \varepsilon(m,\eta). \quad (19)$$

In addition, by inequality (18), we conclude that for the same η as above, and for the function $f_{\mathcal{H}}$ that minimizes the expected error $\mathcal{E}(f)$ over $f \in \mathcal{H}$, the inequality

$$\mathcal{E}(f_{\mathcal{H}}) > \mathcal{E}_m(f_{\mathcal{H}}) - \frac{(ab+1)}{2}\sqrt{\frac{\ln(C_1/\eta)}{\ln(1/\rho_1)^{\frac{1}{2}} m^{\frac{1}{2}}}} \quad (20)$$

holds with probability $1-\eta$. Note that

$$\mathcal{E}_m(f_{\mathcal{H}}) \ge \mathcal{E}_m(f_{\mathbf{z}}). \quad (21)$$

From (7), (19), (20), and (21), we deduce that with probability at least $1-2\eta$, the inequality

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le{}& \varepsilon(m,\eta) + \frac{(ab+1)}{2}\sqrt{\frac{\ln(C_1/\eta)}{\ln(1/\rho_1)^{\frac{1}{2}} m^{\frac{1}{2}}}} \\ &+ \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho)\end{aligned}$$

is valid. In addition, if

$$m \ge \max\left\{\frac{1}{6(ab+1)}\sqrt{\frac{\ln(C_1/\eta)}{\ln(1/\rho_1)^{\frac{1}{2}}}}, \frac{1}{6}\sqrt{\frac{2^{\frac{3}{2}}\ln(C_1/\eta)}{\ln(1/\rho_1)^{\frac{1}{2}}}}, \frac{ab}{9(ab+1)}\sqrt{\frac{2^{\frac{3}{2}} C_0}{\ln(1/\rho_1)^{\frac{1}{2}}}}\right\}$$

then we have $\varepsilon \le \min\{3(ab+1), 6(ab+1)^2\}$. This leads to Theorem 1.

Proof of Theorem 2: By the definition of $\tilde{D}(\lambda)$, we have

$$\lambda ||\tilde{f}_\lambda||_2^2 \le \mathcal{E}(\tilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda ||\tilde{f}_\lambda||_2^2 = \tilde{D}(\lambda).$$

It follows that $||\tilde{f}_\lambda||_2 \le \sqrt{\tilde{D}(\lambda)/\lambda}$, and $\ell(\tilde{f}_\lambda, z) \le \left(a\sqrt{\tilde{D}(\lambda)/\lambda} + 1\right)^2 \doteq C_2$. The final inequality follows from the fact that $||\mathbf{x}||_2 \le a < \infty$.

By Lemma 6, we have, for any $\varepsilon, 0 < \varepsilon \le 3C_2$

$$P\big\{\mathcal{E}_m(\tilde{f}_\lambda) - \mathcal{E}(\tilde{f}_\lambda) \ge \varepsilon\big\} \le (1 + \gamma e^{-2}) \exp\left\{\frac{-m^{(\beta)}\varepsilon^2}{2C_2^2}\right\}.$$

It follows that for any $\delta \in (0,1)$, there exists a subset V_2 of $\mathcal{Z}^m$ such that for any $\tilde{f}_\lambda \in \mathcal{H}$ and for any $\mathbf{z} \in V_2$, inequality

$$\mathcal{E}_m(\tilde{f}_\lambda) - \mathcal{E}(\tilde{f}_\lambda) \le C_2\sqrt{\frac{2\ln[(1+\gamma e^{-2})/\delta]}{m^{(\beta)}}} \quad (22)$$

is valid with probability at least $1-\delta$. In addition, by Lemma 7 we deduce that for the same δ above, there exists a subset $V'(R)$ of $\mathcal{Z}^m$ such that for any $\mathbf{z} \in V'(R)$, the inequality

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_m(f_{\mathbf{z},\lambda}) \le \varepsilon(m,\delta) \quad (23)$$

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

$$\varepsilon \doteq \varepsilon(m,\delta) \le \max\left\{2(ab+1)^2\Big[\frac{2\ln\left((1+\gamma e^{-2})/\delta\right)}{m^{(\beta)}}\Big]^{\frac{1}{2}}, \times 2(ab+1)\Big[\frac{8C_0 a^2 b^2 (ab+1)^2}{m^{(\beta)}}\Big]^{\frac{1}{4}}\right\}.$$

Let $W'(R) = \{\mathbf{z} \in V_2 : f_{\mathbf{z},\lambda} \in \mathcal{H}\}$. By inequalities (22) and (23), we deduce that for any $\mathbf{z} \in V'(R) \cap W'(R)$, with probability at least $1 - 2\delta$

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_m(f_{\mathbf{z},\lambda}) + \mathcal{E}_m(\tilde{f_\lambda}) - \mathcal{E}(\tilde{f_\lambda}) \leq \varepsilon(m,\delta) + C_2\sqrt{\frac{2\ln[(1+\gamma e^{-2})/\delta]}{m^{(\beta)}}}.$$

Combine the above inequality with inequality (11), we have

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \leq C_2\sqrt{\frac{2\ln[(1+\gamma e^{-2})/\delta]}{m^{(\beta)}}} + \varepsilon(m,\delta) + \tilde{D}(\lambda). \tag{24}$$

Since for all $\lambda > 0$, and almost all $\mathbf{z} \in \mathcal{Z}^m$, and the definition of $f_{\mathbf{z},\lambda}$, we have

$$\lambda||f_{\mathbf{z},\lambda}||_2^2 \leq \mathcal{E}_m(f_{\mathbf{z},\lambda}) + \lambda||f_{\mathbf{z},\lambda}||_2^2 \leq \mathcal{E}_m(0) + 0 \leq 1.$$

It follows that $||f_{\mathbf{z},\lambda}||_2 \leq 1/\sqrt{\lambda}$ for almost all $\mathbf{z} \in \mathcal{Z}^m$. This implies that $W(1/\sqrt{\lambda}) = \mathcal{Z}^m$. Replacing ab by $1/\sqrt{\lambda}$ in inequality (24), we complete the proof of Theorem 2.

ACKNOWLEDGMENT

The authors would like to thank Prof. D. Liu, Editor-in-Chief, the handling Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and the three anonymous referees, whose careful comments and valuable suggestions led to a significant improvement in the presentation of this paper.

REFERENCES

[1] I. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations" *J. Multivariate Anal.*, vol. 100, no. 1, pp. 175–194, 2009.
[2] M. Vidyasagar, *Learning and Generalization with Applications to Neural Networks*. London, U.K.: Springer-Verlag, 2003.
[3] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *Ann. Probab.*, vol. 22, no. 1, pp. 94–116, 1994.
[4] S. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," *IEEE Trans. Inform. Theory*, vol. 42, no. 6, pp. 2133–2145, Nov. 1996.
[5] A. Lozano, S. Kulkarni, and R. Schapire, "Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 819–826.
[6] 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.
[7] M. Mohri and A. Rostamizadeh, "Rademarcher complexity bounds for non-i.i.d. processes" in *Advances in Neural Information Processing Systems*. Vancouver, BC, Canada: MIT Press, 2009.
[8] I. Steinwart and A. Christmann, "Fast learning from non-i.i.d. observations," in *Advances in Neural Information Processing Systems*, vol. 22. Cambridge, MA: MIT Press, 2009, pp. 1768–1776.
[9] S. Smale and D. X. Zhou, "Online learning with Markov sampling," *Anal. Appl.*, vol. 7, no. 1, pp. 87–113, 2009.
[10] H. W. Sun and Q. Wu, "Regularized least square regression with dependent samples," *Adv. Comput. Math.,* vol. 32, no. 2, pp. 175–189, 2010.
[11] B. Zou, L. Q. Li, and Z. B. Xu, "The generalization performance of ERM algorithm with strongly mixing observations," *Mach. Learn.*, vol. 75, no. 3, pp. 275–295, Jun. 2009.
[12] K. Fukunaga, "Nonparametric discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 6, pp. 671–678, Nov. 1983.
[13] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.
[14] T. Hastie, R. Tibshirani, and A. Buja, "Flexible discriminant analysis by optimal scoring," *J. Amer. Stat. Assoc.* vol. 89, no. 428, pp. 1255–1270, 1994.
[15] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *J. Royal Stat. Soc. (Series B)*, vol. 58, no. 1, pp. 155–176, 1996.
[16] M. Loog, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 762–766, Jul. 2001.
[17] O. C. Hamsici and A. M. Martinez, "Bayes optimality in linear discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 647–657, Apr. 2008.
[18] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.
[19] A. Stuhlsatz, J. Lippel, and T. Zielke, "Feature extraction with deep neural networks by a generalized discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 4, pp. 596–608, Apr. 2012.
[20] Y. Hou, L. Song, H. K. Min, and C. H. Park, "Complexity-reduced scheme for feature extraction with linear discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 1003–1009, Jun. 2012.
[21] G. O. Roberts and J. S. Rosenthal, "General state space Markov chains and MCMC algorithms," *Probab. Surv.*, vol. 1, pp. 20–71, Apr. 2004.
[22] B. Zou, H. Zhang, and Z. B. Xu, "Learning from uniformly ergodic Markov chain samples," *J. Complex.*, vol. 25, no. 2, pp. 188–200, Apr. 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 distribution and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.
[25] T. G. Dirtterich, *Ensemble Methods in Machine Learning* (Lecture Notes in Computer Science), vol. 1857. New York: Springer-Verlag, 2000, pp. 1–15.
[26] S. P. Meyn and R. L. Tweedie, *Markov chains and Stochastic Stability*. New York: Springer-Verlag, 1993.
[27] D. Aldous, L. Lovász, and P. Winkler, "Mixing times for uniformly ergodic Markov chains," *Stochast. Process Appl.*, vol. 71, no. 2, pp. 165–185, Nov. 1997.
[28] R. O. Duda and P. E. Hart, *Pattern classification and Scene Analysis*. New York: Wiley, 1973.
[29] K. Fukunaga, *Introduction to Statistical Pattern Classification*. San Francisco, CA: Academic, 1990.
[30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag, 2001.
[31] I. C. Goknar, M. Yildiz, S. Minaei, and E. Deniz, "Neural CMOS-integrated ciruit and its application to data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 717–724, May 2012.
[32] P. Zhang and N. Riedel, "Discriminant analysis: A unified approach," in *Proc. 5th IEEE Int. Conf. Data Mining*, Nov. 2005, pp. 514–521.
[33] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[34] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.
[35] S. Mika, "Kernal Fisher discriminants," Ph.D. dissertation, Dept. Electr. Eng. Inf. Technol., Univ. Technol., Berlin, Germany, Oct. 2002.
[36] R. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, no. 2, pp. 178–188, 1936.
[37] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, no. 39, pp. 1–49, 2001.
[38] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
[39] F. Cucker and S. Smale, "Best choices for regularization parameters in learning theory: On the bias-variance problem," *Found. Comp. Math.*, vol. 2, no. 4, pp. 413–428, 2002.
[40] Q. Wu, "Classification and regularization in learning theory," Ph.D. dissertation, Facul. Math., City Univ. Hong Kong, Hong Kong, 2005.
[41] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, 2000.
[42] Q. Wu, Y. M. Ying, and D. X. Zhou, "Learning rates of least-square regularized regression," *Found. Comput. Math.*, vol. 6, no. 2, pp. 171–192, Apr. 2006.
[43] A. E. Gelfand and A. F. Smith, "Sampling-based approaches to calculating marginal densities," *J. Amer. Stat. Assoc.*, vol. 85, no. 410, pp. 398–409, 1990.

[44] N. Metropolis and S. Ulam, "The Monte Carlo method," *J. Amer. Stat. Asso.*, vol. 44, no. 247, pp. 335–341, Sep. 1949.
[45] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. Teller, and H. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1091, 1953.
[46] C. Andrieu, N. D. Freitas, A. Doucet, and M. Jordan, "An introduction to MCMC for machine learning," *Mach. Learn.*, vol. 50, no. 1, pp. 5–43, 2003.
[47] T. Zhang, "Covering number bounds of certain regularized linear function classes," *J. Mach. Learn. Res.*, vol. 2, pp. 527–550, Mar. 2002.
[48] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, Mar. 1963.

Bin Zou received the Ph.D. degree 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, Hubei Province, and the Faculty of Mathematics and Computer Science, Hubei University, where he has been an Associate Professor since 2007. His current research interests include statistical learning theory, machine learning, and pattern recognition.

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 a Professor with the Faculty of Mathematics and Computer Science, Hubei University. His current research interests include approximation theory, learning theory, image processing, and pattern recognition.

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

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

He is currently the Vice President of Xi'an Jiaotong University, where he is the Chief Scientist of the National Basic Research Program of China for the 973 Project and the Director of the Institute for Information and System Sciences. He gave a talk at the 2010 International Congress of Mathematicians. 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.

Tao Luo received the B.Sc. degree in applied mathematics from the University of Electronic Science and Technology of China, Chengdu, China, in 2010. He is currently pursuing the M.Sc. degree with the Institute for Information and System Science, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China.

His current research interests include learning to rank, hyperspectral image processing, machine learning, and statistical learning theory.

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

He is currently a Chair Professor with the Faculty of Science and Technology, University of Macau, Macau, China, and a Professor, an Adjunct Professor, or an Honorary Professor with several institutes, including Chongqing University, Concordia University, and Hong Kong Baptist University, Hong Kong. He has authored or co-authored more than 360 technical papers in journals and conferences, and has authored or co-authored over 20 monographs, books, and book chapters on electrical engineering and computer science. His current research interests include wavelet theory and applications, pattern recognition, image processing, document processing, artificial intelligence, and Chinese computing.

Dr. Tang is the Founder and the Editor-in-Chief of the *International Journal on Wavelets*, *Multiresolution, and Information Processing*, an Associate Editor-in-Chief of the *International Journal on Frontiers of Computer Science*, and an Associate Editor of several international journals related to Pattern Recognition and Artificial Intelligence. He is the Founder and the Chair of Pattern Recognition Committee of the IEEE SMC. He is the Founder and the General Chair of the series International Conferences on Wavelets Analysis and Pattern Recognition. He is Fellow of the IAPR. He was the General Chair of the Program Chair or a Committee Member of many international conferences including the 18th International Conference on Pattern Recognition (ICPR'06).