Journal of Complexity 25 (2009) 188–200

Contents lists available at ScienceDirect

Journal of Complexity

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

Learning from uniformly ergodic Markov chains☆

Bin Zou[a,b,*], Hai Zhang[a], Zongben Xu[a]

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

ARTICLE INFO

Article history:
Received 16 April 2008
Accepted 13 January 2009
Available online 30 January 2009

Keywords:
ERM algorithms
Uniform ergodic Markov chain samples
Generalization bound
Uniform convergence
Relative uniform convergence

ABSTRACT

Evaluation for generalization performance of learning algorithms has been the main thread of machine learning theoretical research. The previous bounds describing the generalization performance of the empirical risk minimization (ERM) algorithm are usually established based on independent and identically distributed (i.i.d.) samples. In this paper we go far beyond this classical framework by establishing the generalization bounds of the ERM algorithm with uniformly ergodic Markov chain (u.e.M.c.) samples. We prove the bounds on the rate of uniform convergence/relative uniform convergence of the ERM algorithm with u.e.M.c. samples, and show that the ERM algorithm with u.e.M.c. samples is consistent. The established theory underlies application of ERM type of learning algorithms.

1. Introduction

Recently Support Vector Machines (SVMs) have become one of the most widely used algorithms in the machine learning community [1]. Besides their good performance in practical applications they also enjoy a good theoretical justification in terms of both universal consistency and learning rates when training samples come from an i.i.d. process. This renewed interest for theory naturally boosted the development of performance bounds for learning algorithms (see [2–6], etc.). However, this i.i.d. assumption cannot often be strictly justified in real-world applications, and many machine learning applications such as market prediction, system diagnosis, and speech recognition are inherently temporal in nature, and consequently not i.i.d. processes [7]. Relaxations of such i.i.d. assumption have

☆ Supported by National 973 project (2007CB311002), NSFC key project (70501030) and Foundation of Hubei Educational Committee (Q200710001).

* Corresponding author at: Faculty of Mathematics and Computer Science, Hubei University, Wuhan, 430062, PR China.
E-mail addresses: zoubin0502@hubu.edu.cn (B. Zou), zhanghai@nwu.edu.cn (H. Zhang), zbxu@mail.xjtu.edu.cn (Z. Xu).

0885-064X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2009.01.001

been considered for quite a while in both machine learning and statistics literatures. For example, Yu [8] established the rates of convergence for empirical processes of stationary mixing sequences. Modha and Masry [9] established the minimum complexity regression estimation with m-dependent observations and strongly mixing observations respectively. Vidyasagar [10] considered the notions of mixing and proved that most of the desirable properties (e.g. PAC or UCEMUP property) of i.i.d. sequence are preserved when the underlying sequence is mixing sequence. Steinwart, Hush and Scovel [7] proved that the SVMs for both classification and regression are consistent only if the data-generating process satisfies a certain type of law of large numbers (e.g. WLLNE, SLLNE). Smale and Zhou [11] considered online regularization learning algorithm based on Markov sampling. Zou and Li [12] established the bounds on the rate of uniform convergence of learning machines with strongly mixing observations. Zou, Li and Xu [13] obtained the generalization bounds of the ERM algorithm with exponentially strongly mixing observations.

There have been many dependent (not i.i.d.) sampling mechanisms studied in machine learning literatures ([14,15] etc.). In the present paper we focus on an analysis in the case when the samples are Markov chains (that is, the Markov chain samples). The Markov chain samples appear so often and naturally in applications, especially in biological (DNA or protein) sequence analysis, speech recognition, character recognition, content-based web search and marking prediction. Two examples are as follows:

Example 1. Consider the problem of an insurance company wanting to draft the amount of insurance money and claim settlement according to the health condition of insurance applicants. In the simplest case, the health condition of an insurance applicant consists of healthy and ill. For an insurance applicant during given age stage, we suppose that the probability that he/she is healthy this year and also next year is given. The probability that he/she is ill this year but healthy next year is also known. Let x_i be the health condition given by the ith year, and y_i be the corresponding profit or loss the insurance company made. Then $\{x_i\}$ is a sequence with Markov property. The insurance company had a data set of past insurance applicants and the profit or loss of the company. To draft the amount of insurance money and claim settlement, one should learn the unknown functional dependency between x_i and y_i from the Markov chain samples $\{\mathbf{z}_i = (x_i, y_i)\}_{i\geq 1}$.

Example 2. We usually have the following quantitative example in the models of random walk and predicting the weather, that is, suppose that $\{x_i\}$ is a Markov chain consisting of five states 1, 2, 3, 4, 5 and having transition probability matrix

$$P = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.1 & 0.3 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.1 & 0.3 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.1 & 0.3 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.1 & 0.3 \end{bmatrix}.$$

By the matrix P and Matlab software, we can create a sequence with Markov property, for example, $x_1 = 1, x_2 = 1, x_3 = 5, x_4 = 3, \ldots$. Through target function $y = f(x) = x^2 + 10x + 3$, we also can produce the corresponding values of x_i, that is, $y_1 = 14, y_2 = 14, y_3 = 78, y_4 = 42, \ldots$. Then a problem is posed: how can we learn the target function $f(x) = x^2 + 10x + 3$ from the Markov chain samples

$$S = \{\mathbf{z}_1 = (1, 14), \mathbf{z}_2 = (1, 14), \mathbf{z}_3 = (5, 78), \mathbf{z}_4 = (3, 42), \ldots\}.$$

Many empirical evidences show that a learning algorithm very often performs well with Markov chain samples ([16,17], etc.). Why it is so, however, has been unknown (particularly, it is unknown how well it performs in terms of consistency and generalization). Answering those questions is the purpose of the present paper. We will provide theoretical justification of the success of the ERM algorithm by establishing a consistency and the generalization bound estimation results of the ERM algorithm with u.e.M.c. samples. Following this schedule, in Section 2 we introduce some notions and notations. In Sections 3 and 4 we derive the bounds on the rate of uniform convergence and relative uniform convergence of the ERM algorithm respectively, and obtain the generalization bounds of the