Available online at www.sciencedirect.com
SCIENCE DIRECT®

NEUROCOMPUTING

Neurocomputing 69 (2005) 62–84

www.elsevier.com/locate/neucom

Algebraic geometry and stochastic complexity of hidden Markov models

Keisuke Yamazaki*, Sumio Watanabe

Precision and Intelligence Laboratory, Tokyo Institute of Technology, 4259 Nagatsuda, Midori-ku, Yokohama 226-8503, Japan

Received 29 March 2004; received in revised form 18 November 2004; accepted 9 February 2005
Available online 8 September 2005

Abstract

Hidden Markov models are now used in many fields, for example, speech recognition, natural language processing, etc. However, the mathematical foundation of analysis for the models is not yet constructed, since the HMM is non-identifiable.

In recent years, we have developed the algebraic geometrical method that allows us to analyze the non-regular and non-identifiable models. In this paper, we apply this method to the HMM and reveal the asymptotic stochastic complexity in a mathematically rigorous way.

Our results show that the Bayesian estimation makes the generalization error small and that the well known BIC is different from the stochastic complexity.

Keywords: Hidden Markov models; Stochastic complexity; Algebraic geometry

1. Introduction

Hidden Markov models [13] are now used in many areas, for example, speech recognition, natural language processing, bioinformatics, and so on. The models are robust to nonlinear time scaling, and used for learning time series. In spite of these

*Corresponding author. Tel.: +81 45 924 5018; fax: +81 45 924 5018.
E-mail address: k-yam@pi.titech.ac.jp (K. Yamazaki).

0925-2312/$ - see front matter
doi:10.1016/j.neucom.2005.02.014

enormous applications and technical learning algorithms, their theoretical properties have not yet been clarified.

All learning models belong to either the category, *identifiable* or *nonidentifiable*. A learning model is generally represented as a probability density function $p(x\,|\,w)$, where w is a parameter. If the mapping from the parameter to the probability density function is one-to-one, the model is *identifiable*, otherwise, *nonidentifiable*.

From the statistical point of view, we cannot analyze the nonidentifiable model by the conventional method. If the learning model attains the true distribution, the parameter space contains the true parameter(s). In nonidentifiable models, the set of true parameters is not one point but an analytic set in the parameter space. Because the set includes many singularities, the Fisher matrices are not positive definite. The log likelihood cannot be approximated by any quadratic form of the parameter in the neighborhood of the singularities [2,23]. This is the reason why the mathematical properties of nonidentifiable model still remain unknown.

The hidden Markov model is also nonidentifiable such as multilayered perceptrons, mixture models, and Boltzmann machines. Let us look at a simple example. Let $q(x)$ be the true distribution, which has one hidden state (Fig. 1a). The datum $x = (y_1, y_2)$ is taken from this distribution. In this example, the length of data is two for brevity. Since this hidden Markov model always remains in the same state, $q(x)$ is written as

$$q(x) = \prod_{t}^{2} f(y_t\,|\,b^*),$$

where $f(y_t\,|\,b^*)$ is a probability density function and $b^* \in R^M$. We assume that the learning machine can attain $q(x)$. Let $p(x\,|\,w)$ be the learning model which has two hidden states (Fig. 1b). There are two probability functions $f_1(y_t\,|\,b_1)$ and $f_2(y_t\,|\,b_2)$ in the states. Because a_{ij} shows the transition probability from the ith state

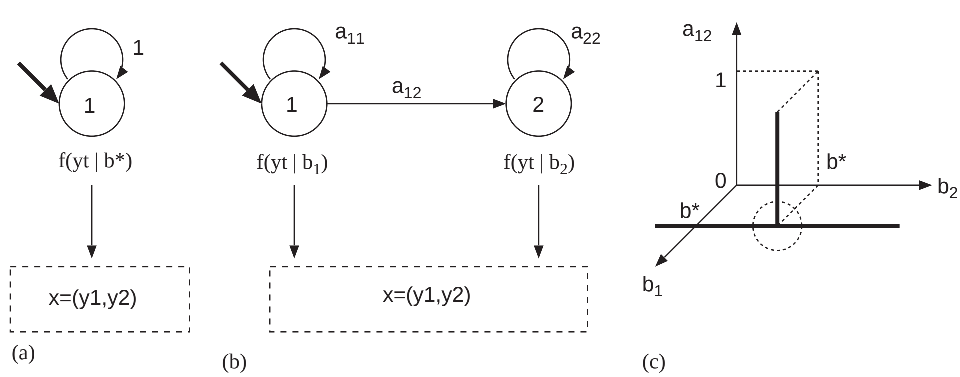

Fig. 1. Example: (a) the true has one hidden state; (b) the learner has two hidden states; (c) the singularity in the parameter space.