情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

代理ベイズ学習と隠れマルコフモデルへの応用
Vicarious Bayes Learning and its Application to HMMs

山崎　啓介*
Keisuke Yamazaki

Abstract: Hierarchical parametric models, such as Gaussian mixture models, Bayesian networks, and hidden Markov models, are widely used in the information engineering fields. These models are generally expressed as probability functions of the given data space, and there are a number of learning algorithms for each model. However, it is still unknown whether the space is suitable and effective for learning of the function. Therefore, the present paper considers a feature map to a different domain space, and investigates how the map changes the generalization error. Then, we proposed the *vicarious learning* in the Bayes estimation, which preserves the error value of the original space in a different space. This new learning framework reduces the computational learning and evaluation costs because a simpler space makes the calculation of the likelihood faster. As one of its applications, we can derive a necessary length of training data for HMMs.

Keywords: Bayes Learning, Feature Selection, Generalization Error

1 Introduction

Hierarchical parametric models, such as Gaussian mixture models, Bayesian networks and hidden Markov models, are used in a number of practical engineering fields. The parameter space of such models can have singularities due to the hierarchical structure or the latent variables. Then, these models are referred to as *singular*. Models are *regular* when the parameter space does not include any singularity.

The conventional statistical analysis is established on the basis of unique probabilities of the regular model. The analysis is not available for singular models. More statistically, the inverse of the Fisher information matrix is required to describe the convergence of the optimal parameter. The matrices are not positive definite on the singularities, which means that the inverse matrices do not exist. Therefore, the algebraic geometrical method has been developed to reveal the Bayesian generalization error of singular models [6]. Based on the method, the errors of the hierarchical models are revealed (e.g. [8, 7]).

The present paper focuses on a relation between the data space and the generalization error. To project the given data to a different space is a common technique in feature extraction and dimensionality reduction. Models dealing with sequential data have large computational cost for learning and evaluation even though they have effective algorithms [4, 3, 2]. A feature map actually seems to reduce the cost because the models can be simplified in the feature space. However, effect of the map on the error has not been studied yet.

The feature space has to be designed properly. Let us consider the following simple example: Discrete data x are assumed to be D dimensional binary vectors. Then, the dimension of the data space is 2^D. The most naive modeling is to provide parameters as probability variables of each x. In this case, $2^D - 1$ variables are required to represent all x. For example, three variables p_1, p_2, p_3 are sufficient for $D = 2$, i.e. $P(00) = p_1, P(01) = p_2, P(10) = p_3$. Note that $P(11) = 1 - p_1 - p_2 - p_3$. A parametric model $p(x|w)$, where w is the parameter, generally has to have less number of parameters than $2^D - 1$. Now, a feature map projects the data to 2^d dimensional space. The feature

*東京工業大学　精密工学研究所, 〒 226-8503 横浜市緑区長津田 4259 R2-5, e-mail k-yam@pi.titech.ac.jp,
Tokyo Institute of Technology, R2-5, 4259 Nagatuta, Midori-ku Yokohama

space is assumed to be much smaller than the original one, $d \ll D$. It is easy to prove that the parameter w cannot be correctly identified when $2^d - 1 < \dim w$. Thus, a theoretical evaluation of a feature map is required to clarify if the feature space is suitable for the parameter learning.

Based on the algebraic geometrical method, the present paper proposes an evaluation method the feature map. The main purpose is to design the model $p(x|w)$ through the feature space. Therefore, one of the expected applications is the model selection with the cross-validation [5]. The selected model will be finally used in the original space. Our task is to design the optimal model available not for the feature space, but for the original one. Then, the feature space preserving the generalization error is desired because the parameters can be completely estimated. The present paper defines training and test procedures in such feature space as the *vicarious Bayes learning*, and the feature map as the *vicarious feature map*. As for the demonstration, the vicarious feature map will be found for hidden Markov models, and a necessary length of data sequences is derived for the complete learning.

The remainder of the paper is organized as follows. Section 2 formalizes the Bayes learning and summarizes important results of the algebraic geometrical method. Section 3 proposes the vicarious Bayes learning. Section 4 shows an application to hidden Markov models (HMMs). Sections 5 and 6 present discussions and our conclusion, respectively.

2 The Bayes Learning and the Algebraic Geometrical Method

Let us formally define the generalization error. A set of training data $X^n = \{X_1, \ldots, X_n\}$ is independently and identically distributed from the true model $q(x)$. The learning model with its parameter w is generative and is represented as $p(x|w)$. The generalization error is the average Kullback divergence from $q(x)$ to the predictive distribution $p(x|X^n)$,

$$G(n) = E_{X^n}\left[\int q(x) \ln \frac{q(x)}{p(x|X^n)} dx\right], \tag{1}$$

where n is the number of the training data and $E_{X^n}[\cdot]$ represents the expectation value over all training samples. The predictive distribution is constructed by $p(x|w)$. For example, the maximum likelihood method gives

$$p(x|X^n) = p(x|\hat{w}) \tag{2}$$

where $\hat{w}$ is the maximum likelihood estimator:

$$\hat{w} = \arg\max_w L(w, X^n), \tag{3}$$

$$L(w, X^n) = \prod_{i=1}^{n} p(X_i|w). \tag{4}$$

The Bayes estimation yields the predictive distribution

$$p(x|X^n) = \int p(x|w)p(w|X^n)dw, \tag{5}$$

where the posterior $p(w|X^n)$ is defined by

$$p(w|X^n) = \frac{1}{Z(X^n)} L(w, X^n)\varphi(w) \tag{6}$$

using a prior $\varphi(w)$ and the normalization factor $Z(X^n)$.

The asymptotic form of Eq.(1) is expressed as

$$G(n) = \frac{\lambda}{n} - \frac{m-1}{n \ln n} + o\Big(\frac{1}{n \ln n}\Big) \tag{7}$$

when $W_t \equiv \{w : p(x|w) = q(x)\} \neq \emptyset$, i.e. $p(x|w)$ can attain $q(x)$ [6]. The coefficients are defined as follows: All poles of the zeta function

$$J(z) = \int H(w)^z \varphi(w) dw \tag{8}$$

are real negative and rational, where the Kullback divergence

$$H(w) = \int q(x) \ln \frac{q(x)}{p(x|w)} dx \tag{9}$$

is analytic. Then, $z = -\lambda$ is the largest pole and m is its order.

Eqs (7)-(9) shows that the generalization error is determined by the relation between $q(x)$ and $p(x|w)$. Behavior of $H(w)$ in the neighborhood of W_t directly affects λ and m. For example, $J(z) = \int w^{2jz} dw$ ($j = 1, 2, \ldots$) when $H(w) = w^{2j}$ and $\varphi(w)$ is closed to uniform around $W_t = \{0\}$. It is easily found that $J(z)$ has a factor $1/(2jz + 1)$ by integrating over w, which implies $\lambda = 1/2j$.

3 Proposed Learning Framework

In this section, we propose the vicarious Bayes learning.

3.1 Mapping to a Feature Space

As can be noticed in Eqs (7)-(9), the error $G(n)$ depends on the probabilities $p(x|w)$ and $q(x)$. Herein, we consider the error value over a different domain space.

Let $\Phi : x \mapsto y$ be a feature map. Based on the map, the models on y are defined by

$$q_\Phi(y) = \int q(x)\delta(y - \Phi(x))dx, \tag{10}$$

$$p_\Phi(y|w) = \int p(x|w)\delta(y - \Phi(x))dx, \tag{11}$$

where $\delta(\cdot)$ is the Dirac's delta function. The likelihood function Eq (4) is given by

$$L_\Phi(w, Y^n) = \prod_{i=1}^{n} p_\Phi(Y_i|w), \tag{12}$$

where $Y^n = \{Y_1, \ldots, Y_n\} = \{\Phi(X_1), \ldots, \Phi(X_n)\}$. The Bayes estimation yields the posterior $p_\Phi(w|Y^n)$ and the predictive distribution $p_\Phi(y|Y^n)$ by replacing $p(x|w)$ of Eqs (5) and (6) with $p_\Phi(y|w)$. Then, the generalization error defined as

$$G_\Phi(n) = E_{Y^n}\left[\int q_\Phi(y) \ln \frac{q_\Phi(y)}{p_\Phi(y|Y^n)} dy\right], \tag{13}$$

has the asymptotic form, whose coefficients are determined by the zeta function of

$$H_\Phi(w) = \int q_\Phi(y) \ln \frac{q_\Phi(y)}{p_\Phi(y|w)} dy. \tag{14}$$

$\sum_x$ and/or $\sum_y$ should be substituted for $\int dx$ and/or $\int dy$ in the discrete space, respectively.

3.2 Vicarious Bayes Learning

Comparing Eq (13) with Eq (1), let us determine how the data space affects the parameter learning. It is known that the generalization error implicitly expresses the tuning cost of all essential parameters. For example, the regular model has the coefficients $\lambda = \dim w/2$ and $m = 1$. In singular models, λ also depends on the number of parameters to be tuned [6, 8]. The learning process is preserved if a feature map Φ does not change the error. Observing the error change, we can theoretically investigate which factor of the data space is necessary for the parameter learning.

We propose the following learning framework:

Definition 1 (Vicarious Bayes Learning) *The Bayesian parameter learning and its evaluation of the generalization error over a feature space are referred to as the vicarious Bayes learning when $G(n)$ and $G_\Phi(n)$ have the common asymptotic form.*

In the present paper, we refer to this preserving map as the *vicarious feature map*. Herein we are not interested in the case $G(n) > G_\Phi(n)$ because our purpose is to investigate model properties on the given space x. The smaller error implies that not all parameters have to be estimated in the feature space, which means that the proper form of $p(x|w)$ cannot be obtained.

3.3 A Theory of the Feature Map on the Error

A novelty of the proposed learning is to restrict the feature map Φ to the vicarious one in the asymptotic manner. Therefore, we study a condition, under which Φ becomes vicarious.

The coefficients of the asymptotic error is determined by behavior of the Kullback divergence $H(w)$ or $H_\Phi(w)$ in the neighborhood of W_t. The divergence is expressed as a polynomial form of w because it is analytic. Based on the Noetherian property of polynomial ring, the divergence consists of bases. For example, if polynomials $f_1(w)$ and $f_2(w)$ are bases, the non-negative function $H(w)$ can have the following form,

$$H(w) = f_1(w)^2 + f_2(w)^2 + f_3(w), \tag{15}$$

where f_3 is a sum of squared polynomials with respect to f_1 and f_2, and is higher order than f_1^2 and f_2^2. More precisely, f_3 consists of terms, such as $(f_1 f_2)^2$, $(f_1 + f_2^2)^2$ and $(f_1^2 + f_2)^2$, with coefficients. It naturally holds that $f_1(w) = f_2(w) = 0$ on W_t. The error $G_\Phi(n)$ will have the same asymptotic form if $H_\Phi(w)$ is given by

$$H_\Phi(w) = f_1(w)^2 + f_2(w)^2 + f_4(w), \tag{16}$$

where f_4 consists of the same terms as f_3 with different coefficients. Based on this example, we derive a condition:

Theorem 1 *A feature map Φ becomes vicarious if $H(w)$ and $H_\Phi(w)$ have the same essential terms for λ and m.*

This theorem is easily proved according to the relation of Eqs (7)-(9). Note that Theorem 1 shows a sufficient condition: The largest poles in the zeta functions of

$H(w)$ and $H_\Phi(w)$ can be at the same position even if the essential parts are different from each other.

A general feature map provides insight of the generalization error even when it is not straightforward to find the vicarious one.

Theorem 2 *For a feature map Φ, it holds that*

$$G(n) \geq G_\Phi(n). \tag{17}$$

The proof is in Appendix.

Theorem 2 intuitively shows that the learning in a feature space is more accurate than in the original space because the domain space y is generally simplified based on the definition of $p_\Phi(y|w)$. For example, let us divide elements of vector x into two sets x_1 and x_2, i.e. $x = (x_1, x_2)$, and define two feature maps as $\pi_i : (x_1, x_2) \mapsto x_i$ for $i = 1, 2$. The map π_i selects the attribute x_i. The original model $p(x|w)$ is a joint probability of x_1 and x_2. According to the definition, $p_{\pi_i}(x_i|w)$ is a marginal probability of x_i. Then, $G(n)$ measures the error over x_1 and x_2 whereas $G_{\pi_i}(n)$ is about only x_i. This derives that $G_{\pi_i}(n)$ should be smaller than $G(n)$. Theorem 2 claims this fact mathematically.

A vicarious feature map purifies the domain space when the dimension of the feature space is less than the original one. If π_1 is a vicarious feature map, the attributes in x_1 are essential for the parameter learning and x_2 is nuisance dimension. Considering the map π_i, we can regard the vicarious learning as a feature selection in a Bayes scenario.

4 An Application to HMMs

In this section, we apply the vicarious Bayes learning to HMMs. We show that a restriction map of a data length can be a vicarious feature map, which derives a necessary length for the parameter learning.

4.1 Model Setting

The present paper focuses on the ergodic HMMs, in which the transition connections among the hidden states construct a complete graph. The number of output alphabets and the length of data are $M + 1$ and L_0, respectively, i.e. $x \in \{1, \ldots, M + 1\}^{L_0}$. The numbers of the hidden states are $K + 1$ and $K_0 + 1$ in the learning and true models. For simplicity, the initial state is always the first one in both models. The parameter w includes the transition probability a_{ij}, indicating the probability of transition from the ith state to the jth state, and the output probability b_{im}, indicating the probability of generating alphabet m at the ith hidden state. These probabilities satisfy the conditions, $a_{ii} = 1 - \sum_{i \neq j}^{K+1} a_{ij}$ for $\forall a_{ij} \geq 0$ and $b_{iM+1} = 1 - \sum_{m=1}^{M} b_{im}$ for $\forall b_{im} \geq 0$.

4.2 Restriction Maps on the Length

We consider the following map of the data space,

$$\Phi_L : \{1, \ldots, M + 1\}^{L_0} \to \{1, \ldots, M + 1\}^L \tag{18}$$

for $L \leq L_0$, which cuts off the data sequences to change the length into L. For example, $\Phi_3(8145924518) = 484$ when $L_0 = 10$, $L = 3$ and $M = 8$. In the similar way to π_i, $p_{\Phi_3}(814|w)$ is the marginal probability over joint probabilities $p(814 * * * * * * * |w)$, where '$*$' means any number from $\{1, \ldots, 9\}$.

4.3 Necessary Length for the Vicarious Bayes Learning

Using the restriction map, we consider a necessary length for the parameter learning.

It can be conjectured that $L = 3$ is sufficient for the learning according to the following reason: The process to generate data with $L = 3$ includes two transitions. All transition parameters a_{ij} are used for two transitions in the complete graph. All output parameters b_{im} are used for generating data at all hidden states. Therefore, the information of all parameters w can be extracted from the output sequences when infinitely large number of sequences are given as training data.

We hereinafter prove that $L = 3$ is not enough even when $n \to \infty$ and derive a necessary length.

Lemma 1 *The Kullback divergence $H_{\Phi_L}(w)$ is expressed as a sum of squared terms, which follows the rules:*

1. *The squared terms monotonically increase with growth of the data length. More precisely, $H_{\Phi_{L2}}(w)$ includes all the terms of $H_{\Phi_{L1}}(w)$ for $L1 \leq L2$.*
2. *The significant number of the squared terms $NST(L, M)$ can be expressed as*

$$NST(L, M) = (M + 1)^{L-1} + M - 1. \tag{19}$$

The proof is in Appendix.

Theorem 3 *For sufficiently large L_0, vicarious feature maps exist in the series of Φ_L.*

Proof: Based on Lemma 1, $NST(L, M)$ monotonically increases and all squared terms in $H_{\Phi_{L1}}(w)$ are included by those in $H_{\Phi_{L2}}(w)$ for $L1 < L2$. We consider the series of the squared terms of $H(w) = H_{\Phi_{L_0}}(w)$. Let S_L be a set of the squared terms $f(w)^2$ defined by

$$S_L = \{f(w)^2 : f(w)^2 \in H_{\Phi_L}(w), f(w)^2 \notin H_{\Phi_{L-1}}(w)\}, \tag{20}$$

where $S_0 = \emptyset$. The order of the series is given by

$$S_1, S_2, \ldots, S_L, \ldots S_{L_0}. \tag{21}$$

The Noetherian property on $H(w)$ shows that there is a constant $L_e < L_0$, which is the number of the essential squared terms for $H_{\Phi_L}(w) = 0$, since L_0 is sufficiently large for the parameter learning. More mathematically, L_e is the minimum number, such that S_{L_e} includes all generating elements for $H(w) = 0$ in terms of the ideal theory on polynomial ring. Then Φ_L for $L_e \leq L < L_0$ is vicarious. **(End of Proof)**

Let $\dim w$ be the dimension of the parameter in the learning model. In the present HMMs,

$$\dim w = (K + 1)(K + M). \tag{22}$$

Comparing Eq (22) with Eq (19), the following theorem indicates a necessary length:

Theorem 4 *A necessary length L_m for the vicarious Bayes learning of HMMs are represented by*

$$L_m = \arg\min_L \{NST(M, L) \geq \dim w\}. \tag{23}$$

Proof: The maximum number of the parameters to be tuned is $\dim w$. According to Eqs (7)-(9), we focus on the case $p_{\Phi_L}(y|w) = q_{\Phi_L}(y)$, i.e., $H_{\Phi_L}(w) = 0$. In the case, the squared terms are all zero, where each term is a polynomial of w. To identify all elements of w, the number of the polynomials $NST(M, L)$ should not be less than $\dim w$ based on the relation between the number of variables and that of equations. **(End of Proof)**

Theorem 4 shows that $L = 3$ can not attain the vicarious Bayes learning. For example, let us assume that $L = 3$, $M = 1$ and $K = 1$.

$$NST(1, 3) = (1 + 1)^2 + 1 - 1 = 4 < \dim w = (1 + 1)(2 + 1) = 6, \tag{24}$$

which does not satisfy the condition of L_m. The necessary length is derived as

$$L_m = \arg\min_L \{(1 + 1)^{L-1} + 1 - 1 \geq 6\} = 4. \tag{25}$$

Note that this length could not be sufficient, i.e. the vicarious Bayes learning requires longer sequences. Let us assume that $H_{\Phi_L}(w)$ consists of the square terms $f_1(w)^2$, $f_2(w)^2$ and $f_3(w)^2$, and that $\dim w = 3$. If $f_3(w)$ is a polynomial with respect to f_1 and f_2, $f_1(w) = f_2(w) = 0$ automatically satisfies $f_3(w) = 0$. Then the actual number of the equations to identify w decreases to two, which is less than $\dim w$. In such case, L should be larger to obtain more squared terms in $H_{\Phi_L}(w)$.

4.4 Experimental Validation of the Minimum Length

As seen in the previous part, L_m is a necessary length, which implies that longer sequences are required for the vicarious Bayes learning. In this part, we experimentally verify the minimum (necessary and sufficient) length.

We suppose that $(K, M) = (2, 2)$ and investigate the generalization error when $L = 1, \ldots, 10$. The original length is $L_0 = 10$. The dimension of parameters is $\dim w = 12$ and the number of the squared terms is $NST(L, 2) = 3^{L-1}+1$. According to Theorem 4, $L_m = 4$. The number of training data sequences is $n = 500$. We used the MCMC method to construct the posterior $p_{\Phi_L}(w|Y^n)$ [1]. The number of parameters to construct the predictive model $p_{\Phi_L}(y|Y^n)$ is $N_w = 500$, i.e.

$$p_{\Phi_L}(y|Y^n) \simeq \frac{1}{N_w} \sum_{i=1}^{500} p_{\Phi_L}(y|w_i), \tag{26}$$

where w_i is taken from the posterior. In the evaluation, the number of the test data sequences is $N = 5000$ and the average $E_{Y^n}[\cdot]$ is taken by 100 training sets.

Figure 1 shows the results. The horizontal axis indicates the lengths L of training and test data. The vertical one does the generalization error. There are three curves for the true model sizes $K_0 = 0, 1, 2$. It is

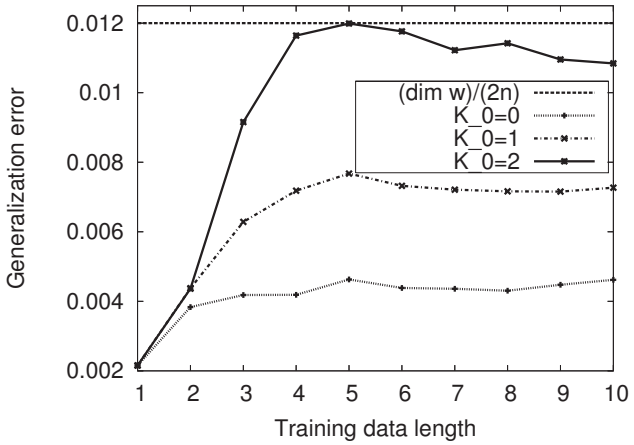

図 1: The generalization error w.r.t. the sequence length.

derived that

$$G(n) = \frac{\dim w}{2n} + O\Big(\frac{1}{n^2}\Big) \tag{27}$$

when $K = K_0$ [6]. The horizontal line described as '$(\dim w)/(2n)$' is the theoretical asymptotic value of the error, $\dim w/(2n) = 12/(2 \times 500) = 0.012$. The error on $K_0 = 2$ has to reach the line for $G(n) = G_{\Phi_L}(n)$. As can be seen in the graph, the curve of $K_0 = 2$ has a gap between $L = 3$ and $L = 4$, and almost reaches the horizontal line at $L = 4$, which implies that $L_m = 4$ will be the minimum length of the vicarious learning.

5 Discussions

The vicarious learning on HMMs reduces the computational cost in both training and test sessions. It is known that the time complexity of $p(x|w)$ is $O((K+1)^2 L)$ in HMMs. Then, the complexity of the likelihood $L(w, X^n)$ is $O(n(K+1)^2 L_0)$ for a given w, which shows the cost $O(n(K+1)^2 L)$ of $L_{\Phi_L}(w, Y^n)$ is much less than that of $L(w, X^n)$ when a number of data exist. The MCMC method requires the calculation of $L(w, X^n)$ in each update of the parameter for training. The computation of $p(x|w)$ is frequently used for the generalization error in the test session. Therefore, to shorten the length of the data sequences saves a number of computational complexity.

The reduction of the computational cost is also effective for cross-validation [5]. In the validation, the given data is divided into training data and validation data. These data sets are used for training and testing respectively. This procedure is then repeated after reversing the roles of the sets, and the generalization error is estimated. This validation method is commonly used for selecting the optimal size of a model, so called the model selection problem. To change the domain space by the vicarious feature map Φ_L for $L \geq L_m$ is a powerful method for the model selection because both the training and testing sessions are enormously repeated in the cross-validation and $G_{\Phi_L}(n)$ has the same value as $G(n)$ in any size.

6 Conclusion

We proposed the vicarious Bayes learning. It is regarded as a feature selection preserving the Bayes generalization error in an asymptotic manner. We also demonstrated its availability in HMMs. The length restriction is a theoretically guaranteed feature selection on the basis of the vicarious learning framework.

Acknowledgement

This research was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 18079007.

Appendix

Proof of Theorem 2

First, let us state the following lemmas without their proofs:

Lemma 2 *Let λ_1, m_1 and λ_2, m_2 be the largest poles and the orders in the zeta functions of $H_1(w)$ and $H_2(w)$, respectively. Then, the following relation holds if $H_1(w) \leq H_2(w)$ for all w [8],*

$$\frac{\lambda_1}{n} - \frac{m_1 - 1}{n \ln n} \leq \frac{\lambda_2}{n} - \frac{m_2 - 1}{n \ln n}. \tag{28}$$

Lemma 3 (Continuous Log-Sum Inequality) *For $\int\int q(x,y)p(y) \ln \frac{q(x,y)}{p(x,y)} dxdy < \infty$,*

$$\int\int q(x,y)p(y) \ln \frac{q(x,y)}{p(x,y)} dxdy \geq \int \Big(\int q(x,y)p(y)dy \Big) \ln \frac{\int q(x,y)p(y)dy}{\int p(x,y)p(y)dy} dx, \tag{29}$$

where $p(y), q(x,y), p(x,y)$ are all probability density functions.

Eq (9) is rewritten as

$$H(w) = \int H_y(w) dy, \tag{30}$$

$$H_y(w) = \int q(x)\delta(y - \Phi(x)) \ln \frac{q(x)}{p(x|w)} dx. \tag{31}$$

Based on Lemma 3,

$$H(w) \geq H_\Phi(w). \tag{32}$$

Lemma 2 indicates that

$$G(n) \geq G_\Phi(n), \tag{33}$$

which proves the theorem. **(End of Proof)**

Proof of Lemma 1

First, we focus on $L = 3$, then generalize the result. We use the following notation:

$$H(w) \geqslant K(w), \tag{34}$$

where there are positive constants C_1, C_2 such that

$$C_1 K(w) \leq H(w) \leq C_2(w) \tag{35}$$

in the neighborhood of $H(w) = 0$. The true model has the true parameter $w^* = \{\{a^*_{ij}\}, \{b^*_{im}\}\}$ for $1 \leq i, j \leq K_0 + 1$ and $1 \leq m \leq M + 1$. It is known that the largest poles in the zeta functions of $H(w)$ and $K(w)$ are the same [8]. It holds that

$$H_{\Phi_3}(w) \geqslant \sum_y \{p_{\Phi_3}(y|w) - q_{\Phi_3}(y)\}^2 \tag{36}$$

because y is discrete [7]. To simplify the descriptions, we use the notation,

$$\sum_{alph} \sum_{path} b_1 abab = \sum_{i,j,k \in \{1,\ldots,M+1\}^3} \sum_{l,m \in \{1,\ldots,K+1\}^2} b_{1i} a_{1l} b_{lj} a_{lm} b_{mk}, \tag{37}$$

which represents that $\sum_{alph}$ and $\sum_{path}$ is the marginalization over all generation of alphabets and all paths of transitions among the hidden states. For the true parameters $\{\{a^*_{ij}\}, \{b^*_{im}\}\}$,

$$\sum_{path} \equiv \sum_{l,m \in \{1,\ldots,K_0+1\}^2}. \tag{38}$$

Then, Eq (36) is rewritten as

$$H_{\Phi_3}(w) \geqslant \sum_{alph} \Big\{ \sum_{path} b_1 abab - \sum_{path} b^*_1 a^* b^* a^* b^* \Big\}^2. \tag{39}$$

Using $b_{iM+1} = 1 - \sum_{m=1}^M b_{im}$,

$$\begin{aligned} &\Big\{ \sum_{path} b_1 abab_{\cdot M+1} - \sum_{path} b^*_1 a^* b^* a^* b^*_{\cdot M+1} \Big\}^2 \\ =&\Big\{ \sum_{path} b_1 aba(1 - \sum_{m=1}^M b_{\cdot m}) \\ &- \sum_{path} b^*_1 a^* b^* a^* (1 - \sum_{m=1}^M b^*_{\cdot m}) \Big\}^2. \end{aligned} \tag{40}$$

Note that the right-hand side of Eq (39) includes terms

$$\Big\{ \sum_{path} b_1 abab_{\cdot m} - \sum_{path} b^*_1 a^* b^* a^* b^*_{\cdot m} \Big\}^2 \tag{41}$$

for $1 \leq m \leq M$. For any constant c, it holds that

$$h_1(w)^2 + \{c h_1(w) + h_2(w)\}^2 \geqslant h_1(w)^2 + h_2(w)^2. \tag{42}$$

Combining Eq (42) and the presence of the terms in Eq (41), the term in Eq (40) is rewritten as

$$\begin{aligned} &\Big\{ \sum_{path} b_1 abab_{\cdot M+1} - \sum_{path} b^*_1 a^* b^* a^* b^*_{\cdot M+1} \Big\}^2 \\ &+ \sum_{m=1}^M \Big\{ \sum_{path} b_1 abab_{\cdot m} - \sum_{path} b^*_1 a^* b^* a^* b^*_{\cdot m} \Big\}^2 \\ \geqslant&\Big\{ \sum_{path} b_1 aba - \sum_{path} b^*_1 a^* b^* a^* \Big\}^2 \\ &+ \sum_{m=1}^M \Big\{ \sum_{path} b_1 abab_{\cdot m} - \sum_{path} b^*_1 a^* b^* a^* b^*_{\cdot m} \Big\}^2. \end{aligned} \tag{43}$$

Based on $\sum_{i=1}^{K+1} a_{\cdot i} = 1$, $\sum_{path} b_1 aba = \sum_{path} b_1 ab$. Then the term in Eq (43) is rewritten as

$$\begin{aligned} &\Big\{ \sum_{path} b_1 aba - \sum_{path} b^*_1 a^* b^* a^* \Big\}^2 \\ =&\Big\{ \sum_{path} b_1 ab - \sum_{path} b^*_1 a^* b^* \Big\}^2. \end{aligned} \tag{44}$$

Applying this procedure to b_{iM+1} of the last factor recursively, we can eliminate the b_{iM+1}, i.e.

$$\begin{aligned} H_{\Phi_3}(w) \geqslant &\sum_{alph}^M (b_1 - b^*_1)^2 + \sum_{alph}^M \{ \sum_{path} b_1 ab - \sum_{path} b^*_1 a^* b^* \}^2 \\ &+ \sum_{alph}^M \{ \sum_{path} b_1 aab - \sum_{path} b^*_1 a^* a^* b^* \}^2 \\ &+ \sum_{alph}^M \{ \sum_{path} b_1 abab - \sum_{path} b^*_1 a^* b^* a^* b^* \}^2, \end{aligned} \tag{45}$$

where $\sum_{alpha}^{M}$ means the marginalization over all generation of alphabets except for $M+1$. We apply a map $b_{1m} = b'_{1m} - b^*_{1m}$ to $H_{\Phi_3}(w)$. For simplicity, we use the same symbol b_{1m} for b'_{1m}.

$$H_{\Phi_3}(w) \geqslant \sum_{alph}^{M} b_1^2 + \sum_{alph}^{M}\{\sum_{path}(b_1+b_1^*)ab - \sum_{path} b_1^* a^* b^*\}^2 + \sum_{alph}^{M}\{\sum_{path}(b_1+b_1^*)aab - \sum_{path} b_1^* a^* a^* b^*\}^2 + \sum_{alph}^{M}\{\sum_{path}(b_1+b_1^*)abab - \sum_{path} b_1^* a^* b^* a^* b^*\}^2. \quad (46)$$

According to Eq (42),

$$H_{\Phi_3}(w) \geqslant \sum_{alph}^{M} b_1^2 + \sum_{alph}^{M}\{\sum_{path} b_1^* ab - \sum_{path} b_1^* a^* b^*\}^2 + \sum_{alph}^{M}\{\sum_{path} b_1^* aab - \sum_{path} b_1^* a^* a^* b^*\}^2 + \sum_{alph}^{M}\{\sum_{path} b_1^* abab - \sum_{path} b_1^* a^* b^* a^* b^*\}^2. \quad (47)$$

Since b^*_{1m} is a constant,

$$H_{\Phi_3}(w) \geqslant \sum_{alph}^{M} b_1^2 + \sum_{alph}^{M}\{\sum_{path} ab - \sum_{path} a^* b^*\}^2 + \sum_{alph}^{M}\{\sum_{path} aab - \sum_{path} a^* a^* b^*\}^2 + \sum_{alph}^{M}\{\sum_{path} abab - \sum_{path} a^* b^* a^* b^*\}^2. \quad (48)$$

The number of terms in $\sum_{alph}^{M}\{\cdot\}^2$ is $M^{\#b}$, where $\#b$ is the number of b_{im} because only the output probability b_{im} is counted in $\sum_{alph}^{M}$. Then $NST(M,3) = M+M+M+M^2 = (M+1)^{3-1} + M - 1$.

Hereinafter, let us generalize the proof for any L. Because of the reduction procedure, the right-hand side of Eq (48) includes the terms

$$\sum_{alph}^{M} b_1^2, \text{ and } \sum_{alph}^{M}\{\sum_{path} ab - \sum_{path} a^* b^*\}^2, \quad (49)$$

which are the term of $L = 2$. Moreover, the first term appears when $L = 1$. This shows that the squared terms of $L = l$ includes those of $L < l$, which proves Lemma 1-1.

We count the squared terms. We define that the squared terms in Eqs (39) and Eq (48) are $ST1_L$ and $ST2_L$, respectively. Let $\#(\cdot)$ be a function to indicate the number of squared terms. Obviously $\#(ST1_L) = (M+1)^L$. The suffix on the summation $\sum_{path}$ does not affect the number of the terms, which implies that the number of the squared terms is determined by the number of b_{im} in $\sum_{alph}$. Let us assume that $b_{\cdot M+1} = 1$ in Eq (39) in order not to count it as a parameter. Under this assumption, $\#(ST1_L) = (M+1)^L - 1$ because the term for the alphabet $y = (M+1)(M+1)(M+1)$ in Eq (39) vanishes i.e. $\{\sum_{path} aa - \sum_{path} a^* a^*\}^2 = 0$. Then, the numbers of $\sum_{alph}$ and b_{im} in $ST1_{L-1}$ are completely the same as those of $\sum_{alph}^{M}$ and b_{im} in $ST2_L$ except for the first term $\sum_{alph}^{M} b_1^2$. Therefore it is easy to confirm that

$$\#(ST2_L) = \#(ST1_{L-1}) + \#(\sum_{alph}^{M} b_1^2) = (M+1)^{L-1} - 1 + M, \quad (50)$$

which proves Lemma 1-2. **(End of Proof)**

参考文献

[1] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

[2] R. E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Engineering*, 82:35–45, 1960.

[3] Karim Lari and Steve J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56, 1990.

[4] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(22):257–2862, 1989.

[5] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–147, 1974.

[6] S. Watanabe. Algebraic analysis for non-identifiable learning machines. *Neural Computation*, 13 (4):899–933, 2001.

[7] K. Yamazaki, M. Aoyagi, and S. Watanabe. Asymptotic analysis of Bayesian generalization error with Newton diagram. *Neural Networks*, to appear.

[8] Keisuke Yamazaki and Sumio Watanabe. Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing*, 69(1-3):62–84, 2005.