Properties of Jeffreys Mixture for Markov Sources

Jun'ichi Takeuchi, *Member, IEEE*, Tsutomu Kawabata, *Member, IEEE*, and Andrew R. Barron, *Fellow, IEEE*

***Abstract*—We discuss the properties of Jeffreys mixture for a Markov model. First, we show that a modified Jeffreys mixture asymptotically achieves the minimax coding regret for universal data compression, where we do not put any restriction on data sequences. Moreover, we give an approximation formula for the prediction probability of Jeffreys mixture for a Markov model. By this formula, it is revealed that the prediction probability by Jeffreys mixture for the Markov model with alphabet $\{0,1\}$ is not of the form $(n_{x|s}+\alpha)/(n_s+\beta)$, where $n_{x|s}$ is the number of occurrences of the symbol x following the context $s\in\{0,1\}$ and $n_s=n_{0|s}+n_{1|s}$. Moreover, we propose a method to compute our minimax strategy, which is a combination of a Monte Carlo method and the approximation formula, where the former is used for earlier stages in the data, while the latter is used for later stages.**

***Index Terms*—Bayes code, Jeffreys prior, minimax regret, stochastic complexity, universal source coding.**

I. Introduction

We discuss the properties of Jeffreys mixture for a Markov model (a class of fixed ordered Markov chains) in the problem of sequential prediction and universal coding. We employ logarithmic regret (which has other names, e.g., coding regret and pointwise redundancy) as a performance measure and show that a modified Jeffreys mixture asymptotically achieves the minimax regret up to constant order. This provides a sense in which the modified Jeffreys mixture is one of the best prediction strategies. Moreover, it implies that the modified Jeffreys mixture achieves the *stochastic complexity* [17] for the class of Markov models, which has various statistical interpretations.

The primary motivation for this investigation is to provide a stochastic model that achieves the universal coding and predictive objectives, including the determination of a sequence of priors for which the corresponding mixtures (for coding) and posterior (for prediction) achieve the approximate minimax regret. This improves understanding of the exact minimax regret procedure (normalized maximum likelihood) as identified by Shtar'kov[18], which seemingly lacks such interpretation. The normalized maximum likelihood distribution and our asymptotically minimax regret mixture distribution are close to each other in total relative entropy. We are extending a line of work in [6], [7], [27] and [28] which were for i.i.d. models, the work in [21] which was for exponential families, and the work in [1] which investigated the regret of Jeffreys mixture for Markov sources for sequences for which the maximum likelihood estimates (MLEs) (the relative frequencies of transition) are located away from zero. This study extends these conclusions to obtain results for regret that are uniformly valid over all sequences.

Although in the i.i.d. case, the Jeffreys mixture corresponds to the Dirichlet $(1/2,\ldots,1/2)$ prior which produces a Laplace-like Jeffreys prediction rule (also called the Krichevsky–Trofimov estimator), in the Markov case the Jeffreys prior does not correspond to independent Dirichlet priors on the transition probabilities for each context, so the corresponding rule is more complex.

The secondary motivation of our investigation is the calculation of the predictive probabilities needed for sequential prediction and universal coding algorithms. We propose an approximation formula in the form of a corrected Laplace estimator. The error of the correction is of order $1/n_s$, where n_s is the number of past occurrences of the current context (state) s. Moreover, we propose a method to compute approximately our minimax strategy, which is a combination of a Monte Carlo method and the approximation formula, where the former is used as long as n_s is not large for earlier stages in the data, while the latter is used once n_s becomes large.

Coding regret is defined as the difference of the loss incurred and the loss of an ideal coding or prediction strategy for each sequence. A coding scheme for sequences of length n is equivalent to a probability mass function $q(x^n)$ on $\mathcal{X}^n$ (the n-fold direct product of an alphabet $\mathcal{X}$). We can use q also for prediction, i.e., its conditionals $q(x_{i+1}|x^i)$ provide a distribution for the coding or prediction of the next symbol given the past. The minimax regret with respect to a family of probability mass functions $S=\{p(\cdot|\eta):\eta\in H\}$ is defined as

$$\min_q \max_{x^n:\hat{\eta}\in K} \max_{\eta\in H}\left(\log\frac{1}{q(x^n)}-\log\frac{1}{p(x^n|\eta)}\right)$$
$$=\min_q \max_{x^n:\hat{\eta}\in K}\left(\log\frac{1}{q(x^n)}-\log\frac{1}{p(x^n|\hat{\eta})}\right)$$

where $\hat{\eta}$ is the MLE of η given x^n. Restriction to a subset $K\subset H$ is used in some developments. Our main results are for the case that the maximum is taken over all strings x^n, i.e., $K=H$.

Here, the regret $\log(1/q(x^n))-\log(1/p(x^n|\hat{\eta}))$ in the data compression context is also called the pointwise redundancy: the difference between the codelength based on q and the minimum of the codelength $\log(1/p(x^n|\eta))$ achieved by distributions in the family. Also, $\log(1/q(x^n))-$

Manuscript received March 25, 2005; revised July 06, 2012; accepted July 23, 2012. Date of publication September 19, 2012; date of current version December 19, 2012. J. Takeuchi was supported in part by the Ministry of Education, Science, Sports, and Culture Grant-in-Aid for Scientific Research (B), 19300051 and (C), 24500018. This paper was presented in part at the 4th Workshop on Information Based Induction Sciences, Tokyo, Japan, 2001.

J. Takeuchi is with the Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan (e-mail: tak@inf.kyushu-u.ac.jp).

T. Kawabata is with the Graduate School of Informatics and Engineering, University of Electro-Communications, Tokyo 182-8585, Japan (e-mail: kawabata@ice.uec.ac.jp).

A. R. Barron is with the Department of Statistics, Yale University, New Haven, CT 06520 USA (e-mail: Andrew.Barron@yale.edu).

Communicated by Vinay A. Vaishampayan, Associate Editor At Large.

Digital Object Identifier 10.1109/TIT.2012.2219171

$\log(1/p(x^n|\eta))$ is the sum of the incremental regrets of prediction $\log(1/q(x_{i+1}|x^i)) - \log(1/p(x_{i+1}|x^i,\eta))$. For our Markov setting, the regret is defined conditionally on an initial state.

When S is the class of discrete memoryless sources, Xie and Barron [28] proved that the minimax regret asymptotically equals

$$\frac{d}{2}\log\frac{n}{2\pi} + \log\int_H \sqrt{\det J(\eta)}d\eta + o(1)$$

where d equals the size of the alphabet minus 1 and $J(\eta)$ is the Fisher information matrix with respect to η. This evaluation is not for a subset of sequences x^n but for the whole set of sequences. To obtain this asymptotically minimax regret, they use sequences of Bayes mixtures with prior distributions that weakly converge to the Jeffreys prior. The reason why one needs such variants of the Jeffreys prior is as follows: if we use the Jeffreys prior, the risk is asymptotically higher than the minimax value, for x^n such that $\hat{\eta}$ is near the boundary of H. They use priors which have higher density near the boundaries than the Jeffreys prior, to give more prior attention to these boundary regions and thereby pull the risk down to the asymptotically minimax level.

In this paper, we generalize the results of [28] to the case where S is a class of the kth-order Markov chains with alphabet size $d+1$. In particular, we give an upper bound on the minimax regret, using variants of the Jeffreys mixture, as

$$\frac{(d+1)^k d}{2}\log\frac{n}{2\pi} + \log\int_H \sqrt{\det J(\eta)}d\eta + o(1). \quad (1)$$

Note that $(d+1)^k d$ equals the number of the parameters of the class S. In [21], we showed that similar mixtures are minimax for (i.i.d.) exponential families and certain near exponential families that permit dependence, but in general those bounds are for the restricted set of sequences for which the MLE locates in a compact set interior to the parameter space. Our result is a generalization of [28] to Markov models and that of [21] to the set of all sequences. (Strictly speaking, the first-order Markov chain with alphabet size 2 is treated in [26]). Concerning Markov models, Atteson [1] obtained both pointwise regret and expected redundancy bound for Jeffreys mixtures with parameter values away from the boundary. Also, Gotoh *et al.* [8] gave an asymptotic upper bound on the regret, which holds almost surely.

It should be noted that *the normalized maximum likelihood* $p(x^n|\hat{\eta})/\sum_{x^n} p(x^n|\hat{\eta})$, proposed by Shtar'kov [18], provides the precise minimax procedure for pointwise regret. In [18], Shtar'kov introduced the pointwise regret and gave upper bounds on the codelength of normalized maximum likelihood for classes of discrete memoryless sources and finite state machines (FSMX model [24], which is an extension of Markov chains). His bound for the FSMX model yields a bound for Markov chain as $((d+1)^k d/2)\log n + C$, where C is a constant depending only on d and k. More recently, Jacquet and Szpankowski [11] evaluated it more precisely and determined the constant term of the minimax regret for the Markov chains. Modified to condition on the initial state (or initial string $x_{-k}\dots x_0$), their evaluation coincides with the form (1) in terms of Fisher information as explained in [20].

Rissanen's stochastic complexity [17] is the codelength having the minimax coding regret. It is used as the main part of model selection criteria by the minimum description length principle. A consequence of this study is that this criterion is approximately a Bayes criterion with modified Jeffreys prior.

To summarize, 1) we show that our modified Jeffreys mixture is asymptotically minimax; 2) consequently, the divergence between this mixture and the normalized maximum likelihood tends to 0 as n goes to infinity; 3) it provides the expression for the stochastic complexity exhibiting the role of the Fisher information; and moreover 4) the expression (1) for the minimax regret holds, even though we do not put any restriction on the sequences.

The Jeffreys mixture for the Bernoulli model induces the Laplace-like estimator $(k+1/2)/(n+1)$ where n is the data size and k is the number of occurrences of the symbol 1. While the Laplace estimator is in a very simple form, the Jeffreys mixture for a Markov model is not, even when the model is the first-order Markov chain. Hence, we give an approximation formula for the prediction probability of Jeffreys mixture for Markov models. This is an extension of the approximation formulas of the Bayes estimator for (i.i.d.) exponential families, shown in [19]. We can see the behavior of Jeffreys mixture by this formula. In particular, the prediction probability by Jeffreys mixture for the first-order Markov chain with alphabet $\{0,1\}$ is not of the Laplace-like form.

II. Preliminaries

Define an alphabet as $\mathcal{X} \stackrel{\text{def}}{=} \{0,1,..,d\}$, and let $\mathcal{X}'$ denote $\{1,2,\dots,d\}$. In this paper, we employ the class of kth-order Markov chains on the alphabet $\mathcal{X}$ as a parametric model. Let L denote $\mathcal{X}^k$ and let $\ell = |L|$. Listing the elements of L by dictionary order, denote $L = \{s_1, s_2, \dots, s_{(d+1)^k}\}$ (e.g., $s_1 = 00\dots0$). We refer to $s \in L$ as a context. For each context $s \in L$, let $\eta_{y|s}$ denote the probability that $y \in \mathcal{X}$ occurs after $s \in L$. So it is assumed that $\sum_{x\in\mathcal{X}} \eta_{x|s} = 1$ and $\eta_{x|s} \geq 0$. Let $\boldsymbol{\eta}_s$ denote the vector $(\eta_{1|s},\dots,\eta_{d|s})^t$ and $\boldsymbol{\eta}$ the vector $(\boldsymbol{\eta}_{s_1}^t, \boldsymbol{\eta}_{s_2}^t, \dots, \boldsymbol{\eta}_{s_\ell}^t)$, where $\ell = (d+1)^k$. Here, $\boldsymbol{\xi}^t$ denotes the transposition of a vector $\boldsymbol{\xi}$. Define the range of $\boldsymbol{\eta}_s$ as

$$H_s \stackrel{\text{def}}{=} \{\boldsymbol{\eta}_s : \forall x \in \mathcal{X}', \eta_{x|s} \geq 0 \text{ and } \sum_{x\in\mathcal{X}'} \eta_{x|s} \leq 1\}.$$

Likewise, the range of $\boldsymbol{\eta}$ is $H \stackrel{\text{def}}{=} \prod_{s\in L} H_s$. Let x_m^n denote a sequence $x_m x_{m+1}\dots x_n$ $(m \leq n)$ and x^n a sequence x_1^n. Note that $\eta_{0|s} = 1 - \sum_{x\in\mathcal{X}'} \eta_{x|s}$.

Assume that we have an initial context $s_0 = x_{-k+1}^0$ in advance. Let $n_{x|s}$ denote the number of occurrences of x as a direct successor of the context s in the sequence x_1^n given s_0, and define $n_s \stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}} n_{x|s}$. Denote the probability mass function for the sequence x^n, determined by $\boldsymbol{\eta}$, by $p(x^n|s_0,\boldsymbol{\eta})$. Let S denote the family of probability mass functions $S \stackrel{\text{def}}{=} \{p(\cdot|\cdot,\boldsymbol{\eta}) :$

$\boldsymbol{\eta} \in H\}$. We usually omit $s_0 = x_{-k+1}^0$ from $p(x^n|s_0, \boldsymbol{\eta})$ and simply denote it $p(x^n|\boldsymbol{\eta})$. Then, we have

$$\begin{aligned}\log p(x^n|\boldsymbol{\eta}) &= \sum_{t=0}^{n-1} \log \eta_{x_{t+1}|\tau(x_{-k+1}^t)} \\ &= \sum_{s\in L,\, x\in\mathcal{X}} n_{x|s} \log \eta_{x|s}\end{aligned} \tag{2}$$

where we let "log" denote the natural logarithm and τ denote the context function $\tau(x^t) \stackrel{\text{def}}{=} x_{t-k+1}^t$ (the last k symbols of x_{-k+1}^t) for $t = 0, 1, \ldots, n-1$. Let $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}(x_{-k+1}^n)$ denote the MLE of $\boldsymbol{\eta}$ given x_{-k+1}^n. We have

$$\hat{\eta}_{x|s} = \hat{\eta}_{x|s}(x_{-k+1}^n) = \frac{n_{x|s}}{n_s}.$$

Here, we introduce the minimax and maximin regret in the Markov setting, where we fix the initial state $s_0 = x_{-k+1}^0$. Let $\mathcal{P}(W_n|\mathcal{X}^k)$ denote the set of all conditional probability mass functions on $W_n \subseteq \mathcal{X}^n$ given x_{-k+1}^0.

The conditional maximum regret given s_0 of $q \in \mathcal{P}(\mathcal{X}^n|\mathcal{X}^k)$ with respect to a family of conditional probability mass functions $S = \{p(x^n|x_{-k+1}^0, \boldsymbol{\eta}) : \boldsymbol{\eta} \in H\}$ and $W_n \subseteq \mathcal{X}^n$ is defined as

$$\begin{aligned}&\bar{r}_n(q, W_n|s_0) \\ &\quad\stackrel{\text{def}}{=} \sup_{x^n\in W_n} \Big(\log \frac{1}{q(x^n|s_0)} - \log \frac{1}{p(x^n|s_0, \hat{\boldsymbol{\eta}})}\Big).\end{aligned}$$

The conditional minimax regret given s_0 with respect to a family of probability mass functions S and a set of the sequences W_n is defined as

$$\begin{aligned}\bar{r}_n(W_n|s_0) &\stackrel{\text{def}}{=} \inf_{q\in\mathcal{P}(W_n|\mathcal{X}^k)} \bar{r}_n(q, W_n|s_0) \\ &= \inf_{q\in\mathcal{P}(W_n|\mathcal{X}^k)} \sup_{x^n\in W_n} \log \frac{p(x^n|s_0, \hat{\boldsymbol{\eta}})}{q(x^n|s_0)}.\end{aligned}$$

The conditional maximin regret given s_0 for a set of sequences W_n is defined as

$$\begin{aligned}&\underline{r}_n(W_n|s_0) \\ &\quad\stackrel{\text{def}}{=} \sup_{q\in\mathcal{P}(W_n|\mathcal{X}^k)} \inf_{r\in\mathcal{P}(W_n|\mathcal{X}^k)} E_{q(\cdot|s_0)} \log \frac{p(x^n|s_0, \hat{\boldsymbol{\eta}})}{r(x^n|s_0)} \\ &\quad= \sup_{q\in\mathcal{P}(W_n|\mathcal{X}^k)} E_{q(\cdot|s_0)} \log \frac{p(x^n|s_0, \hat{\boldsymbol{\eta}})}{q(x^n|s_0)}\end{aligned}$$

where we let $E_{q(\cdot|s_0)}$ denote the conditional expectation with respect to q given $s_0 = x_{-k+1}^0$. As the consequence of the definitions, $\bar{r}_n(W_n|s_0) \geq \underline{r}_n(W_n|s_0)$ holds. For logarithmic regret, it can be shown that $\bar{r}_n(W_n|s_0) = \underline{r}_n(W_n|s_0)$ holds in the same manner as in [18] and [28].

Now we introduce the Fisher information and empirical Fisher information. Empirical Fisher information is the Hessian of $-(1/n)\log p(x^n|\boldsymbol{\eta})$. We denote its component with respect to $\eta_{x|s}$ and $\eta_{y|t}$ $(x, y \in \mathcal{X}')$, by $\hat{J}_{sx,ty}(\boldsymbol{\eta}) = \hat{J}_{sx,ty}(x^n, \boldsymbol{\eta})$. Then, one can derive from (2) that

$$\hat{J}_{sx,ty}(x^n, \boldsymbol{\eta}) = \delta_{st}\hat{p}_s\Big(\frac{\delta_{xy}\hat{\eta}_{x|s}}{(\eta_{x|s})^2} + \frac{\hat{\eta}_{0|s}}{(\eta_{0|s})^2}\Big) \tag{3}$$

where we let $\hat{p}_s = \hat{p}_s(x_{-k+1}^n) \stackrel{\text{def}}{=} n_s/n$ and let δ_{xy} and δ_{st} be Kronecker's delta. The Fisher information is defined as

$$\begin{aligned}J_{sx,ty}(\boldsymbol{\eta}) &= \lim_{n\to\infty} E_\eta \hat{J}_{sx,ty}(x^n, \boldsymbol{\eta}) \\ &= \delta_{st}\mu_s\Big(\frac{\delta_{xy}}{\eta_{x|s}} + \frac{1}{\eta_{0|s}}\Big)\end{aligned} \tag{4}$$

where $\mu_s = \mu_s(\boldsymbol{\eta})$ denotes the stationary probability of the state s determined by $p(\cdot|\boldsymbol{\eta})$, and the symbol E_η the expectation with respect to $p(\cdot|\boldsymbol{\eta})$.

Define the Jeffreys prior density with respect to the Lebesgue measure $d\boldsymbol{\eta} = \prod_{s\in L} d\eta_{1|s} d\eta_{2|s} \cdots d\eta_{d|s}$ as

$$\rho_J(\boldsymbol{\eta}) \stackrel{\text{def}}{=} \sqrt{\det J(\boldsymbol{\eta})}/C_J$$

where $C_J \stackrel{\text{def}}{=} \int_H \sqrt{\det J(\boldsymbol{\eta})}d\boldsymbol{\eta}$ is the normalization constant. Let $D_{(\alpha)}(\boldsymbol{\eta}_s) \stackrel{\text{def}}{=} \prod_{x\in\mathcal{X}}(\eta_{x|s})^{-(1-\alpha)}$ be the Dirichlet function. Then, from (4), we have

$$\rho_J(\boldsymbol{\eta}) = \frac{1}{C_J}\prod_{s\in L}\frac{\mu_s^{d/2}}{\sqrt{\prod_{x\in\mathcal{X}}\eta_{x|s}}} = \frac{\prod_{s\in L}\mu_s^{d/2}D_{(1/2)}(\boldsymbol{\eta}_s)}{C_J}. \tag{5}$$

Let m_J denote the mixture by ρ_J (Jeffreys mixture) which is $m_J(x^n|s_0) \stackrel{\text{def}}{=} \int_H p(x^n|s_0, \boldsymbol{\eta})\rho_J(\boldsymbol{\eta})d\boldsymbol{\eta}$. We also define the Dirichlet (α) prior density as

$$\rho_{(\alpha)}(\boldsymbol{\eta}) \stackrel{\text{def}}{=} \frac{\prod_{s\in L} D_{(\alpha)}(\boldsymbol{\eta}_s)}{(C_{(\alpha)})^\ell}$$

where $C_{(\alpha)} \stackrel{\text{def}}{=} \int D_{(\alpha)}(\boldsymbol{\eta}_s)d\boldsymbol{\eta}_s$. This is a product of Dirichlet prior densities, one for each context, reflecting independence *à priori* between the contexts. In contrast, $\rho_J(\boldsymbol{\eta})$ is not of product form because $\mu_s(\boldsymbol{\eta})$ depends on all of $\boldsymbol{\eta}$ for each s. Note that $\rho_{(\alpha)}(\boldsymbol{\eta})/\rho_J(\boldsymbol{\eta}) \to \infty$ holds as $\boldsymbol{\eta}$ approaches the boundaries of H, if $0 < \alpha < 1/2$ holds.

III. Results

A. Minimax Regret

We establish a tight upper bound on the minimax regret for Markov model by the following theorem.

Theorem 1: Let $S = \{p(\cdot|s, \boldsymbol{\eta})|\eta \in H, s \in L\}$ be a class of kth-order Markov chains with alphabet $\{0, 1, \ldots, d\}$. Define a modified Jeffreys prior density for H as

$$\rho_n \stackrel{\text{def}}{=} (1 - \kappa_n)\rho_J + \kappa_n\rho_{(\alpha)}$$

where $0 < \alpha < 1/2$ is assumed and $\kappa_n = ((\ell - 1)/n)^b$. Let m_n be a mixture of Markov sources as

$$m_n(x^n|s_0) \stackrel{\text{def}}{=} \int_H p(x^n|s_0, \boldsymbol{\eta})\rho_n(\boldsymbol{\eta})d\boldsymbol{\eta}.$$

Then, for an arbitrary $b : 0 < b < (1/2 - \alpha)/(k(2\ell - 1))$, the following bound on $\bar{r}_n(m_n) = \bar{r}_n(m_n, \mathcal{X}^n|s_0)$ holds for any $s_o \in L$:

$$\begin{aligned}&\bar{r}_n(m_n) \\ &\quad\leq \frac{\ell d}{2}\log\frac{n}{2\pi} + \log\int_H \sqrt{\det J(\boldsymbol{\eta})}d\boldsymbol{\eta} + o(1)\end{aligned} \tag{6}$$

where $o(1)$ converges to 0 as n goes to infinity.

The complete proof is given in Section IV, but we give the intuition here.

The main tool for the proof is the Laplace approximation, by which we have the following asymptotics:

$$\frac{\int p(x^n|\boldsymbol{\eta})\rho_J(\boldsymbol{\eta})d\boldsymbol{\eta}}{p(x^n|\hat{\boldsymbol{\eta}})} \sim \frac{\sqrt{\det(J(\hat{\boldsymbol{\eta}}))}}{C_J\sqrt{\det(\hat{J}(\hat{\boldsymbol{\eta}}))}}\frac{(2\pi)^{d\ell/2}}{n^{d\ell/2}}. \tag{7}$$

This is obtained by writing $p(x^n|\boldsymbol{\eta})$ as the exponential of $\log p(x^n|\boldsymbol{\eta})$, and taking a second-order Taylor expansion for $\boldsymbol{\eta}$ near $\hat{\boldsymbol{\eta}}$. In this way, one approximates $p(x^n|\boldsymbol{\eta})$ with a Gaussian density function for $\boldsymbol{\eta}$.

When the model is an exponential family, $\hat{J}(\hat{\boldsymbol{\eta}}) = J(\hat{\boldsymbol{\eta}})$ holds. Then, if S were an exponential family, our task would be to control the accuracy of approximation (7) only. Though the stationary Markov model is not exponential type, it converges to an exponential family, when the sample size goes to infinity (see [29] for example). Moreover, for the Markov model, the empirical Fisher information converges to the Fisher information

$$|\hat{J}(x^n, \hat{\boldsymbol{\eta}}) - J(\hat{\boldsymbol{\eta}})| \to 0. \tag{8}$$

This convergence holds uniformly for x^n with $\hat{\boldsymbol{\eta}}$ in a set in the interior of H. As a consequence, it is possible to make the regret of the modified Jeffreys mixture converge to the minimax one.

This task is accomplished by a case argument concerning whether the MLE is near the boundary of the parameter space or not. When we restrict the sequence x^n so that the MLE $\hat{\boldsymbol{\eta}}(x_{-k+1}^n)$ belongs to a compact set K included in the interior of H, then we can prove that the convergence of (7) and (8) is uniform for those sequences, since neighborhoods of $\hat{\boldsymbol{\eta}}$ are guaranteed to be included in H. The Laplace approximation is valid as long as neighborhoods of $\hat{\boldsymbol{\eta}}$ of radius of larger order than $1/\sqrt{n}$ are included in H. Consequently, it is possible to prove the uniform convergence of the regret, even if we moderate the restriction on the sequences. Instead of sequences being restricted to have MLE in a fixed set K, we allow more generally for sequences with the MLE in $H^{(n^{-a})}$ $(0 < a < 1/2)$, where we let

$$H_s^{(\epsilon)} \stackrel{\text{def}}{=} \{\boldsymbol{\eta}_s \in H_s : \forall x \in \mathcal{X},\ \eta_{x|s} \geq \epsilon\}$$

and

$$H^{(\epsilon)} \stackrel{\text{def}}{=} \prod_{s\in L} H_s^{(\epsilon)}.$$

For the sequences with $\hat{\boldsymbol{\eta}}$ within order $1/\sqrt{n}$ of the boundary of H, we cannot use the Laplace approximation. The shape of $p(x^n|\boldsymbol{\eta})$ becomes that of a truncated Gaussian, with reduced value of the integral in (7). A similar reduction to the integral occurs if $\hat{\boldsymbol{\eta}}$ is on the boundary. Hence, the regret would be larger by some amount. Indeed, it has been shown for the memoryless case that the regret of Jeffreys mixture is larger than the asymptotic minimax value by the amount $(d/2)\log 2$, when $\hat{\boldsymbol{\eta}}$ is located at a vertex of H (Lemma 3, [28]). Hence, we need the contribution from the second term of ρ_n, which is $n^{-b}\rho_{(\alpha)}(\boldsymbol{\eta})$. With the help from it, for $\hat{\boldsymbol{\eta}}$ on or near the boundary, we can obtain smaller regret than the minimax value. For the proof, we use Lemma 4 of [28].

The need to consider the difference between $\hat{J}(\hat{\boldsymbol{\eta}})$ and $J(\hat{\boldsymbol{\eta}})$ as in (8) makes the proof about the interior region harder (this problem does not exist for the memoryless case [28] and one-dimensional exponential family [21]).

B. Lower Bound

It is possible to directly obtain a lower bound on the maximin regret which asymptotically matches the upper bound in the previous section. Here, we will give an outline of the proof. Let K be an arbitrary compact subset of H°, and define for each $s_0 = x_{-k+1}^0$

$$\mathcal{K}_{n,s_0} \stackrel{\text{def}}{=} \{x^n : s_0 x^n \in \mathcal{X}^{n+k}, \hat{\boldsymbol{\eta}} \in K\}.$$

In a fashion similar to the upper bound, by Laplace approximation, it is possible to show that

$$\log \frac{p(x^n|s_0, \hat{\boldsymbol{\eta}})}{m_J(x^n|s_0)} = \frac{d\ell}{2}\log\frac{n}{2\pi} + \log\int_H \sqrt{\det J(\boldsymbol{\eta})}d\boldsymbol{\eta} + o(1) \tag{9}$$

uniformly for all $s_0 \in \mathcal{X}^k$ and for all $x^n \in \mathcal{K}_{n,s_0}$. Let m_J is the Jeffreys mixture of $p(x^n|\boldsymbol{\eta})$ for H. Define the restriction of m_J to $\mathcal{K}_{n,s_0}$ as

$$m_J^K(x^n|s_0) \stackrel{\text{def}}{=} \frac{m_J(x^n|s_0)1_{\mathcal{K}_{n,s_0}}(x^n)}{M(\mathcal{K}_{n,s_0}|s_0)}$$

where

$$M(\mathcal{K}_{n,s_0}|s_0) \stackrel{\text{def}}{=} \sum_{x^n\in\mathcal{K}_{n,s_0}} m_J(x^n|s_0).$$

By the definition of $\underline{r}_n = \underline{r}_n(\mathcal{K}_{n,s_0}|s_0)$, it is at least

$$\inf_{r\in\mathcal{P}(\mathcal{X}^n|\mathcal{X}^k)} E_{m_J^K(\cdot|s_0)}\log\frac{p(x^n|s_0,\hat{\boldsymbol{\eta}})}{r(x^n|s_0)} = E_{m_J^K(\cdot|s_0)}\log\frac{p(x^n|s_0,\hat{\boldsymbol{\eta}})}{m_J^K(x^n|s_0)}$$

which by the approximation (9) is of the form

$$\frac{d\ell}{2}\log\frac{n}{2\pi} + \log\int_H\sqrt{\det J(\boldsymbol{\eta})}d\boldsymbol{\eta} + \log M(\mathcal{K}_{n,s_0}|s_0) + o(1),$$

uniformly in $\mathcal{K}_{n,s_0}$. Consequently

$$\underline{r}_n(\mathcal{K}_{n,s_0}|s_0) \geq \frac{d\ell}{2}\log\frac{n}{2\pi} + \log\int_H\sqrt{\det J(\boldsymbol{\eta})}d\boldsymbol{\eta} + \log M(\mathcal{K}_{n,s_0}|s_0) + o(1).$$

Now, let $\{K_i\}$ be a sequence of compact subsets of H° such that $K_i \subset K_{i+1}^\circ$ and $\lim_{i\to\infty}\int_{K_i} d\boldsymbol{\eta} = 1$ holds (K_i converges to H). Let $\mathcal{K}_{n,s_0,i}$ denote the set $\{x^n : s_0x^n \in \mathcal{X}^{n+k}, \hat{\boldsymbol{\eta}} \in K_i\}$. Then, it is possible to prove $\lim_{i\to\infty}\lim_{n\to\infty} M(\mathcal{K}_{n,s_0,i}|s_0) = 1$. This implies

$$\underline{r}_n(\mathcal{X}^n|s_0) \geq \frac{d\ell}{2}\log\frac{n}{2\pi} + \log\int_H\sqrt{\det J(\boldsymbol{\eta})}d\boldsymbol{\eta} + o(1).$$

The right-hand side matches our upper bound on the minimax regret. Another way is to utilize Rissanen's result (Theorem 1, [17]) for a compact K interior to H.

Remark 1: Theorem 1 is a generalization of the result about the first-order Markov chain with alphabet size 2 in [26], but the proof is not its straightforward extension.

Remark 2: A similar bound for Markov chains is obtained in [1], but it is not demonstrated to be uniformly valid over all $\mathcal{X}^n$. In [17], [21] and [31], upper bounds of the same form on the minimax regret are obtained for more general models, but they hold under the restriction on the sequences that MLE is located in a compact set included in the interior of the parameter space (an exception is one-dimensional exponential family in [21]). Under that condition, Jeffreys mixture (we do not need modification) is asymptotically minimax.

Remark 3: It is possible to apply our minimax procedure to the universal prediction problem, using the conditionals $m_n(x_{t+1}|x^t) = m_n(x^{t+1}|s_0)/m_n(x^t|s_0)$, which equals $\int p(x_{t+1}|x^t,\boldsymbol{\eta})\rho_n(\boldsymbol{\eta}|x^t)d\boldsymbol{\eta} = \int \eta_{x_{t+1}|\tau(x^t_{-k+1})}\rho_n(\boldsymbol{\eta}|x^t)d\boldsymbol{\eta}$, where $\rho_n(\boldsymbol{\eta}|x^t)$ denotes the posterior density of $\boldsymbol{\eta}$ given x^t_{-k+1} (recall $\tau(x^t) \stackrel{\text{def}}{=} x^t_{t-k+1}$). The conditionals are essential also for universal coding, since it is needed for arithmetic coding.

Remark 4: The $m_n(x_{t+1}|x^t)$ depends on n because of the modification of Jeffreys prior. Thus, we have to know the length of the sequence in advance, in order to use m_n for the prediction, while the Laplace estimator does not depend on the total length of the sequence. However, it is possible to calculate $m_n(x_{t+1}|x^t)$ even for $t > n$, and use it for prediction, though the minimax property is lost.

C. Computation of Posterior Updates

Although for a product of Dirichlet priors, posterior predictive densities and mixture densities are easy to compute (using the fact that the posterior densities is also Dirichlet), there are additional challenges in computation of the Jeffreys mixture and its modified forms.

The general form of the product of Dirichlet densities is $\bar{D}_{\boldsymbol{\lambda}}(\boldsymbol{\eta}) = \prod_{s\in L}[(\prod_{x=0}^d \eta_{x|s}^{\lambda_{x|s}-1})/C_{\boldsymbol{\lambda}_s}]$, where the normalizing factors are $C_{\boldsymbol{\lambda}_s} = \int \prod_{x=0}^d \eta_{x|s}^{\lambda_{x|s}-1} d\boldsymbol{\eta}_s$ which are called Dirichlet integrals (given as a ratio of products of Gamma functions). For a Dirichlet (α) prior, it is known that the posterior distributions given data x^n are Dirichlet $\bar{D}_{\alpha\mathbf{1}+\boldsymbol{n}}(\boldsymbol{\eta})$, where $\mathbf{1}$ denotes the $(d+1)\ell$-dimensional vector with all entries are 1 and $\boldsymbol{n}$ is the $(d+1)\ell$-dimensional vector with entries $(n_s, s \in L)$. Its predictive distribution follows Laplace update rules for evaluation of

$$\hat{\eta}^{(\alpha)}_{x|s} = \int \boldsymbol{\eta}_{x|s}\bar{D}_{\alpha\mathbf{1}+\boldsymbol{n}}(\boldsymbol{\eta}) = \frac{n_{x|s}+\alpha}{n_s+\alpha(d+1)}.$$

In particular, with $\alpha = 1/2$, this provides what is also called the Krichevsky–Trofimov estimator.

In contrast, the Jeffreys posterior is more involved because of the $\prod_s \mu_s(\boldsymbol{\eta})^{d/2}$ factor in the prior as in (5) where $\mu_s(\boldsymbol{\eta})$ depends on all $\boldsymbol{\eta}_{s'}$ $(s' \in L)$. The posterior density $\rho_J(\boldsymbol{\eta}|x^n)$ is proportional to $\big(\prod_s \mu_s(\boldsymbol{\eta})^{d/2}\big)\bar{D}_{\mathbf{1}/2+\boldsymbol{n}}(\boldsymbol{\eta})$ as described in Appendix A. For the computation of the Jeffreys predictive probabilities and Jeffreys mixture, define the unnormalized estimates

$$\hat{\eta}^J_{x|s} = \int \eta_{x|s}\Big(\prod_{s\in L}\mu_s(\boldsymbol{\eta})^{d/2}\Big)\bar{D}_{(\mathbf{1}/2+\boldsymbol{n})}(\boldsymbol{\eta})d\boldsymbol{\eta}. \tag{10}$$

Then, the Jeffreys predictive probabilities for possible next symbols $x \in \mathcal{X}'$ given data x^n with $\tau(x^n) = s$ are proportional to these $\hat{\eta}^J_{x|s}$. That is, $p_J(x_{n+1} = x|x^n) = \int \eta_{x|s}\rho_J(\boldsymbol{\eta}|x^n)d\boldsymbol{\eta}$ is equal to

$$\frac{\hat{\eta}^J_{x|s}}{\sum_{x\in\mathcal{X}}\hat{\eta}^J_{x|s}}.$$

The successive predictive probabilities $p_J(x_{t+1}|x^t)$ are computed in the same way, where in place of $\boldsymbol{n}$ we use the vector $\boldsymbol{t}$ of counts $t_{x|s}$ for $x \in \{0,\dots,d\}$ and $s \in L$, based on the data segment x^t for each $t \geq 0$.

The Jeffreys mixture $m_J(x^n)$ is computed from successive products of such predictive probabilities.

Also, without the $\prod_s \mu_s(\boldsymbol{\eta})^{d/2}$ factor, one has the $\mathrm{Dirichlet}(\alpha)$ mixture $m_{(\alpha)}(x^n)$. Our modified Jeffreys mixture is thus

$$m_n(x^n) = (1-\kappa_n)m_J(x^n) + \kappa_n m_{(\alpha)}(x^n).$$

The associated marginals are $m_n(x^t) = (1-\kappa_n)m_J(x^t) + \kappa_n m_{(\alpha)}(x^t)$ for $t \leq n$. The posterior weight it gives to the Jeffreys mixture is

$$\pi(J|x^t) = (1-\kappa_n)m_J(x^t)/m_n(x^t).$$

The associated predictive probabilities $p_n(x_{t+1}|x^t) = m_n(x^{t+1})/m_n(x^t)$ are

$$\pi(J|x^t)\,p_J(x_{t+1}|x^t) + (1-\pi(J|x^t))p_{(\alpha)}(x_{t+1}|x^t).$$

This method of computing the mixture needs the computation of

$$\hat{\eta}^J_{x|s} = \int \eta_{x|s}\big(\prod_{s\in L}\mu_s(\boldsymbol{\eta})^{d/2}\big)\bar{D}_{(\mathbf{1}/2+\boldsymbol{t})}(\boldsymbol{\eta})d\boldsymbol{\eta}$$

where $\boldsymbol{t} = (t_s, s \in L)$ is the vector of context counts for each initial segment of length $t \leq n$.

It lacks the explicit form of $\hat{\eta}^{(1/2)}_{x|s} = \int \eta_{x|s}\bar{D}_{(\mathbf{1}/2+\boldsymbol{t})}(\boldsymbol{\eta})d\boldsymbol{\eta}$. Nevertheless, comparison of these integrals leads to advocacy of a Monte Carlo evaluation. To compute $\hat{\eta}^J_{x|s}$ from data x^t, one way is to average $\eta_{x|s}\prod_{s\in L}(\mu_s(\boldsymbol{\eta}))^{d/2}$ with a large number (a million) of independent $\boldsymbol{\eta}$ each drawn according to the Dirichlet $\bar{D}_{(\mathbf{1}/2+\boldsymbol{t})}$ distribution. A refinement to this Monte Carlo evaluation is given in Section VI for the two-state first-order Markov case.

An alternative to Monte Carlo evaluation is an approximation formula appropriate for long strings with $\hat{\boldsymbol{\eta}} \in K$, as developed next.

D. Approximation Formula

As stated in the preceding section, the Jeffreys mixture for the Markov model is nearly a best strategy, but it is hard to calculate it in general, because it is a multi-integral with respect to the parameter $\boldsymbol{\eta}$. The following theorem provides its approximation formula, which is easier to calculate than the original form. Here, note that $\int p(x|x^n)\rho_J(\boldsymbol{\eta}|x^n)d\boldsymbol{\eta} = \int \eta_{x|s}\rho_J(\boldsymbol{\eta}|x^n)d\boldsymbol{\eta}$ with $s = \tau(x_{n-k+1}^n)$.

Theorem 2: Let $S = \{p(\cdot|s,\boldsymbol{\eta})|\eta \in H, s \in L\}$ be a class of kth-order Markov chains with the alphabet $\{0,1,\ldots,d\}$. Let K be a compact set included in the interior of H and n_0 be an arbitrary natural number. Then, for all $x \in \mathcal{X}$ and for all $s \in L$

$$\begin{aligned}\int \eta_{x|s}\rho_J(\boldsymbol{\eta}|x^n)d\boldsymbol{\eta} &= \frac{n_{x|s}+0.5}{n_s+(d+1)/2} \\ &+ \sum_{y\in\mathcal{X}',\, t\in L} \frac{d\hat{\eta}_{y|s}(\delta_{xy}-\hat{\eta}_{x|s})}{2n_s+d+1}\left.\frac{\partial \log \mu_t}{\partial \eta_{y|s}}\right|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}} \\ &+ O\Big(\frac{\sqrt{\log n}}{n\sqrt{n}}\Big)\end{aligned} \tag{11}$$

holds, uniformly for all infinite sequences $x_{-k+1}\ldots x_1 x_2\ldots$ such that $\hat{\boldsymbol{\eta}} \in K$ holds for all $n \geq n_0$ for some $n_0 \geq 1$.

This represents the approximation formula as an additive modification to the estimator $(n_{x|s}+0.5)/(n_s+(d+1)/2)$. Note that the following multiplicative form is equally valid:

$$\begin{aligned}\int \eta_{x|s}\rho_J(\boldsymbol{\eta}|x^n)d\boldsymbol{\eta} &= \frac{n_{x|s}+0.5}{n_s+(d+1)/2} \\ &\cdot \exp\Bigg(\frac{d}{2}\sum_{y\in\mathcal{X}',\, t\in L}\frac{\hat{\eta}_{y|s}(\delta_{xy}-\hat{\eta}_{x|s})}{n_{x|s}+0.5}\left.\frac{\partial \log \mu_t}{\partial \eta_{y|s}}\right|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}} \\ &+ O\Big(\frac{\sqrt{\log n}}{n\sqrt{n}}\Big)\Bigg).\end{aligned}$$

This theorem is proved from Theorem 3 given later.

Theorem 2 shows how we should correct the Laplace-like estimator (the first factor of the right-hand side) in order to decrease the worst case logarithmic regret. The correction (second term) contains the derivative of the stationary probabilities, which are rational functions of the parameter $\boldsymbol{\eta}$ as shown in Appendix B.

The following example is the simplest case.

Example 1: Let $\mathcal{X} = \{0,1\}$ and $d = 1$ ($L = \{0,1\}$). We have $\mu_1 = \eta_{1|0}/(\eta_{1|0}+\eta_{0|1})$ and $\mu_0 = \eta_{0|1}/(\eta_{1|0}+\eta_{0|1})$. Let $s = 0$ and $x = 1$; then, the sum in exponent in the third line of (11) equals

$$\begin{aligned}&\sum_{y\in\mathcal{X}',\, t\in L}\frac{\eta_{y|s}(\delta_{xy}-\eta_{x|s})}{n_{x|s}+1/2}\frac{\partial \log \mu_t}{\partial \eta_{y|s}} \\ &= \sum_{t\in L}\frac{\eta_{1|0}(1-\eta_{1|0})}{n_{1|0}+1/2}\frac{\partial \log \mu_t}{\partial \eta_{1|0}} \\ &= \frac{\eta_{1|0}(1-\eta_{1|0})}{n_{1|0}+1/2}\Big(\frac{1}{\eta_{1|0}}-\frac{2}{\eta_{1|0}+\eta_{0|1}}\Big).\end{aligned}$$

Let $x_n = 0$ ($\tau(x^n) = 0$); then, by Theorem 2, the approximation of the Jeffreys mixture for this case is given by

$$\begin{aligned}&\int \eta_{1|0}\rho_J(\boldsymbol{\eta}|x^n)d\boldsymbol{\eta} \\ &\quad\approx \frac{n_{1|0}+0.5}{n_0+1} \\ &\quad\quad+\frac{1}{n_0+1}\Big(\frac{1-\hat{\eta}_{1|0}}{2}-\frac{\hat{\eta}_{1|0}(1-\hat{\eta}_{1|0})}{\hat{\eta}_{1|0}+\hat{\eta}_{0|1}}\Big) \\ &\quad= \frac{n_{1|0}+0.5+(1-\hat{\eta}_{1|0})(0.5-\hat{\mu}_1)}{n_0+1}\end{aligned} \tag{12}$$

where the error term is $O(\sqrt{\log n}/n\sqrt{n})$ and $\hat{\mu}_1$ denotes $\hat{\eta}_{1|0}/(\hat{\eta}_{0|1}+\hat{\eta}_{1|0})$.

Note that this depends not only on $\hat{\eta}_{1|0}$ but on $\hat{\eta}_{0|1}$ and that the difference between this and the Laplace-like estimator is of order $\Omega(1/n_0)$ (where "$\Omega(x)$" denotes negation of "$o(x)$"). It is known that the Jeffreys mixture for the i.i.d. case induces the Laplace-like estimator (Krichevsky–Trofimov estimator), which is widely used in many data compression or prediction methods, e.g., in the CONTEXT [16] and the CTW [25] method, even though these methods are for non-i.i.d. sources. The reason is that it is in a very simple form and is believed to have good coding performance. Theorem 2 shows that for Markov sources, the Laplace estimator is different from the minimax strategy in terms of second-order efficiency. Moreover, the theorem suggests that we may have to calculate the derivative of stationary probabilities every time a datum is input to achieve the minimax regret in sequential prediction or data compression with Markov models. If we employ a naive algorithm to calculate them, the computational cost is of order $O(\ell^3)$, since it includes the eigenvalue problem. Note that it can be reduced to $O(\ell^2)$ by making use of the Sherman–Morrison formula (see [15] for example).

We can show a more general approximation formula (Theorem 3), from which Theorem 2 is obtained as a corollary. To state it, we need some preliminaries. First, we introduce another parameter $\boldsymbol{\theta}$ than $\boldsymbol{\eta}$. Note that $p(x^n|\boldsymbol{\eta})$ is rewritten as follows:

$$\begin{aligned}p(x^n|\boldsymbol{\eta}) &= \prod_{s\in L, x\in\mathcal{X}}(\eta_{x|s})^{n_{x|s}} \\ &= \prod_{s\in L, x\in\mathcal{X}}((\eta_{x|s})^{n_{x|s}/n_s})^{n_s} \\ &= \prod_{s\in L}\exp(n_s(\sum_{x\in\mathcal{X}'}\theta_{x|s}\hat{\eta}_{x|s}-\psi(\boldsymbol{\theta}_s)))\end{aligned} \tag{13}$$

where we let $\theta_{x|s} = \theta_{x|s}(\boldsymbol{\eta}_s) = \log(\eta_{x|s}/\eta_{0|s})$, $\boldsymbol{\theta}_s = \boldsymbol{\theta}_s(\boldsymbol{\eta}_s) = (\theta_{1|s},\ldots,\theta_{d|s})^t$, and $\psi(\boldsymbol{\theta}_s) = -\log\eta_{0|s} = \log(1+\sum_{x\in\mathcal{X}'}\exp\theta_{x|s})$. Recall that $\hat{\eta}_{x|s} = n_{x|s}/n_s$, where $n_{x|s}$ denotes the number of occurrences of x at the state s in the sequence x^n, and $n_s \stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}} n_{x|s}$. Define $\boldsymbol{\theta} = (\boldsymbol{\theta}_{s_1}^t,\boldsymbol{\theta}_{s_2}^t,\ldots,\boldsymbol{\theta}_{s_\ell}^t)^t$ similarly as $\boldsymbol{\eta}$. Let $\Theta_s \stackrel{\text{def}}{=} \{\boldsymbol{\theta}_s(\boldsymbol{\eta}_s) : \boldsymbol{\eta}_s \in H_s^\circ\}$; then, $\Theta_s = \Re^{|\mathcal{X}'|}$ holds. Let $\Theta \stackrel{\text{def}}{=} \prod_{s\in L}\Theta_s = \Re^{\ell\cdot|\mathcal{X}'|}$. It is known that the map $\boldsymbol{\eta}_s \mapsto \boldsymbol{\theta}_s(\boldsymbol{\eta}_s)$ on H_s° is one to one and analytic (see [3]). Note that $(\partial/\partial\theta_{x|s})\psi(\boldsymbol{\theta}_s) = \eta_{x|s}$ holds. Define functions g_{xy} as

$$g_{xy}(\boldsymbol{\theta}_s) \stackrel{\text{def}}{=} \frac{\partial^2\psi(\boldsymbol{\theta}_s)}{\partial\theta_{y|s}\partial\theta_{x|s}}. \tag{14}$$

Then, we have $\partial\eta_{x|s}/\partial\theta_{y|s} = g_{xy}(\boldsymbol{\theta}_s)$. Let $g(\boldsymbol{\theta}_s)$ denote the matrix whose (x,y) component is $g_{xy}(\boldsymbol{\theta}_s)$. Note that $g(\boldsymbol{\theta}_s)$ is positive definite for any $\boldsymbol{\theta}_s \in \Theta_s$. Let $I(\boldsymbol{\eta}_s)$ denote the inverse matrix of $g(\boldsymbol{\theta}_s)$. Then, the following holds:

$$I_{xy}(\boldsymbol{\eta}_s) = \frac{\delta_{xy}}{\eta_{x|s}} + \frac{1}{\eta_{0|s}}. \tag{15}$$

That is, $I_{xy}(\boldsymbol{\eta}_s)$ equals the Fisher information matrix for the multinomial Bernoulli model. Note that $\mu_s I_{x,y}(\boldsymbol{\eta}_s) = J_{sx,sy}(\boldsymbol{\eta})$ holds. Since $\theta_{x|s} = \log(\eta_{x|s}/\eta_{0|s})$, $\eta_{x|s} = e^{\theta_{x|s}}/(1+\sum_{z\in\mathcal{X}'} e^{\theta_{x|s}})$ holds. Hence, we have

$$\begin{aligned}\frac{\partial\eta_{y|s}}{\partial\theta_{x|s}} &= -\frac{\eta_{y|s}\eta_{x|s}}{(1+\sum_{z\in\mathcal{X}'} e^{\theta_{z|s}})^2} + \frac{\delta_{xy}\eta_{y|s}}{1+\sum_{z\in\mathcal{X}'} e^{\theta_{z|s}}} \\ &= -\eta_{x|s}\eta_{y|s} + \delta_{xy}\eta_{y|s} = \eta_{x|s}(\delta_{xy} - \eta_{y|s}).\end{aligned}$$

Therefore, we have

$$\frac{\partial}{\partial\theta_{x|s}} = \sum_{y\in\mathcal{X}'} \eta_{x|s}(\delta_{xy} - \eta_{y|s})\frac{\partial}{\partial\eta_{y|s}}. \tag{16}$$

Given the prior measure $\rho(\boldsymbol{\eta})d\boldsymbol{\eta}$, denote the prior density function with respect to $d\boldsymbol{\theta}$ as

$$\begin{aligned}w(\boldsymbol{\theta}) &= \rho(\boldsymbol{\eta})\left|\det\left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\theta}}\right)\right| \\ &= \rho(\boldsymbol{\eta})\prod_{s\in L}\det(g(\boldsymbol{\theta}_s)) = \rho(\boldsymbol{\eta})\prod_{s\in L, y\in\mathcal{X}}\eta_{y|s}.\end{aligned}$$

For the Jeffreys prior and the Dirichlet prior, let

$$\begin{aligned}w_J(\boldsymbol{\theta}) &= \frac{\prod_{s\in L}\mu_s^{d/2}D_{(1/2)}(\boldsymbol{\eta}_s)\prod_{y\in\mathcal{X}}\eta_{y|s}}{C_J} \\ &= \frac{\prod_{s\in L}\mu_s^{d/2}D_{(3/2)}(\boldsymbol{\eta}_s)}{C_J} \\ w_{(\alpha)}(\boldsymbol{\theta}) &= \frac{\prod_{s\in L}D_{(\alpha)}(\boldsymbol{\eta}_s)\prod_{y\in\mathcal{X}}\eta_{y|s}}{(C_{(\alpha)})^\ell} \\ &= \frac{\prod_{s\in L}D_{(\alpha+1)}(\boldsymbol{\eta}_s)}{(C_{(\alpha)})^\ell}.\end{aligned}$$

The following is our assumption for a prior density w.

Assumption 1: For a compact set K included in H°, there exists a certain integer $m_{\min}$, such that for all $\boldsymbol{\eta}' \in K$, for all $x \in \mathcal{X}'$, and for all $s \in L$

$$\frac{\partial\log w(\boldsymbol{\theta})}{\partial\theta_{x|s}}\prod_{t\in L, y\in\mathcal{X}}(\eta_{y|t})^{m_{min}\eta'_{y|t}}\cdot w(\boldsymbol{\theta})$$

is integrable over Θ.

Suppose that Assumption 1 holds for a prior w. Then, $\partial\log w(\boldsymbol{\theta})/\partial\theta_{x|s}\cdot w(\boldsymbol{\theta})\prod_{t\in L,y\in\mathcal{X}}(\eta_{y|t})^{m_{\min}\eta'_{y|t}}$ is integrable for any $\boldsymbol{n}$ such that for all $s \in L$, $n_s \geq m_{\min}$. This assumption holds for Jeffreys prior (see Lemma 5 in Appendix C).

The following theorem provides an approximation formula for a general prior density.

Theorem 3: Let K be a compact set included in the interior of H. Suppose that the prior density w satisfies Assumption 1 for a certain $m_{\min}$. Then, the following holds, uniformly for all infinite sequences $x_{-k+1}\ldots x_1x_2\ldots$ such that $\hat{\boldsymbol{\eta}} \in K$ holds for all $n \geq m_{\min}$:

$$\begin{aligned}&\int \eta_{x|s}w(\boldsymbol{\eta}|x^n)d\boldsymbol{\eta} \\ &\quad= \hat{\eta}_{x|s} + \frac{1}{n_s}\left.\frac{\partial\log w(\boldsymbol{\theta})}{\partial\theta_{x|s}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + O(\frac{\sqrt{\log n}}{n\sqrt{n}}).\end{aligned} \tag{17}$$

Alternatively, the following holds uniformly for all sequences $x_{-k+1}\ldots x_1x_2\ldots$ such that $\hat{\boldsymbol{\eta}}^{(1/2)} \in K$ holds for all $n \geq m_{\min}$, where we recall $\hat{\boldsymbol{\eta}}^{(1/2)}_{x|s} \stackrel{\text{def}}{=} (n_{x|s}+0.5)/(n_s+(d+1)/2)$

$$\begin{aligned}&\int \eta_{x|s}w(\boldsymbol{\eta}|x^n)d\boldsymbol{\eta} \\ &\quad= \hat{\eta}^{(\frac{1}{2})}_{x|s} + \frac{1}{n_s+(d+1)/2}\left.\frac{\partial\log(w(\boldsymbol{\theta})/w_{(\frac{1}{2})}(\boldsymbol{\theta}))}{\partial\theta_{x|s}}\right|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^{(\frac{1}{2})}} \\ &\quad+ O(\frac{\sqrt{\log n}}{n\sqrt{n}}).\end{aligned} \tag{18}$$

The proof is given in Section V.

Noting $|\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}^{(1/2)}| = O(1/n)$, Theorem 2 is easily derived from (18) and (16).

E. Simulation

Concerning the simplest case (Example 1), where the target model is the first-order Markov chain with binary alphabet, we evaluate the coding regret of the strategy using the algorithm described in Sections III-C and III-D.

In the following experiments, we used the Monte Carlo method when $n_s \leq 20$ holds for the current state s, and otherwise we used the approximation formula (12), when computing the successive factors of $m_J(x^n)$ and $q(x^n)$.

We generated data sequences of length $n = 10^7$, which was according to Markov sources with various parameter settings. The parameter settings are $\alpha = 0.019$ and $b = 0.16$ ($\kappa_n = n^{-0.16}$), which satisfies the assumption of Theorem 1. Note that b must be smaller than $(0.5-\alpha)/3 \leq 0.5/3 = 0.16\cdots$; hence, the setting $b = 0.16$ is nearly optimal for rapid convergence of κ_n. In fact, we have $n^{-0.16} = 0.0759\cdots$ when $n = 10^7$. We set Monte Carlo sample size to 1 000 000. Table I here shows the results of our experiment. In each line, we list the MLE for $(\eta_{0|1}, \eta_{1|0})$ and the computed values of the regret of the procedures based on m_J and q. The regret of q is $\tilde{r}(q) = \tilde{r}(q, x^n_{-k+1})$ defined by

$$\tilde{r}(q, x^n_{-k+1}) \stackrel{\text{def}}{=} \log\frac{1}{q(x^n|x^0_{-k+1})} - \log\frac{1}{p(x^n|\hat{\boldsymbol{\eta}})} - \log\frac{n}{2\pi}.$$

The column heading $\tilde{r}(m_J)$ lists the regrets by the genuine Jeffreys mixture and the column heading $\tilde{r}(q)$ lists the regrets by the modified Jeffreys mixture computed by the proposed method. Here, the $m_J(x^n)$ and $q(x^n)$ are computed by as a product of successive factors required for prediction and for arithmetic coding. If q is the minimax strategy, $\tilde{r}(q, x^n_{-k+1})$ converges to $\log C_J$, which approximately equals 1.2985 (see Appendix D). The regrets by q in Table I are approximately 0.08 nat larger than this value. It coincides with the fact that $-\log(1-\kappa_n) \approx -\log(1-0.0759) = 0.078\cdots$. For the ordinary (nonextremal) cases, we see that the regrets of the genuine

TABLE I
REGRET (ORDINARY CASES)

$\hat{\eta}_{0\|1}$	$\hat{\eta}_{1\|0}$	$\tilde{r}(m_J)$	$\tilde{r}(q)$
0.0999	0.0998	1.297	1.376
0.0999	0.3002	1.300	1.378
0.0999	0.5004	1.300	1.378
0.1000	0.6997	1.298	1.377
0.0999	0.8998	1.300	1.376
0.3000	0.3000	1.303	1.382
0.3002	0.4999	1.301	1.379
0.3002	0.7001	1.299	1.377
0.3000	0.9000	1.298	1.376
0.4998	0.5004	1.299	1.378
0.5000	0.7001	1.297	1.376
0.5000	0.8998	1.299	1.377
0.7001	0.7003	1.302	1.380
0.7001	0.9001	1.303	1.382
0.9000	0.9000	1.300	1.378

Jeffreys mixture (the case of $\kappa_n = 0$) are between 1.297 and 1.303. For each line, the data x_0^n are generated according to a Markov source with $\eta_{0|1}$ and $\eta_{1|0}$ equal to the two digit values which the reported $\hat{\eta}_{0|1}$ and $\hat{\eta}_{1|0}$ clearly estimate. While the expected regret depends only on the $\hat{\eta}_{0|1}$ and $\hat{\eta}_{1|0}$ based on the whole sample, our approximation uses the $q(x_{t+1}|x^t)$ based on partial samples of sizes $t \leq n$. Consequently, different realizations of x_0^n of the same Markov types $\hat{\eta}_{0|1}$ and $\hat{\eta}_{1|0}$ will have slightly different computed regret.

Here, we used a Monte Carlo sample size of 1 000 000 for near three digit accuracy. A Monte Carlo size of 10 000 would be sufficient for two digit accuracy.

IV. PROOF OF THEOREM 1

In this section, we give the proof of Theorem 1. As described in Section III-A, a key of the proof is the convergence rate of the determinant of empirical Fisher information to that of Fisher information. Comparing (3) with (4), we realize that our main task is to evaluate the ratio $\hat{p}_s/\mu_s(\hat{\boldsymbol{\eta}})$ for $s \in L$ where $\hat{p}_s \stackrel{\text{def}}{=} n_s/n$. Hence, we first give a lemma about this item in the next section. After that we will prove Theorem 1.

A. Convergence of State Frequency to Stationary Probability

We can show the following Lemma.

Lemma 1: Let $r = k(\ell - 1)$. For all $\epsilon \in (0, 1]$, if $\hat{\boldsymbol{\eta}}(x^n) \in H^{(\epsilon)}$ is satisfied, then the following two inequalities hold:

$$n_s \geq n(\epsilon/2)^k \tag{19}$$

$$|\log \frac{\hat{p}_s}{\mu_s(\hat{\boldsymbol{\eta}})}| < \frac{C_1}{n\epsilon^{r+k}} \tag{20}$$

where C_1 is a certain positive constant independent of ϵ and n.

Remark: When the model is the first-order Markov chain with alphabet $\{0, 1\}$, the proposition which corresponds to Lemma 1 is easy to show, since the explicit forms of μ_s are very simple.

Let $\epsilon_n = n^{-a}$, where we assume

$$0 < a < \frac{1}{2r + k}.$$

Then, we have $n\epsilon_n^{r+k} > n^{1-(r+k)/(2r+k)} \to \infty$ as n goes to infinity. Hence, by Lemma 1, we have $\hat{p}_s/\mu_s(\hat{\boldsymbol{\eta}}) \to 1$ uniformly for all x_{-k+1}^n such that $\hat{\boldsymbol{\eta}} \in H^{(\epsilon_n)}$. Hence, Lemma 1 implies that empirical Fisher information converges to Fisher information, uniformly for all x_{-k+1}^n such that $\hat{\boldsymbol{\eta}} \in H^{(\epsilon_n)}$.

In the remainder of this section, we describe the proof of Lemma 1.

The sequence of states of x_{-k+1}^n are the successive overlapping segments of length k shifting by just 1. Thus, there is a length n sequence of states arising from x^n after the initial state.

Define $n_{u|t}$ for every pair of strings $t, u \in L$ as the number of transitions from the state $t \in L$ to the state $u \in L$ in the sequence x_{-k+1}^n. Likewise, for $s \in L$, we let $n_{x|s}$ denote the number of occurrences of an individual symbol $x \in \mathcal{X}$ after the state s in the sequence x_{-k+1}^n. Then, $n_{x|s}$ equals $n_{\tau(sx)|s}$.

Similarly, we are to define the parameter $\eta_{u|t}$ for every $u, t \in L$. First, define

$$D_s \stackrel{\text{def}}{=} \{\tau(sx) : x \in \mathcal{X}\}.$$

The set D_s consists of the states which are reached by one transition from the state s. Note that for $u \in D_s$, there exists a unique $x \in \mathcal{X}$ such that $u = \tau(sx)$. Let $\xi(s, u)$ denote such x for every $s \in L$ and $u \in D_s$. Then, for every $u, t \in L$, define

$$\eta_{u|s} = \begin{cases} \eta_{\xi(s,u)|s}, & \text{when } u \in D_s \\ 0, & \text{otherwise.} \end{cases}$$

Then, let Π be a matrix whose (t, u) component is $\Pi_{tu} = \eta_{t|u}$; then, it is the state transition probability matrix, and let Π^k be its kth power.

First, we will show the following.

Proposition 1: Let ϵ be a nonnegative real number. If $\eta_{x|t} > \epsilon$ holds for each $t \in L = \mathcal{X}^k$ and each $x \in \mathcal{X}$, then $\mu_t > \epsilon^k$ holds.

Proof: Note that the stationary probabilities μ_t $(t \in L)$ satisfy the following linear equations:

$$\mu_t = \sum_{t' \in L} \Pi_{tt'} \mu_{t'}. \tag{21}$$

For each $t \in L$ and $x \in \mathcal{X}$, we have $\eta_{x|t} > \epsilon$ by the assumption. Now $\tau(t't) = t$ holds for each pair $(t, t') \in L^2$. This implies that it is possible to get to any state from any state by k transitions. Further, $\eta_{x|t} > \epsilon$ holds for all $(t, x) \in L \times \mathcal{X}$,

This implies that each element of Π^k is larger than ϵ^k, i.e., each μ_t is larger than ϵ^k. This completes the *proof of Proposition* 1.

Proposition 2: There exists a certain positive number C_1, such that

$$\left|\frac{\partial \log \mu_s}{\partial \eta_{x|t}}\right| \leq \frac{C_1}{\epsilon^r}$$

holds for all $s \in L$, for all $t \in L$, for all $x \in \mathcal{X}'$, for all $\epsilon > 0$, and for all $\boldsymbol{\eta} \in H^{(\epsilon)}$, where $r = k(\ell - 1)$.

Proof: Renumber the states as $L = \{s_1, s_2, \ldots s_\ell\}$. Define a matrix A as $A_{ij} \stackrel{\text{def}}{=} (\Pi^k)_{s_i s_j}$ and a vector $\boldsymbol{\mu}$ as $\boldsymbol{\mu} \stackrel{\text{def}}{=} (\mu_{s_1}, \ldots, \mu_{s_\ell})^t$. By Lemma 4 in Appendix B, we have

$$\mu_i = \frac{\Delta_{ii}}{\sum_j \Delta_{jj}}. \tag{22}$$

and $\Delta_{ii} \geq \epsilon^{k(\ell-1)}$ for all $\boldsymbol{\eta} \in H^{(\epsilon)}$ and for all $s \in L$, where Δ_{ij} denotes the (i,j)th cofactor of $I - A$, where I is the identity matrix. Hence, we have

$$\frac{\partial \log \mu_s}{\partial \eta_{x|t}} = \frac{1}{\Delta_{ii}} \frac{\partial \Delta_{ii}}{\partial \eta_{x|t}} - \frac{1}{\sum_j \Delta_{jj}} \frac{\partial \sum_j \Delta_{jj}}{\partial \eta_{x|t}}.$$

Note that the derivative of Δ_{jj} is bounded from above when $\boldsymbol{\eta} \in H^{(\epsilon)}$. Therefore, we have for all $t \in L$, for all $x \in \mathcal{X}'$, and for all $\boldsymbol{\eta} \in H^{(\epsilon)}$

$$\left| \frac{\partial \log \mu_s}{\partial \eta_{x|t}} \right| \leq \frac{C}{\epsilon^{k(\ell-1)}}.$$

This completes the proof of Proposition 2.

Remark: By using Lemma 4 in Appendix B, which gives an explicit form of the stationary probabilities $\mu_s(\boldsymbol{\eta})$, we can write down the Jeffreys prior as

$$\rho_J(\boldsymbol{\eta}) = \frac{1}{C_J} \prod_j \left(\frac{\bar{\Delta}_{jj}}{\sum_{l=1}^{\ell} \bar{\Delta}_{ll}} \right)^{d/2} D_{(1/2)}(\boldsymbol{\eta}_{s_j})$$

where $\bar{\Delta}_{ij}$ denotes the (i,j)th cofactor of the matrix whose entries are $\delta_{ij} - \Pi_{s_i s_j}$, and C_J is the normalization constant.

Now, we can prove Lemma 1.

Proof of Lemma 1: Let $s_0 \stackrel{\text{def}}{=} \tau(x_{-k+1}^0)$ be the initial state, and $s_e \stackrel{\text{def}}{=} \tau(x^n)$ be the final state. First, we treat a special case in which $s_0 = s_e$ holds. In this case, we have

$$\forall s \in L, \quad \sum_{t \in L} n_s^t = \sum_{t \in L} n_t^s \tag{23}$$

since the number of all transition from the state s equals the number of all transition to the state s. Hence, we have

$$\sum_{t \in L} \hat{\eta}_{s|t}\, \hat{p}_t = \sum_{t \in L} \frac{n_{s|t}}{n_t} \frac{n_t}{n} = \sum_{t \in L} \frac{n_{s|t}}{n} = \frac{n_s}{n} = \hat{p}_s. \tag{24}$$

This implies $\hat{p}_s = \mu_s(\hat{\boldsymbol{\eta}})$.

When $s_0 \neq s_e$, let $x_{n+1}^{n+\alpha}$ be a minimum path from state s_e to s_0 (α does not exceed k). By adding a sequence $x_{n+1}^{n+\alpha}$ to the sequence x^n, we have $\tau(x^{n+\alpha}) = s_0$. Then, we have $\hat{p}_s(x^{n+\alpha}) = \mu_s(\boldsymbol{\eta}(x^{n+\alpha}))$. Let $\phi_{t|s}$ denote the number of transition from state s to state t in the sequence $x_n \ldots x_{n+\alpha}$, and let $\phi_s = \sum_t \phi_{t|s}$. Here, $\phi_s = 0$ or 1, since x_{n+1}^{α} is the minimum path from s_e to s_0. We have $\hat{\eta}_{t|s}(x^{n+\alpha}) = (n_{t|s} + \phi_{t|s})/(n_s + \phi_s)$. Hence

$$\hat{\eta}_{t|s}(x^{n+\alpha}) \geq \frac{n_{t|s}}{n_s + 1} = \frac{\hat{\eta}_{t|s}(x^n)}{1 + 1/n_s} \geq \frac{\hat{\eta}_{t|s}(x^n)}{2} > \frac{\epsilon}{2}$$

where we use the fact that $n_s = \sum_t n_{t|s} \geq 1$ for sufficiently large n. This can be shown as follows. If $n_{t|s} = 0$ holds for all $t \in L$, then $n_{s|t} \leq 1$ for all $t \in L$. Since there exists one $u \in L$ at least such that $n_u \geq n/\ell$, we have $\hat{\eta}_{s|u}(x^n) = n_{s|u}/n_u \leq \ell/n$, which is smaller than ϵ for sufficiently large n. This contradicts the assumption $\hat{\boldsymbol{\eta}}(x^n) \in H^{(\epsilon)}$.

By $\hat{\eta}_{t|s}(x^{n+\alpha}) > \epsilon/2$ and Proposition 1, we have $\hat{p}_s(x^{n+\alpha}) = \mu_s(\hat{\boldsymbol{\eta}}(x^{n+\alpha})) > (\epsilon/2)^k$, i.e., $(n_s + \phi_s)/(n + \alpha) > \epsilon^k$. Therefore, $n_s > n(\epsilon/2)^k - 1$ holds, which means $n_s \geq n(\epsilon/2)^k$. This is (19).

Hence, we have

$$\hat{\eta}_{t|s}(x^{n+\alpha}) \geq \frac{\hat{\eta}_{t|s}(x^n)}{1 + 1/n_s} > \frac{\hat{\eta}_{t|s}(x^n)}{1 + 2^k/(n\epsilon^k)}.$$

Hence

$$\begin{aligned} \hat{\eta}_{t|s}(x^n) &< \hat{\eta}_{t|s}(x^{n+\alpha})(1 + \frac{2^k}{n\epsilon^k}) \\ &\leq \hat{\eta}_{t|s}(x^{n+\alpha}) + \frac{2^k}{n\epsilon^k}. \end{aligned}$$

Also, we have $\hat{\eta}_{t|s}(x^{n+\alpha}) < (n_{t|s} + 1)/n_s = \hat{\eta}_{t|s}(x^n) + 1/n$. Therefore, we have

$$|\hat{\eta}_{t|s}(x^{n+\alpha}) - \hat{\eta}_{t|s}(x^n)| < \frac{2}{n\epsilon^k}.$$

By Taylor's theorem, we have

$$\begin{aligned} &\log \mu_s(\hat{\boldsymbol{\eta}}(x^{n+\alpha})) - \log \mu_s(\hat{\boldsymbol{\eta}}(x^n)) \\ &\quad = \sum_{t \in L,\ x \in \mathcal{X}'} \left. \frac{\partial \log \mu_s}{\partial \eta_{x|t}} \right|_{\boldsymbol{\eta} = \boldsymbol{h}} (\hat{\eta}_{x|t}(x^{n+\alpha}) - \hat{\eta}_{x|t}(x^n)) \end{aligned}$$

where $\boldsymbol{h}$ is a point between $\hat{\boldsymbol{\eta}}(x^{n+\alpha})$ and $\hat{\boldsymbol{\eta}}(x^n)$. Since $\hat{\boldsymbol{\eta}}(x^{n+\alpha}), \hat{\boldsymbol{\eta}}(x^n) \in H^{(\epsilon)}$, $\boldsymbol{h} \in H^{(\epsilon)}$ holds. Hence, by Proposition 2, we have

$$\left| \left. \frac{\partial \log \mu_s}{\partial \eta_{x|t}} \right|_{\boldsymbol{\eta} = \boldsymbol{h}} \right| \leq \frac{C 2^r}{\epsilon^r}.$$

Hence, we have

$$-\frac{2^k C k \ell}{n\epsilon^{2r+k}} \leq \log \frac{\mu_s(\hat{\boldsymbol{\eta}}(x^{n+\alpha}))}{\mu_s(\hat{\boldsymbol{\eta}}(x^n))} \leq \frac{2^k C k \ell}{n\epsilon^{2r+k}}. \tag{25}$$

Since $\hat{p}_s(x^{n+\alpha}) = (n_s + \phi_s)/(n + \alpha)$, we have

$$\hat{p}_s(x^{n+\alpha}) \geq \frac{n_s}{n + \alpha} = \frac{\hat{p}_s(x^n)}{1 + \alpha/n} \geq \frac{\hat{p}_s(x^n)}{1 + k/n}$$

and

$$\begin{aligned} \hat{p}_s(x^{n+\alpha}) &\leq \frac{n_s + 1}{n} = \hat{p}_s(x^n) + \frac{1}{n} \\ &= \hat{p}_s(x^n)(1 + \frac{1}{n\hat{p}_s}) = \hat{p}_s(x^n)(1 + \frac{1}{n_s}). \end{aligned}$$

Hence

$$\frac{1}{1 + 1/n_s} \leq \frac{\hat{p}_s(x^n)}{\hat{p}_s(x^{n+\alpha})} \leq 1 + \frac{k}{n}$$

that is

$$-\frac{1}{n_s} \leq \log \frac{\hat{p}_s(x^n)}{\hat{p}_s(x^{n+\alpha})} \leq \frac{k}{n}$$

holds. Together with (25) and $\hat{p}_s(x^{n+\alpha}) = \mu_s(\hat{\boldsymbol{\eta}}(x^{n+\alpha}))$, we have

$$\begin{aligned} -\frac{C_2}{n\epsilon^{r+k}} &< -\frac{2^k C_1 k \ell}{n\epsilon^{r+k}} - \frac{1}{n_s} \\ &< \log \frac{\hat{p}_s(x^n)}{\mu_s(\hat{\boldsymbol{\eta}}(x^n))} \leq \frac{2^k C_1 k \ell}{n\epsilon^{r+k}} + \frac{k}{n} < \frac{C_2}{n\epsilon^{r+k}} \end{aligned}$$

where we use (19) and let $C_2 = 2\max\{2^k C_1 k\ell,\ k\}$. This completes the Proof of Lemma 1.

B. Proof of Theorem 1

Now, we can prove Theorem 1.

Proof of Theorem 1: Let a denote a certain constant which satisfies

$$b < a(1/2-\alpha) < \frac{1/2-\alpha}{2r+k}.$$

There exists such a, since b is smaller than

$$\frac{1/2-\alpha}{k(2\ell-1)} = \frac{1/2-\alpha}{2r+k}$$

by the assumption of the theorem (recall $r = k(\ell-1)$). Define $\epsilon_n = n^{-a}$. Suppose that $\hat{\boldsymbol{\eta}} \in H^{(\epsilon_n)}$ holds. Then, by Lemma 1, we have

$$n_s \geq n(\epsilon_n/2)^k \geq \frac{n^{1-k/(r+k)}-1}{2^k} \to \infty$$

when n goes to infinity.

Since $r = k(\ell-1)$, we have $r \geq 1$ and

$$\frac{1}{2r+k} = \frac{1}{k(2\ell-1)}.$$

Since $r \geq 1$, it follows that $a < 1/(2+k)$ holds. This implies $(1-ak)/2 > a$.

In this proof, let κ_n denote $((\ell-1)/n)^b$.

Part I (interior points): First, we treat sequences with $\hat{\boldsymbol{\eta}} \in H^{(\epsilon_n)}$. Note the inequality

$$\frac{m_n(x^n)}{p(x^n|\hat{\boldsymbol{\eta}})} \geq \frac{(1-\kappa_n)\int p(x^n|\boldsymbol{\eta})\rho_J(\boldsymbol{\eta})d\boldsymbol{\eta}}{p(x^n|\hat{\boldsymbol{\eta}})}.$$

We evaluate the ratio $\int p(x^n|\boldsymbol{\eta})\rho_J(\boldsymbol{\eta})d\boldsymbol{\eta}/p(x^n|\hat{\boldsymbol{\eta}})$.

We can write

$$\begin{aligned} p(x^n|\boldsymbol{\eta}) &= \prod_{s\in L,\ x\in\mathcal{X}} \exp(n_{x|s}\log\eta_{x|s}) \qquad (26)\\ &= \prod_{s\in L,\ x\in\mathcal{X}} \exp(n\frac{n_s}{n}\frac{n_{x|s}}{n_s}\log\eta_{x|s})\\ &= \prod_{s\in L}\exp(n\hat{p}_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\eta_{x|s}). \end{aligned}$$

Therefore, we have

$$\begin{aligned} &\frac{\int p(x^n|\boldsymbol{\eta})\rho_J(\boldsymbol{\eta})d\boldsymbol{\eta}}{p(x^n|\hat{\boldsymbol{\eta}})}\\ &= \int\Big(\prod_s \frac{e^{n\hat{p}_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\eta_{x|s}}}{e^{n\hat{p}_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\hat{\eta}_{x|s}}}\Big)\rho_J(\boldsymbol{\eta})d\boldsymbol{\eta}. \end{aligned}$$

Here, recall

$$\rho_J(\boldsymbol{\eta}) = \frac{1}{C_J}\prod_s \mu_s(\boldsymbol{\eta})^{d/2}D_{(1/2)}(\boldsymbol{\eta}_s).$$

We evaluate this integration (denoted as V) by Laplace approximation. We define a neighborhood $B_{n,s}$ of $\hat{\boldsymbol{\eta}}_s$ as

$$B_{n,s} = \{\boldsymbol{\eta}_s : \hat{p}_s(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s) \leq \frac{4d\log n}{n}\}$$

where $I(\hat{\boldsymbol{\eta}})$ is the same one as (15). We show that for sufficiently large n, $B_{n,s}$ is included in $H_s^{(\epsilon_n)}$. Note that all eigenvalues of $I(\boldsymbol{\eta}_s)$ are larger than 1 for arbitrary $\boldsymbol{\eta}_s \in H_s^{(0)}$. Hence, the minimum eigenvalue of $\hat{p}_s I(\hat{\boldsymbol{\eta}}_s)$ is larger than $\hat{p}_s$, which by Lemma 1 satisfies

$$\begin{aligned} \hat{p}_s &> \mu_s(\hat{\boldsymbol{\eta}})\exp(-\frac{C_1}{n\epsilon_n^{r+k}})\\ &= \mu_s(\hat{\boldsymbol{\eta}})\exp(-\frac{C_1}{n^{1-a(r+k)}})\\ &> \mu_s(\hat{\boldsymbol{\eta}})\exp(-\frac{C_1}{n^{r/(2r+k)}}). \end{aligned}$$

The second inequality here follows from

$$1-a(r+k) > 1-(r+k)/(2r+k) = r/(2r+k)$$

(recall $a < 1/(2r+k)$). Then, by Proposition 1, the minimum eigenvalue of $\hat{p}_s I(\hat{\boldsymbol{\eta}}_s)$ is larger than ϵ_n^k/e for all n such that $n^{r/(2r+k)} > C_1$. Therefore, with Lemma 1, the diameter of $B_{n,s}$ is smaller than

$$\frac{\sqrt{4de\log n}}{\sqrt{n\epsilon_n^k}} = C_4 n^{-(1-ak)/2}\sqrt{\log n}. \qquad (27)$$

Its ratio to $\epsilon_n = n^{-a}$ converges to 0 as n goes to infinity, since $(1-ak)/2 > a$. Hence, $B_n \stackrel{\text{def}}{=} \prod_s B_{n,s}$ is included in $H^{(\epsilon_n)}$ for sufficiently large n. Hence, we have

$$\begin{aligned} V &\geq \int_{B_n}\rho_J(\boldsymbol{\eta})\prod_s \frac{e^{n\hat{p}_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\eta_{x|s}}}{e^{n\hat{p}_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\hat{\eta}_{x|s}}}d\boldsymbol{\eta}\\ &\geq \inf_{\boldsymbol{\eta}\in B_n}\rho_J(\boldsymbol{\eta})\int_{B_n}\prod_s \frac{e^{n\hat{p}_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\eta_{x|s}}}{e^{n\hat{p}_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\hat{\eta}_{x|s}}}d\boldsymbol{\eta}\\ &= \beta_n\prod_s\int_{B_{n,s}}\frac{e^{n\hat{p}_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\eta_{x|s}}}{e^{n\hat{p}_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\hat{\eta}_{x|s}}}d\boldsymbol{\eta}_s\\ &\geq \beta_n\prod_s\int_{B_{n,s}} e^{-n\hat{p}_s e^{\gamma_n}(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s)^t I(\hat{\eta}_s)(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s)/2}d\boldsymbol{\eta}_s \end{aligned}$$

where we have used Taylor's theorem in the manipulation from the third line to the fourth line, and we let

$$\beta_n \stackrel{\text{def}}{=} \inf_{\boldsymbol{\eta}\in B_n}\rho_J(\boldsymbol{\eta}) \qquad (28)$$

$$\hat{I}_{xy}(\boldsymbol{\eta}_s) \stackrel{\text{def}}{=} \frac{\delta_{xy}\hat{\eta}_{x|s}}{(\eta_{x|s})^2} + \frac{\hat{\eta}_{0|s}}{(\eta_{0|s})^2}$$

$$e^{\gamma_n} \stackrel{\text{def}}{=} \sup\frac{(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s)^t\hat{I}(\boldsymbol{\eta}_s')(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s)}{(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s)}. \qquad (29)$$

In (29), the supremum is taken for all x_{-k+1}^n: $\hat{\boldsymbol{\eta}} \in H^{(\epsilon_n)}$, for all $\boldsymbol{\eta}_s \in B_{n,s}\setminus\{\hat{\boldsymbol{\eta}}_s\}$, and for all $\boldsymbol{\eta}_s' \in B_{n,s}$. The quantities $\hat{I}_{xy}(\boldsymbol{\eta}_s)$ provide the empirical Fisher information for the Bernoulli sources. Note that $\hat{I}_{xy}(\hat{\boldsymbol{\eta}}_s) = I_{xy}(\hat{\boldsymbol{\eta}}_s)$ holds.

We are going to show that the following two inequalities uniformly hold for all $x_{-k+1}^n : \hat{\boldsymbol{\eta}} \in H^{(\epsilon_n)}$:

$$\beta_n \geq (1-o(1))\rho_J(\hat{\boldsymbol{\eta}}) \qquad (30)$$

$$\int_{B_{n,s}} e^{-n\hat{p}_s e^{\gamma_n}(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s-\hat{\boldsymbol{\eta}}_s)/2}d\boldsymbol{\eta}_s \qquad (31)$$

$$\geq (1-o(1))\frac{(2\pi)^{d/2}}{n^{d/2}\mu_s(\hat{\boldsymbol{\eta}})^{d/2}D_{(1/2)}(\hat{\boldsymbol{\eta}}_s)}$$

where $o(1)$ converges to 0 as n goes to infinity. These inequalities imply $V \geq (1-o(1))(2\pi/n)^{d\ell/2}/C_J$.

As for (30), note that we have

$$\begin{aligned}\inf_{\hat{\boldsymbol{\eta}}\in H^{(\epsilon_n)}} \inf_{\boldsymbol{\eta}\in B_n} \frac{D_{1/2}(\boldsymbol{\eta}_s)}{D_{1/2}(\hat{\boldsymbol{\eta}}_s)} &= \inf_{\hat{\boldsymbol{\eta}}\in H^{(\epsilon_n)}} \inf_{\boldsymbol{\eta}\in B_n} \prod_x \frac{\sqrt{\hat{\eta}_{x|s}}}{\sqrt{\eta_{x|s}}} \\ &\geq \prod_x \frac{\sqrt{2\epsilon_n}}{\sqrt{2\epsilon_n + d(B_{n,s})}}\end{aligned} \tag{32}$$

where we let $d(B_{n,s})$ denote the diameter of $B_{n,s}$. Since $d(B_{n,s})/\epsilon_n$ converges to 0 (recall (27)), the last expression converges to 1 as n goes to infinity. We can also show

$$\sup_{\boldsymbol{\eta}\in B_n} \left|\log \frac{\mu_s(\boldsymbol{\eta})}{\mu_s(\hat{\boldsymbol{\eta}})}\right| \leq \frac{C_5 n^{-(1-ak)/2}}{\epsilon_n^r} = C_5 n^{-(1-ak-2ra)/2}$$

in the same way as we obtain (25). (Recall that $C_4 n^{-(1-ak)/2}$ is an upper bound on the diameter of $B_{n,s}$.) Hence

$$\inf_{\boldsymbol{\eta}\in B_n} \frac{\mu_s(\boldsymbol{\eta})}{\mu_s(\hat{\boldsymbol{\eta}})} \geq \exp(-C_5 n^{-(1-ak-2ar)/2}) \to 1 \ \ (n\to\infty)$$

since $a < 1/(2r+k)$. Together with (32), we have (30).

As for (31), first we are going to show $\gamma_n \to 0$ as n goes to infinity. Note that

$$(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)^t \hat{I}(\boldsymbol{\eta}'_s)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s) = \sum_{x\in\mathcal{X}} \frac{\hat{\eta}_{x|s}}{(\eta'_{x|s})^2}(\eta_{x|s} - \hat{\eta}_{x|s})^2$$

holds. Let $a_x \stackrel{\text{def}}{=} \hat{\eta}_{x|s}/(\eta'_{x|s})^2$ and $b_x \stackrel{\text{def}}{=} 1/\hat{\eta}_{x|s}$. Then, we can write

$$\frac{(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)^t \hat{I}(\boldsymbol{\eta}'_s)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)}{(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)} = \frac{\sum_{x\in\mathcal{X}} a_x(\eta_{x|s} - \hat{\eta}_{x|s})^2}{\sum_{x\in\mathcal{X}} b_x(\eta_{x|s} - \hat{\eta}_{x|s})^2}.$$

Since

$$\frac{\sum_{x\in\mathcal{X}} a_x(\eta_{x|s} - \hat{\eta}_{x|s})^2}{\sum_{x\in\mathcal{X}} b_x(\eta_{x|s} - \hat{\eta}_{x|s})^2} \leq \max_{x\in\mathcal{X}} \frac{a_x}{b_x}$$

holds, we have

$$e^{\gamma_n} \leq \sup_{\hat{\boldsymbol{\eta}}\in H^{(\epsilon_n)}} \sup_{\boldsymbol{\eta}'_s\in B_{n,s}} \max_{x\in\mathcal{X}} \Big(\frac{\hat{\eta}_{x|s}}{\eta'_{x|s}}\Big)^2.$$

In a manner similar to the evaluation of (32), we have

$$\gamma_n \leq 2\log\big(1 + \frac{C_3}{\sqrt{4\epsilon_n^2 n_s}}\big) \leq \frac{2C_3}{\sqrt{4\epsilon_n^2 n_s}}.$$

This converges to 0 as n goes to infinity. Next we will show

$$\begin{aligned}&\int_{B_{n,s}} e^{-n\hat{p}_s e^{\gamma_n}(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)/2} d\boldsymbol{\eta}_s \\ &\qquad \geq \frac{e^{-d\gamma_n}(2\pi)^{d/2}(1-n^{-d/2})}{n^{d/2}\sqrt{\det(\hat{p}_s I(\hat{\boldsymbol{\eta}}_s))}}.\end{aligned} \tag{33}$$

The integral over $B_{n,s}$ is equal to the integral over the whole space minus the compliment $B^c_{n,s}$. Thus, the integral on the left-hand side is equal to

$$\sqrt{\frac{(2\pi)^d}{\det(n\hat{p}_s e^{\gamma_n} I(\hat{\boldsymbol{\eta}}_s))}} - I_2 \tag{34}$$

which is

$$\sqrt{\frac{e^{-d\gamma_n}(2\pi)^d}{n^d \det(\hat{p}_s I(\hat{\boldsymbol{\eta}}_s))}} - I_2 \tag{35}$$

where

$$I_2 \stackrel{\text{def}}{=} \int_{B^c_{n,s}} e^{-e^{\gamma_n} n\hat{p}_s(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)/2} d\boldsymbol{\eta}_s. \tag{36}$$

Abbreviate $Q(\boldsymbol{\eta}_s) = \hat{p}_s(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)/2$. In $B^c_{n,s}$, we have $Q(\boldsymbol{\eta}_s) > 4d\log n/n$ and hence the exponent in the integral (36) satisfies $e^{\gamma_n} nQ(\boldsymbol{\eta}_s) \geq e^{\gamma_n} Q(\boldsymbol{\eta}_s) + (n-1)e^{\gamma_n} 4d\log n$, where we have reserved the $e^{\gamma_n} Q(\boldsymbol{\eta}_s)$ part to preserve integrability. Accordingly, we have

$$\begin{aligned}I_2 \leq \exp(-\frac{e^{\gamma_n}(n-1)2d\log n}{n}) \\ \cdot \int_{B^c_{n,s}} e^{-e^{\gamma_n}\hat{p}_s(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)/2} d\boldsymbol{\eta}_s.\end{aligned}$$

Bounding it further by enlarging the last factor, integrating over $\Re^d$, yields

$$I_2 \leq \exp(-\frac{e^{\gamma_n}(n-1)2d\log n}{n}) \sqrt{\frac{e^{-d\gamma_n}(2\pi)^d}{\det(\hat{p}_s I(\hat{\boldsymbol{\eta}}_s))}}.$$

Since $\gamma_n = o(1)$, the inequality $e^{\gamma_n}(n-1)2d/n \geq d$ holds for sufficiently large n, and hence

$$I_2 \leq \frac{1}{n^{d/2}} \sqrt{\frac{e^{-d\gamma_n}(2\pi)^d}{n^d \det(\hat{p}_s I(\hat{\boldsymbol{\eta}}_s))}}$$

holds for sufficiently large n. Therefore, (34) yields

$$\begin{aligned}&\int_{B_{n,s}} e^{-n\hat{p}_s \exp(\gamma_n)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)/2} d\boldsymbol{\eta}_s \\ &\qquad = (1-n^{-d/2}) \sqrt{\frac{\exp(-d\gamma_n)(2\pi)^d}{n^d \det(\hat{p}_s I(\hat{\boldsymbol{\eta}}_s))}}.\end{aligned}$$

This is (33) as desired.

Since $\sqrt{\det(\hat{p}_s I(\hat{\boldsymbol{\eta}}_s))} = \hat{p}_s^{d/2} D_{(1/2)}(\hat{\boldsymbol{\eta}}_s)$, (33) yields

$$\begin{aligned}&\int_{B_{n,s}} e^{-n\hat{p}_s e^{\gamma_n}(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)^t I(\hat{\boldsymbol{\eta}}_s)(\boldsymbol{\eta}_s - \hat{\boldsymbol{\eta}}_s)/2} d\boldsymbol{\eta}_s \\ &\qquad \geq \frac{e^{-d\gamma_n}(2\pi)^{d/2}(1-n^{-d/2})}{n^{d/2}\hat{p}_s^{d/2} D_{(1/2)}(\hat{\boldsymbol{\eta}}_s)}.\end{aligned}$$

By Lemma 1 and since $\gamma_n \to 0$, this implies (31).

Since (30) and (31) hold, we have

$$V \geq \frac{1-o(1)}{C_J}\Big(\frac{2\pi}{n}\Big)^{d\ell/2}$$

where $o(1)$ converges to 0 as n goes to infinity. Therefore, we have

$$\sup_{x^n:\hat{\boldsymbol{\eta}}\in H^{(\epsilon_n)}} \frac{p(x^n|\hat{\boldsymbol{\eta}})}{m_n(x^n)} \leq \frac{C_J}{(1-o(1))((1-\kappa_n))}\frac{n^{d\ell/2}}{(2\pi)^{\ell d/2}}.$$

This implies

$$\sup_{x^n:\hat{\boldsymbol{\eta}}\in H^{(\epsilon_n)}} \log\frac{p(x^n|\hat{\boldsymbol{\eta}})}{m_n(x^n)} \leq \frac{d\ell}{2}\log\frac{n}{2\pi} + \log C_J + o(1). \tag{37}$$

Part II (near boundaries): Now, we consider the case in which $\hat{\boldsymbol{\eta}} \notin H^{(\epsilon_n)}$. We use the second term in the mixture m_n as

$$\frac{m_n(x^n)}{p(x^n|\hat{\boldsymbol{\eta}})} \geq \frac{\kappa_n m_{(\alpha)}(x^n)}{p(x^n|\hat{\boldsymbol{\eta}})} = \frac{\kappa_n \int p(x^n|\boldsymbol{\eta})\prod_s D_{(\alpha)}(\eta_s)d\boldsymbol{\eta}_s}{p(x^n|\hat{\boldsymbol{\eta}}) \quad C_{(\alpha)}^{\ell}}.$$

With the prior of product form, this becomes a product of integrals. Use

$$\frac{p(x^n|\boldsymbol{\eta})}{p(x^n|\hat{\boldsymbol{\eta}})} = \prod_{s\in L}\frac{\exp(n_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\eta_{x|s})}{\exp(n_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\hat{\eta}_{x|s})} = \prod_{s\in L}\exp\Big(-n_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\frac{\hat{\eta}_{x|s}}{\eta_{x|s}}\Big).$$

Then, we have

$$\frac{\int p(x^n|\boldsymbol{\eta})\prod_s D_{(\alpha)}(\boldsymbol{\eta}_s)d\boldsymbol{\eta}_s}{p(x^n|\hat{\boldsymbol{\eta}}) \quad C_{(\alpha)}^{\ell}} = \prod_{s\in E} L_s$$

where

$$L_s \stackrel{\text{def}}{=} \int \exp\Big(-n_s\sum_{x\in\mathcal{X}}\hat{\eta}_{x|s}\log\frac{\hat{\eta}_{x|s}}{\eta_{x|s}}\Big)\frac{D_{(\alpha)}(\boldsymbol{\eta}_s)}{C_{(\alpha)}}d\boldsymbol{\eta}_s = \frac{\int \exp(\sum_{x\in\mathcal{X}} n_{x|s}\log\eta_{x|s})D_{(\alpha)}(\boldsymbol{\eta}_s)d\boldsymbol{\eta}_s}{\exp(\sum_{x\in\mathcal{X}} n_{x|s}\log\hat{\eta}_{x|s}) \quad C_{(\alpha)}}$$

and E is the set of states such that $n_s > 0$.

Split the states in E into subsets $E_1 = \{s|\hat{\boldsymbol{\eta}}_s \notin H_s^{(\epsilon_n)}\} \cap E$ and $E_2 = \{s|\hat{\boldsymbol{\eta}}_s \in H_s^{(\epsilon_n)}\} \cap E$. For $\hat{\boldsymbol{\eta}} \notin H^{(\epsilon_n)}$, we are assured that E_1 is not empty.

Note that $-\log L_s$ is in a form of regret of the mixture by the Dirichlet prior with $\alpha < 1/2$ for the memoryless case. Since the Dirichlet prior with $\alpha < 1/2$ has higher value than the Jeffreys prior near boundaries of H, the quantity $-\log L_s$ for $s \in E_1$ is smaller than $(d/2)\log n$. Indeed, if $s \in E_1$, then there is a symbol x such that $\hat{\eta}_{x|s} \leq 1/n^a \leq 1/n_s^a$, so $n_{x|s} \leq n_s^{1-a}$. Consequently, adapting Xie and Barron's Lemma ([28, Lemma 4]) for the present case and for $s \in E_1$

$$-\log L_s \leq \Big(\frac{d}{2} - (\frac{1}{2}-\alpha)a\Big)\log n_s + K_d\log\frac{1}{\alpha} \tag{38}$$

holds, where K_d is a constant depending on only d.

As for $s \in E_2$, we use the following bound, which holds for all $s \in E$:

$$-\log L_s \leq \frac{d}{2}\log n_s + C_9. \tag{39}$$

This inequality (39) is derived by Lemma 1 of [28]. The lemma is a uniform bound on the regret of the Jeffreys mixture for memoryless case, and can be applied to our case by noting $D_{(\alpha)}(\boldsymbol{\eta}_s)d\boldsymbol{\eta}_s \geq D_{(1/2)}(\boldsymbol{\eta}_s)d\boldsymbol{\eta}_s$ and then

$$L_s > \frac{\int \exp(\sum_{x\in\mathcal{X}} n_{x|s}\log\eta_{x|s})D_{(1/2)}(\boldsymbol{\eta}_s)d\boldsymbol{\eta}_s}{\exp(\sum_{x\in\mathcal{X}} n_{x|s}\log\hat{\eta}_{x|s})C_{(\alpha)}} = \frac{C_{(1/2)}}{C_{(\alpha)}}\frac{\int \exp(\sum_{x\in\mathcal{X}} n_{x|s}\log\eta_{x|s})D_{(1/2)}(\boldsymbol{\eta}_s)d\boldsymbol{\eta}_s}{\exp(\sum_{x\in\mathcal{X}} n_{x|s}\log\hat{\eta}_{x|s})C_{(1/2)}}.$$

Here, $D_{(1/2)}(\boldsymbol{\eta}_s)/C_{(1/2)}$ is the Jeffreys prior for the multinomial Bernoulli model.

Hence, we have

$$\log\frac{p(x^n|\hat{\boldsymbol{\eta}})}{m_{(\alpha)}(x^n)} = \sum_{s\in E}(-\log L_s) \leq \sum_{s\in E_1}\Big(\Big(\frac{d}{2} - (\frac{1}{2}-\alpha)a\Big)\log n_s + K_d\log\frac{1}{\alpha}\Big) + \sum_{s\in E_2}\Big(\frac{d}{2}\log n_s + C_9\Big)$$

which is not more than

$$\sum_{s\in E_1}\Big(\frac{d}{2}-\iota\Big)\log n_s + \sum_{s\in E_2}\frac{d}{2}\log n_s + C_{10}, \tag{40}$$

where

$$C_{10} \stackrel{\text{def}}{=} \ell\max\{K_d\log\frac{1}{\alpha}, C_9\}$$
$$\iota \stackrel{\text{def}}{=} (1/2-\alpha)a.$$

We claim that (40) is less than

$$\Big(\frac{d\ell}{2}-\iota\Big)\log\frac{n}{\ell-2\iota/d} + C_{10}. \tag{41}$$

Since (40) is maximized when $|E_1| = 1$ for any configuration of $\{n_s\}$, it is the worst case. Then, the maximum of (40) is achieved when $n_s = n(d/2-\iota)/(|E|d/2-\iota)$ for $s \in E_1$ and $n_s = (nd/2)/(|E|d/2-\iota)$ for $s \in E_2$. This provides an upper bound which is no more than

$$\Big(\frac{d}{2}-\iota\Big)\log\frac{n(d/2-\iota)}{|E|d/2-\iota} + \frac{d|E_2|}{2}\log\frac{nd/2}{|E|d/2-\iota} + C_{10}$$

whose dependence on $|E|$ is of the form

$$\Big(\frac{d}{2}-\iota\Big)\log\frac{1}{|E|d/2-\iota} + \frac{d(|E|-1)}{2}\log\frac{dn/2}{|E|d/2-\iota} = \frac{d|E|-2\iota}{2}\log\frac{1}{|E|d/2-\iota} + \frac{d(|E|-1)}{2}\log\frac{dn}{2}.$$

Its derivative with respect to $|E|$ is positive when

$$n \geq e|E| - \frac{2e\iota}{d}.$$

Whence for $n \geq e\ell$, the largest $|E| = \ell$ is the worst case, which provides the following upper bound on (40):

$$\left(\frac{d}{2} - \iota\right) \log \frac{n(d/2-\iota)}{d\ell/2 - \iota} + \frac{d(\ell-1)}{2} \log \frac{nd/2}{d\ell/2-\iota} + C_{10}$$

which is less than (41) by $(d\ell/2-\iota)\cdot\log(d/(d-2\iota))$. Therefore, we have

$$\begin{aligned} &\sup_{x^n:\hat{\boldsymbol{\eta}} \notin H^{(\epsilon_n)}} \log \frac{p(x^n|\hat{\boldsymbol{\eta}})}{m_n(x^n)} \\ &\quad \leq \left(\frac{d\ell}{2} - \iota\right) \log \frac{n}{\ell - 2\iota/d} + \log \frac{1}{\kappa_n} + C_{10} \\ &\quad \leq \left(\frac{d\ell}{2} - \iota + b\right) \log \frac{n}{\ell - 2\iota/d} + C_{10} \\ &\quad \leq \left(\frac{d\ell}{2} - \iota + b\right) \log n + C_{11} \end{aligned} \tag{42}$$

where $C_{11} = C_{10} - (d\ell/2 - \iota + b)\log(\ell - 2\iota/d)$.

Since $b < \iota = (1/2 - \alpha)a$ is assumed, the expression (42) is smaller than the right-hand side of (37), when $(\iota - b)\log n$ exceeds the constant $(d\ell/2)\log\pi - \log C_J + C_{11}$. This completes the proof of Theorem 1.

V. Proof of the Approximation Formula

In this section, we give the proof of Theorem 2. For the purpose of abbreviation, we define two functions F and G as follows. Define F on $\Re \times \Theta_s \times H_s$ as

$$\begin{aligned} F(m, \boldsymbol{\theta}_s, \boldsymbol{\eta}'_s) &\stackrel{\text{def}}{=} \exp(m(\sum_{x \in \mathcal{X}'} \theta_{x|s} \eta'_{x|s} - \psi(\boldsymbol{\theta}_s))) \\ &= \prod_{x \in \mathcal{X}} (\eta_{x|s})^{m\eta'_{x|s}} \end{aligned} \tag{43}$$

where ψ and $\boldsymbol{\theta}_s$ are the same ones as in (13). In particular, recall $\theta_{x|s} = \log(\eta_{x|s}/\eta_{0|s})$. Note that the following holds: for each $s \in L$, let m_s denote a real number and $\boldsymbol{m}$ denote a vector $(m_{s_1}, \ldots, m_{s_\ell})$. Define G on $\Re^\ell \times \Theta \times H$ as

$$G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') \stackrel{\text{def}}{=} \prod_{s \in L} F(m_s, \boldsymbol{\theta}_s, \boldsymbol{\eta}'_s) = \prod_{s \in L, x \in \mathcal{X}} (\eta_{x|s})^{m_s \eta'_{x|s}}.$$

Then, defining $\boldsymbol{n} \stackrel{\text{def}}{=} (n_{s_1}, \ldots, n_{s_\ell})$, we have

$$p(x^n | x^0_{-k+1}, \boldsymbol{\eta}) = G(\boldsymbol{n}, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}).$$

Since $(\partial/\partial\theta_{x|s})\psi(\boldsymbol{\theta}_s) = \eta_{x|s}$ and

$$\frac{\partial \log G(\boldsymbol{n}, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}})}{\partial \theta_{x|s}} = \frac{\partial \log F(m_s, \boldsymbol{\theta}_s, \boldsymbol{\eta}'_s)}{\partial \theta_{x|s}}$$

we have

$$\frac{\partial \log G(\boldsymbol{n}, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}})}{\partial \theta_{x|s}} = m_s(\eta'_{x|s} - \eta_{x|s}). \tag{44}$$

Also, recalling the definition of g(14), we have

$$\frac{\partial^2}{\partial \theta_{y|s} \partial \theta_{x|s}} \log G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') = -m_s g_{xy}(\boldsymbol{\theta}_s). \tag{45}$$

Then, since $g(\boldsymbol{\theta}_s)$ is positive definite, $\log G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}')$ is strictly concave with respect to $\boldsymbol{\theta}$, whenever each m_s is positive.

Finally, we let

$$W^{(w)}(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') \stackrel{\text{def}}{=} \frac{w(\boldsymbol{\theta}) G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}')}{\int w(\boldsymbol{\theta}) G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') d\boldsymbol{\theta}}. \tag{46}$$

Then, we have

$$q(x|x^n) = \int \eta_{x|\tau(x^n)} W^{(w)}(\boldsymbol{n}, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) d\boldsymbol{\theta}.$$

Note that w need not be normalized in this expression, since it remains unchanged when we multiply w by a positive constant. Hence, we assume that w does not have to be a probability density hereafter.

First, we prove the following lemma.

Lemma 2: Let K be a compact set included in the interior of H. Let $W^{(w)}$ be the function defined as (46). Suppose that w in $W^{(w)}$ be a positive-valued function, which is integrable over Θ. We assume that w satisfies Assumption 1 and that $\min_{s \in L} m_s \geq m_{\min}$ holds. Then, for all $s \in L$ and for all $x \in \mathcal{X}'$, the following holds, uniformly for $\boldsymbol{\eta}' \in K$:

$$\begin{aligned} &\int \eta_{x|s} W^{(w)}(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') d\boldsymbol{\theta} \\ &\quad = \eta'_{x|s} + \frac{1}{m_s} \left.\frac{\partial \log w(\boldsymbol{\theta})}{\partial \theta_{x|s}}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}'} + O\left(\frac{\sqrt{\log m}}{m_s \sqrt{m}}\right). \end{aligned}$$

Proof: In this proof, we let W denote $W^{(w)}$, omitting (w). Partial differentiating $G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}')w(\boldsymbol{\theta})$ with respect to $\theta_{x|s}$, we have

$$\begin{aligned} &\frac{\partial [G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}')w(\boldsymbol{\theta})]}{\partial \theta_{x|s}} \\ &\quad = m_s(\eta'_{x|s} - \eta_{x|s}) G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') w(\boldsymbol{\theta}) \\ &\quad + \frac{\partial \log w(\boldsymbol{\theta})}{\partial \theta_{x|s}} G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') w(\boldsymbol{\theta}) \end{aligned} \tag{47}$$

where we have used (44). The second term on the right-hand side is integrable because of Assumption 1 and the first term on the right-hand side is integrable because $|\eta_{x|s}|$ is bounded. Therefore, the left-hand side is also integrable. Integrating both sides over Θ, and doing some manipulation, we have

$$\begin{aligned} &\int \eta_{x|s} G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') w(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\quad = \eta'_{x|s} \int G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') w(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\quad + \frac{1}{m_s} \int \frac{\partial \log w(\boldsymbol{\theta})}{\partial \theta_{x|s}} G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') w(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\quad - \frac{1}{m_s} \int \frac{\partial G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}') w(\boldsymbol{\theta})}{\partial \theta_{x|s}} d\boldsymbol{\theta}. \end{aligned} \tag{48}$$

We can show that the third term on the right-hand side is zero. Indeed, by the Fubini's theorem, we have

$$\begin{aligned} &\int \frac{\partial (w(\boldsymbol{\theta}) G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}'))}{\partial \theta_{x|s}} d\boldsymbol{\theta} \\ &\quad = \int [w(\boldsymbol{\theta}) G(\boldsymbol{m}, \boldsymbol{\theta}, \boldsymbol{\eta}')]_{\theta_{x|s} = -\infty}^{\theta_{x|s} = \infty} d\bar{\boldsymbol{\theta}} = 0 \end{aligned}$$

where $\bar{\boldsymbol{\theta}}$ is $\ell \cdot (d-1)$-dimensional vector which is obtained by removing the element $\theta_{x|s}$ from the vector $\boldsymbol{\theta}$. Hence, dividing both sides of (48) by $\int w(\boldsymbol{\theta})G(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}$, we have

$$\int \eta_{x|s} W(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\theta = {\eta'}_{x|s} + \frac{1}{m_s}\int \frac{\partial \log w(\boldsymbol{\theta})}{\partial \theta_{x|s}} W(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}.$$

Therefore, it suffices for obtaining the claim of the Lemma to show that

$$\int \frac{\partial \log w(\boldsymbol{\theta})}{\partial \theta_{x|s}} W(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta} = \left.\frac{\partial \log w(\boldsymbol{\theta})}{\partial \theta_{x|s}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} + O\Big(\frac{\sqrt{\log m}}{\sqrt{m}}\Big) \tag{49}$$

holds uniformly for $\boldsymbol{\eta}' \in K$. We use Laplace integration to prove this. Let

$$h(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\partial \log w(\boldsymbol{\theta})}{\partial \theta_{x|s}}.$$

Since $w(\boldsymbol{\theta}) > 0$ for $\boldsymbol{\theta} \in \Theta$ and since $w(\boldsymbol{\theta})$ is of class C^2 in Θ, $\log w(\boldsymbol{\theta})$ is of class C^2 in Θ. Therefore, $h(\boldsymbol{\theta})$ is of class C^1 in Θ.

Define a neighborhood of $\boldsymbol{\theta}'_s$ ($\theta'_{x|s} \stackrel{\text{def}}{=} \log(\eta'_{x|s}/\eta'_{0|s})$) in $\Re^d$ as

$$N_\delta(\boldsymbol{\theta}'_s|s) \stackrel{\text{def}}{=} \{\boldsymbol{\theta}_s : (\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)^t g(\theta'_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s) \le \delta^2\}.$$

Further define

$$N'_\delta = N_\delta(\boldsymbol{\theta}') \stackrel{\text{def}}{=} \prod_{s\in L} N_\delta(\boldsymbol{\theta}'_s|s)$$

where we assume

$$\delta^2 = \frac{d\cdot \ell \log m}{m}.$$

From (46), we have

$$W(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}') = \frac{w(\boldsymbol{\theta})G(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')}{\int w(\boldsymbol{\theta})G(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}} = \frac{w(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')}{\int w(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}}$$

where we let

$$\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}') \stackrel{\text{def}}{=} G(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')/G(\boldsymbol{m},\boldsymbol{\theta}',\boldsymbol{\eta}').$$

Then, we will evaluate $\int h(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')w(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $\int \bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')w(\boldsymbol{\theta})d\boldsymbol{\theta}$.

Let $v(\theta)$ denote a function $h(\boldsymbol{\theta})w(\boldsymbol{\theta})$ or $w(\boldsymbol{\theta})$. Assume $v(\boldsymbol{\theta}') \ge 0$ without loss of generality; then, we have

$$\int v(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta} = \int_{N'_\delta} v(\theta)\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta} + \int_{\Theta\setminus N'_\delta} v(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}. \tag{50}$$

Using Taylor's theorem, we have

$$\log \bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}') = \sum_s \log \frac{F(m_s,\boldsymbol{\theta}_s,\boldsymbol{\eta}'_s)}{F(m_s,\boldsymbol{\theta}'_s,\boldsymbol{\eta}'_s)} = -\frac{\sum_s m_s(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)^t g(\boldsymbol{q}_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)}{2}$$

where $\boldsymbol{q}_s = \epsilon\boldsymbol{\theta}_s + (1-\epsilon)\boldsymbol{\theta}'_s$ with $\epsilon \in [0,1]$ ($s = s_1,\ldots,s_{(d+1)^k}$). Hence, we have

$$G(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}') = \exp\Big(-\frac{\sum_s m_s(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)^t g(\boldsymbol{q}_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)}{2}\Big).$$

Since $\{\boldsymbol{\theta}(\boldsymbol{\eta}) : \boldsymbol{\eta} \in K\}$ is compact

$$1 - C_1\delta \le \frac{\sum_s m_s(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)}{\sum_s m_s(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}_s{}')(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)} \le 1 + C_1\delta$$

holds for sufficiently large m (small δ), for all $\boldsymbol{\theta}' \in \{\boldsymbol{\theta}(\boldsymbol{\eta}) : \boldsymbol{\eta} \in K\}$, and for all $\boldsymbol{\theta} \in N_\delta(\boldsymbol{\theta}')$, where C_1 is a certain constant. (Hereafter, let C_i ($i = 1,2,\ldots$) denote a certain positive constant.) Hence, we have

$$\begin{aligned}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}') &\le \exp\Big(-\frac{(1-C_1\delta)\sum_s m_s(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}_s{}')(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)}{2}\Big)\\ \bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}') &\ge \exp\Big(-\frac{(1+C_1\delta)\sum_s m_s(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}_s{}')(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s)}{2}\Big).\end{aligned} \tag{51}$$

Using these inequalities, we evaluate the second term of (50). Let $\mathbf{1}$ denote the ℓ-dimensional vector $(1,..,1)$ and $\tilde{\boldsymbol{m}} \stackrel{\text{def}}{=} \boldsymbol{m} - m_{\min}\mathbf{1}$. Noting that $\log \bar{G}(\tilde{\boldsymbol{m}},\boldsymbol{\theta},\boldsymbol{\eta}') = \log G(\tilde{\boldsymbol{m}},\boldsymbol{\theta},\boldsymbol{\eta}') - \log G(\tilde{\boldsymbol{m}},\boldsymbol{\theta}',\boldsymbol{\eta}')$ is strictly concave with respect to $\boldsymbol{\theta} \in \Theta$, we have

$$\begin{aligned}&\sup_{\boldsymbol{\theta}\in\Theta\setminus N'_\delta} \bar{G}(\tilde{\boldsymbol{m}},\boldsymbol{\theta},\boldsymbol{\eta}')\\ &= \sup_{\boldsymbol{\theta}\in\partial N'_\delta} \bar{G}(\tilde{\boldsymbol{m}},\boldsymbol{\theta},\boldsymbol{\eta}')\\ &\le \sup_{\boldsymbol{\theta}\in\partial N'_\delta} e^{-\sum_s \tilde{m}_s(1-C_1\delta)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}_s{}')(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)/2}\\ &= \exp\Big(-\frac{\sum_s (m_s - m_{\min})(1-C_1\delta)\delta^2}{2}\Big)\\ &\le \exp(-(m - m_{\min}\ell)(1-C_1\delta)\delta^2/2)\\ &\le C_2\exp(-m\delta^2/2).\end{aligned}$$

Hence, we have

$$\begin{aligned}&\Big|\int_{\Theta\setminus N'_\delta} v(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')w(\boldsymbol{\theta})d\boldsymbol{\theta}\Big|\\ &= \Big|\int_{\Theta\setminus N'_\delta} v(\boldsymbol{\theta})\bar{G}(m_{\min}\mathbf{1},\boldsymbol{\theta},\boldsymbol{\eta}')\bar{G}(\tilde{\boldsymbol{m}},\boldsymbol{\theta},\boldsymbol{\eta}')w(\boldsymbol{\theta})d\boldsymbol{\theta}\Big|\\ &\le C_2\exp(-m\delta^2/2)\\ &\quad\cdot\Big|\int_{\Theta\setminus N'_\delta} v(\boldsymbol{\theta})\bar{G}(m_{\min}\mathbf{1},\boldsymbol{\theta},\boldsymbol{\eta}')w(\boldsymbol{\theta})d\boldsymbol{\theta}\Big|\\ &\le C_3\cdot\exp(-m\delta^2/2).\end{aligned} \tag{52}$$

Next, we evaluate the first term of (50). We have

$$\begin{aligned}&\left|\int_{N'_\delta}(v(\boldsymbol{\theta})-v(\boldsymbol{\theta}'))\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}\right|\\&\le \sup_{\boldsymbol{\theta}\in N'_\delta}|v(\boldsymbol{\theta})-v(\boldsymbol{\theta}')|\int_{N'_\delta}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}\\&=O(\delta)\int_{N'_\delta}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}.\end{aligned}$$

Hence, we have

$$\begin{aligned}&\int_{N'_\delta}v(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}\\&\quad=(v(\boldsymbol{\theta}')+O(\delta))\int_{N'_\delta}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}.\end{aligned}\tag{53}$$

For the upper bound on $\int_{N'_\delta}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}$, from (1) we have

$$\begin{aligned}&\int_{N'_\delta}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}\\&\quad\le\int_{N'_\delta}e^{-\sum_s m_s(1-C_1\delta)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)g(\boldsymbol{\theta}'_s)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)/2}d\boldsymbol{\theta}\\&\quad\le\int_{\Theta}e^{-\sum_s m_s(1-C_1\delta)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)g(\boldsymbol{\theta}'_s)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)/2}d\boldsymbol{\theta}\\&\quad=\prod_{s\in L}\frac{1}{\sqrt{(2\pi m_s(1-C_1\delta))^d\det(g(\boldsymbol{\theta}'_s))}}\\&\quad=\frac{1+O(\delta)}{\prod_{s\in L}\sqrt{(2\pi m_s)^d\det(g(\boldsymbol{\theta}'_s))}}.\end{aligned}\tag{54}$$

For the lower bound on $\int_{N'_\delta}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}$, from (51) we have

$$\begin{aligned}&\int_{N'_\delta}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}\\&\quad\ge\int_{N'_\delta}e^{-(1+C_1\delta)\sum_s m_s(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}'_s)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)/2}d\boldsymbol{\theta}\\&\quad=\int_{\Theta}e^{-(1+C_1\delta)\sum_s m_s(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}'_s)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)/2}d\boldsymbol{\theta}\\&\quad\quad-\int_{\Theta\setminus N'_\delta}e^{-(1+C_1\delta)\sum_s m_s(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}'_s)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)/2}d\boldsymbol{\theta}\\&\quad=\prod_{s\in L}\frac{1}{\sqrt{(2\pi m_s(1+C_1\delta))^d\det(g(\boldsymbol{\theta}'_s))}}\\&\quad\quad-\int_{\Theta\setminus N'_\delta}e^{-(1+C_1\delta)\sum_s m_s(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}'_s)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)/2}d\boldsymbol{\theta}.\end{aligned}$$

In the same manner as obtaining (52), we have

$$\begin{aligned}&\int_{\Theta\setminus N'_\delta}e^{-(1+C_1\delta)\sum_s m_s(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)^t g(\boldsymbol{\theta}'_s)(\boldsymbol{\theta}_s-\boldsymbol{\theta}'_s)/2}d\boldsymbol{\theta}\\&\quad\le C_4\exp(-m\delta^2/2).\end{aligned}$$

Hence, we have

$$\begin{aligned}&\int_{N'_\delta}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}\\&\quad\ge\frac{1}{\prod_s\sqrt{(2\pi m_s(1+C_1\delta))^d\det(g(\boldsymbol{\theta}'_s))}}-C_4e^{-m\delta^2/2}\\&\quad=\frac{1+O(\delta)}{\prod_s\sqrt{(2\pi m_s)^d\det(g(\boldsymbol{\theta}'_s))}}-C_4\cdot\exp(-m\delta^2/2)\\&\quad=\frac{1+O(\delta)-O(\exp(-m\delta^2/2))\cdot\prod_s\sqrt{m_s{}^d}}{\prod_s\sqrt{(2\pi m_s)^d\det(g(\boldsymbol{\theta}'_s))}}\\&\quad=\frac{1+O(\delta)+O(m^{d\cdot\ell/2}\cdot\exp(-m\delta^2/2))}{\prod_s\sqrt{(2\pi m_s)^d\det(g(\boldsymbol{\theta}'_s))}}.\end{aligned}\tag{55}$$

Hence, with (54), we have

$$\begin{aligned}&\int_{N'_\delta}\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}\\&\quad=\frac{1+O(\delta)+O(m^{d\cdot\ell/2}\cdot\exp(-m\delta^2/2))}{\prod_s\sqrt{(2\pi m_s)^d\det(g(\boldsymbol{\theta}'_s))}}.\end{aligned}$$

From this and (53), we have

$$\begin{aligned}&\int_{N'_\delta}v(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}\\&\quad=\frac{(v(\boldsymbol{\theta}')+O(\delta))\cdot(1+O(\delta)+O(m^{d\cdot\ell/2}e^{-m\delta^2/2})}{\prod_s\sqrt{(2\pi m_s)^d\det(g(\boldsymbol{\theta}'_s))}}\\&\quad=\frac{v(\boldsymbol{\theta}')+O(m^{d\cdot\ell/2}\cdot\exp(-m\delta^2/2))+O(\delta)}{\prod_s\sqrt{(2\pi m_s)^q\det(g(\boldsymbol{\theta}'_s))}}.\end{aligned}$$

By this equation and (52), we have

$$\begin{aligned}&\int_{\Theta}v(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}\\&\quad=\frac{v(\boldsymbol{\theta}')+O(m^{d\cdot\ell/2}\cdot\exp(-m\delta^2/2))+O(\delta)}{\prod_{s\in L}\sqrt{(2\pi m_s)^d\det(g(\boldsymbol{\theta}'_s))}}\\&\quad=\frac{v(\boldsymbol{\theta}')+O(\delta)}{\prod_{s\in L}\sqrt{(2\pi m_s)^d\det(g(\boldsymbol{\theta}'_s))}}.\end{aligned}$$

The last equality is obtained since $m^{d\cdot\ell/2}e^{-m\delta^2/2}=m^{-d\cdot\ell/2}\le\delta$ holds for large m. Recall that this has been proved for $v(\boldsymbol{\theta})=h(\boldsymbol{\theta})w(\boldsymbol{\theta})$ and $v(\boldsymbol{\theta})=w(\boldsymbol{\theta})$. Hence, we have

$$\begin{aligned}\int_{\Theta}h(\boldsymbol{\theta})W(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}&=\frac{\int_{\Theta}h(\boldsymbol{\theta})w(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\boldsymbol{\theta}}{\int_{\Theta}w(\boldsymbol{\theta})\bar{G}(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{\eta}')d\theta}\\&=\frac{h(\boldsymbol{\theta}')w(\boldsymbol{\theta}')+O(\delta)}{w(\boldsymbol{\theta}')+O(\delta)}\\&=h(\boldsymbol{\theta}')+O(\delta).\end{aligned}$$

This completes the proof of Lemma 2.

Proof of Theorem 3: First, we will prove (17). In Lemma 2, plug in $\boldsymbol{n}$ and $\hat{\boldsymbol{\eta}}$ into $\boldsymbol{m}$ and $\boldsymbol{\eta}'$, respectively. Then, since $\hat{\boldsymbol{\eta}} \in K$ holds for large n by the assumption, we have $O(1/n_s) = O(1/n)$ for all $s \in L$. Hence, we obtain (17).

Next, we prove (18). Let $\tilde{w}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} w(\boldsymbol{\theta})/w_{(1/2)}(\boldsymbol{\theta})$. We can prove that the density $\tilde{w}$ satisfies Assumption 1, provided w satisfies it. In fact

$$\begin{aligned}
&\frac{\partial \log \tilde{w}(\boldsymbol{\theta})}{\partial \theta_{x|s}} \prod_{t\in L, y\in\mathcal{X}} (\eta_{y|t})^{m\eta'_{y|t}} \cdot \tilde{w}(\boldsymbol{\theta}) \\
&\quad = \frac{\partial(\log w(\boldsymbol{\theta}) - \log w_{(1/2)}(\boldsymbol{\theta}))}{\partial \theta_{x|s}} \\
&\quad\quad \cdot \prod_{t\in L, y\in\mathcal{X}} (\eta_{y|t})^{m\eta'_{y|t}} \cdot \frac{w(\boldsymbol{\theta})}{w_{(1/2)}(\boldsymbol{\theta})} \\
&\quad = \frac{\partial(\log w(\boldsymbol{\theta}) - \log w_{(1/2)}(\boldsymbol{\theta}))}{\partial \theta_{x|s}} \\
&\quad\quad \cdot \prod_{t\in L, y\in\mathcal{X}} (\eta_{y|t})^{m\eta'_{y|t}-1/2} \cdot w(\boldsymbol{\theta}) \\
&= \frac{\partial \log w(\boldsymbol{\theta})}{\partial \theta_{x|s}} \prod_{t\in L, y\in\mathcal{X}} (\eta_{y|t})^{m\eta'_{y|t}-1/2} \cdot w(\boldsymbol{\theta}) \\
&- \frac{\partial \log w_{(1/2)}(\boldsymbol{\theta})}{\partial \theta_{x|s}} \prod_{t\in L, y\in\mathcal{X}} (\eta_{y|t})^{m\eta'_{y|t}-1/2} \cdot w(\boldsymbol{\theta})
\end{aligned}$$

and both terms in the last line are integrable when

$$m \geq m_{\min} + \frac{1}{2} \max_{s\in L, y\in\mathcal{X}} \frac{1}{\eta'_{y|s}}.$$

(For the second term, see the proof of Lemma 5 in Appendix C.)

Note that

$$\begin{aligned}
&\log\Big(\prod_{s\in L, y\in\mathcal{X}} (\eta_{y|s})^{n_s\hat{\eta}_{y|s}} w(\boldsymbol{\theta})\Big) \\
&\quad = \sum_{s\in L, y\in\mathcal{X}} n_s\hat{\eta}_{y|s} \log \eta_{y|s} + \log w(\boldsymbol{\theta}) \\
&\quad = \sum_{s\in L, y\in\mathcal{X}} (n_s\hat{\eta}_{y|s} + 1/2) \log \eta_{y|s} + \log \frac{w(\boldsymbol{\theta})}{w_{(1/2)}(\boldsymbol{\theta})} \\
&\quad = \sum_{s\in L, y\in\mathcal{X}} n_s(\hat{\eta}_{y|s} + 1/2n_s) \log \eta_{y|s} + \log \tilde{w}(\boldsymbol{\theta}) \\
&\quad = \sum_{s\in L, y\in\mathcal{X}} \Big(n_s + \frac{d+1}{2}\Big) \frac{\hat{\eta}_{y|s} + 1/2n_s}{1 + (d+1)/2n_s} \log \eta_{y|s} \\
&\quad + \log \tilde{w}(\boldsymbol{\theta}) \\
&\quad = \sum_{s\in L, y\in\mathcal{X}} \Big(n_s + \frac{d+1}{2}\Big) \frac{n_{y|s} + 1/2}{n_s + (d+1)} \log \eta_{y|s} \\
&\quad + \log \tilde{w}(\boldsymbol{\theta}) \\
&\quad = \log \prod_{s\in L, y\in\mathcal{X}} (\eta_{y|s})^{(n_s+(1+d)/2)\hat{\eta}^L_{y|s}} \tilde{w}(\boldsymbol{\theta})
\end{aligned}$$

where we have defined Laplace estimator as

$$\hat{\eta}^L_{y|s} \stackrel{\text{def}}{=} \frac{n_{y|s} + 1/2}{n_s + (d+1)}.$$

This implies

$$G(\boldsymbol{n}, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}))w(\boldsymbol{\theta}) = G(\boldsymbol{n} + (d+1)/2 \cdot \mathbf{1}, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}^L)\tilde{w}(\boldsymbol{\theta}). \quad (56)$$

Hence, we have

$$\begin{aligned}
&\int \eta_{x|s} W^{(w)}(\boldsymbol{n}, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) d\boldsymbol{\theta} \\
&\quad = \int \eta_{x|s} W^{(\tilde{w})}(\boldsymbol{n} + (d+1)/2 \cdot \mathbf{1}, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}^L) d\boldsymbol{\theta}.
\end{aligned} \quad (57)$$

By assumption, $\hat{\boldsymbol{\eta}}^L \in K$ holds for all large n. Hence, by Lemma 2 and (57), we have for all $s \in L$ and for all $x \in \mathcal{X}'$

$$\begin{aligned}
&\int \eta_{x|s} w(\boldsymbol{\theta}|x^n) d\boldsymbol{\theta} \\
&\quad = \int \eta_{x|s} W^{(\tilde{w})}(\boldsymbol{n} + (d+1)/2 \cdot \mathbf{1}, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}^L) d\boldsymbol{\theta} \\
&\quad = \hat{\eta}^L_{x|s} + \frac{1}{n_s + \frac{d+1}{2}} \frac{\partial \log \tilde{w}(\boldsymbol{\theta})}{\partial \theta_{x|s}}\bigg|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^L} + O\Big(\frac{\sqrt{\log n}}{n\sqrt{n}}\Big).
\end{aligned}$$

This completes the proof of Theorem 3.

VI. Refinement for the Two-State First-Order Markov Case

As we have seen, the Jeffreys prior differs from a product of Dirichlet $(1/2, \ldots, 1/2)$ priors by the factor

$$\prod_{s\in L} \mu_s^{d/2}$$

where μ_s is the stationary probability of the state s associated with $p(\cdot|\boldsymbol{\eta})$. In the two-state first-order Markov chain case, these stationary probabilities are

$$(\mu_0, \mu_s) = \Big(\frac{a}{a+b}, \frac{b}{a+b}\Big)$$

where $a = \eta_{0|1}$ and $b = \eta_{1|0}$, which yields the Jeffreys factor

$$\frac{\sqrt{ab}}{a+b}.$$

Accordingly, the Jeffreys mixture probability $m_J(x^n) = m_J(x^n|x_0)$ takes the form

$$\frac{1}{C_J} \int_{[0,1]^2} \frac{1}{a+b} a^{n_{0|1}} \bar{a}^{n_{1|1}-0.5} b^{n_{1|0}} \bar{b}^{n_{1|1}-0.5} da db$$

where $\bar{a} = 1 - a = \eta_{1|1}$ and $\bar{b} = 1 - b = \eta_{0|0}$. The factor $1/(a+b)$ prevents the integral from decoupling as a product of integrals for a and for b.

A. Refined Approximation to $m_J(x^n)$

The following lemma obtains tight upper and lower bounds on $m_J(x^n)$ for this two-state first-order Markov case. The idea of the lemma is to obtain approximate decoupling of a and b in the integral.

Lemma 3: For (a,b) and $(\hat{a},\hat{b})$ in $(0,1)^2$

$$\frac{1}{\hat{a}+\hat{b}}\Big(1-\frac{a-\hat{a}}{\hat{a}+\hat{b}}-\frac{b-\hat{b}}{\hat{a}+\hat{b}}\Big) \le \frac{1}{a+b} \le \frac{1}{\hat{a}+\hat{b}}\Big(\frac{\hat{a}}{a}\Big)^{\hat{\mu}_0}\Big(\frac{\hat{b}}{b}\Big)^{\hat{\mu}_1} \tag{58}$$

where $\hat{\mu}_0=\hat{a}/(\hat{a}+\hat{b})$ and $\hat{\mu}_1=\hat{b}/(\hat{a}+\hat{b})$.

Consequently, we have the upper bound on $m_J(x^n)$ of

$$\frac{\hat{a}^{\hat{\mu}_0}\hat{b}^{\hat{\mu}_1}}{(\hat{a}+\hat{b})C_J}\int a^{n_{0|1}-\hat{\mu}_0}\bar{a}^{n_{1|1}-0.5}b^{n_{1|0}-\hat{\mu}_1}\bar{b}^{n_{0|0}-0.5}dadb = \frac{\hat{a}^{\hat{\mu}_0}\hat{b}^{\hat{\mu}_1}}{(\hat{a}+\hat{b})C_J}B(n_{0|1}+1-\hat{\mu}_0,n_{1|1}+0.5)\cdot B(n_{1|0}+1-\hat{\mu}_1,n_{0|0}+0.5)$$

where $B(m_1,m_2)=\Gamma(m_1)\Gamma(m_2)/\Gamma(m_1+m_2)$ is the Beta function. This upper bound is valid for any $(\hat{a},\hat{b})$ in $(0,1)^2$. Moreover, we have the lower bound on $m_J(x^n)$ of

$$\frac{1}{(\hat{a}+\hat{b})C_J}\cdot\int\Big(1-\frac{a-\hat{a}}{\hat{a}+\hat{b}}-\frac{b-\hat{b}}{\hat{a}+\hat{b}}\Big)a^{n_{0|1}}\bar{a}^{n_{1|1}-0.5}b^{n_{1|0}}\bar{b}^{n_{0|0}-0.5}dadb. \tag{59}$$

With the choice $\hat{a}=(n_{0|1}+1)/(n_1+1.5)$ and $\hat{b}=(n_{1|0}+1)/(n_0+1.5)$, the $a-\hat{a}$ and $b-\hat{b}$ contributions to the integral vanish, yielding the lower bound on $m_J(x^n)$ of

$$\frac{1}{(\hat{a}+\hat{b})C_J}B(n_{0|1}+1,n_{1|1}+0.5)B(n_{1|0}+1,n_{0|0}+0.5).$$

These upper and lower bounds hold for all nonnegative counts $n_{0|1},n_{1|1},n_{0|0},n_{1|0}$ and the ratio of the upper and lower bounds tends to 1 when these four counts get large.

Proof of Lemma 3: The function $1/(a+b)$ is convex on $\Re^2$ and so it is greater than or equal to the left-hand side of (58) which is its first-order Taylor expansion, tangent to the function at $(\hat{a},\hat{b})$. Likewise, interpret $1/(a+b)=e^{g(\alpha,\beta)}$ with $g(\alpha,\beta)=-\log(e^{\alpha}+e^{\beta})$ and $a=e^{\alpha}$, $b=e^{\beta}$. The function $g(\alpha,\beta)$ is concave on $\Re^2$ and so it is less than or equal to its first-order Taylor expansion, tangent to it at $(\hat{\alpha},\hat{\beta})$, which yields the right-hand side of (58).

From Sterling's formula, the Gamma function has the property that the ratio $R_n(x)=\Gamma(n+1+x)/(n^x\Gamma(n+1))$ converges to 1 for each x as $n\to\infty$ (see, e.g., [10, p. 80] or [9, p. 886]). The ratio of the upper bound of $m_J(x^n)$ to the lower bound at the chosen $\hat{a}$, $\hat{b}$ is seen to equal

$$\frac{R_{n_{0|1}}(-\hat{\mu}_0)R_{n_{1|0}}(-\hat{\mu}_1)}{R_{n_1}(-\hat{\mu}_0)R_{n_0}(-\hat{\mu}_1)}$$

which accordingly approaches 1 as $n_{0|1}$, $n_{1|0}$, n_1, n_0 get large. This completes the proof of Lemma 3.

Remark: The variance of a $\mathrm{Beta}(m_0,m_1)$ distribution is $m_0m_1/((m_0+m_1+1)(m_0+m_1)^2)$ near $m_0m_1/(m_0+m_1)^3$ which is typically of order $1/(m_0+m_1)$. However, if either m_0 or m_1 stays bounded and the sum m_0+m_1 gets large, then the variance is of the smaller order $1/(m_0+m_1)^2$.

In the integral (59), the remainder of the Taylor expansion $1-(a-\hat{a})/(\hat{a}+\hat{b})-(b-\hat{b})/(\hat{a}+\hat{b})$ is of order $(a-\hat{a})^2/(\hat{a}+\hat{b})^2+(b-\hat{b})^2/(\hat{a}+\hat{b})^2$.

Neglecting effects from a and b far from $\hat{a}$, $\hat{b}$ which do not contribute substantially unless n_0 and n_1 are large, it reveals that $m_J(x^n)$ matches its lower bound approximation to within a factor of order

$$1+O\Big(\frac{1}{n_0}+\frac{1}{n_1}\Big)\frac{1}{(\hat{a}+\hat{b})^2}.$$

B. Improved Monte Carlo Calculation of Predictive Probabilities

The Jeffreys predictive probabilities $m_J(x_{n+1}=0|x^n)$ in the $s=x_n=1$ case arise as the ratio of integrals. As we have seen, the numerator integral is

$$\int_{[0,1]^2}(a)\frac{1}{a+b}a^{n_{0|1}}\bar{a}^{n_{1|1}-0.5}b^{n_{1|0}}\bar{b}^{n_{0|0}-0.5}dadb$$

and the denominator integral is the same but without the factor (a). If we multiply and divide in the integral by the expression $(\hat{a}+\hat{b})(a/\hat{a})^{\hat{\mu}_0}(b/\hat{b})^{\hat{\mu}_1}$, then these integrals can be expressed via expectation forms of appropriate Beta densities. For the numerator, we use

$$\mathrm{Num}_n=\int\Big[\frac{\hat{a}+\hat{b}}{a+b}\Big(\frac{a}{\hat{a}}\Big)^{\hat{\mu}_0}\Big(\frac{b}{\hat{b}}\Big)^{\hat{\mu}_1}\Big]\cdot\mathrm{B}_{n_{0|1}+2-\hat{\mu}_0,n_{1|1}+0.5}(a)\cdot\mathrm{B}_{n_{1|0}+1-\hat{\mu}_1,n_{0|0}+0.5}(b)dadb \tag{60}$$

where B_{m_0,m_1} denotes a $\mathrm{Beta}(m_0,m_1)$ probability density function. For the denominator Den_n, we use the same expression but with the 2 replaced by 1. Here, we have incorporated the normalizing constants of these Beta densities. Accordingly, when we compute the predictive probabilities, we compensate for the ratio of the normalizing constants which is $\mathrm{B}(n_{0|1}+2-\hat{\mu}_0,n_{1|1}+0.5)/\mathrm{B}(n_{0|1}+1-\hat{\mu}_0,n_{1|1}+0.5)$ equal to $(n_{0|1}+1-\hat{\mu}_0)/(n_1+1.5-\hat{\mu}_0)$. Consequently, with $x_n=s=1$, the Jeffreys predictive probability is

$$m_J(x_{n+1}=0|x^n)=\frac{n_{0|1}+\hat{\mu}_1}{n_1+0.5+\hat{\mu}_1}\frac{\mathrm{Num}_n}{\mathrm{Den}_n}$$

where we have used $1-\hat{\mu}_0=\hat{\mu}_1$.

To interpret this expression, the $(n_{0|1}+\hat{\mu}_1)/(n_1+0.5+\hat{\mu}_1)$ is the approximation to the predictive probability, which is asymptotically equivalent to the approximation formula (12) given by Lemma 3. It is accurate when the counts are very large, and then

TABLE II
REGRET (BOUNDARY CASES)

$\hat{\eta}_{0\|1}$	$\hat{\eta}_{1\|0}$	$\tilde{r}_{low}$	$\tilde{r}_{up}$	$\tilde{r}(m_J)$	$\tilde{r}(q)$
0.00001	0.00001	1.2996	1.3133	1.292	-0.068
0.0001	0.0001	1.2986	1.3001	1.300	1.059
0.0010	0.0010	1.2985	1.2987	1.298	1.337
0.0099	0.0100	1.2985	1.2986	1.299	1.374
0.0001	1.0000	1.6448	1.6449	1.657	-3.510
0.0010	0.9990	1.3054	1.3055	1.307	0.763
0.0100	0.9900	1.2985	1.2986	1.297	1.353
0.0001	0.5175	1.2988	1.2994	1.297	1.092
0.0010	0.5037	1.2985	1.2987	1.303	1.349
0.0100	0.5005	1.2985	1.2986	1.300	1.375
0.9900	0.9900	1.2985	1.2986	1.302	1.377
0.9990	0.9990	1.2985	1.2986	1.299	1.336
0.9999	0.9999	1.2986	1.2987	1.299	1.058
0.9901	0.4999	1.2985	1.2986	1.298	1.376
0.9990	0.5000	1.2985	1.2986	1.299	1.375
0.9999	0.4999	1.2986	1.2987	1.300	1.369
0.99999	0.4997	1.3000	1.3001	1.304	1.353

Num_n and Den_n are near 1. When the counts are small or to improve the precision when the counts are moderate, evaluation of Num_n and Den_n is appropriate.

We suggest Monte Carlo evaluation in which the exact integrals Num_n and Den_n are replaced by sample averages of the quantity in brackets (see the first line of (60) for Num_n) using independent draws from the respective Beta distributions. The expression $[(\hat{a}+\hat{b})\,(a+b)^{-1}\,(a/\hat{a})^{\hat{\mu}_0}(b/\hat{b})^{\hat{\mu}_1}]$ in brackets is always less than or equal to 1 and it is near to 1 when the Beta distribution has sufficient counts to make the distribution peaked near $\hat{a}$ and $\hat{b}$. This expression $[(\hat{a}+\hat{b})\,(a+b)^{-1}\,(a/\hat{a})^{\hat{\mu}_0}(b/\hat{b})^{\hat{\mu}_1}]$ arises as the exponential of the remainder of a first-order Taylor expansion used in the proof of Lemma 3, so its drop from 1 is of the order $(a-\hat{a})^2+(b-\hat{b})^2$. The crux is it has considerably reduced variance compared to the previously suggested Monte Carlo. As a result, one does not use as large a Monte Carlo sample size to produce accurate computations.

Table II here shows computation results for the regrets using $m_J(x^n)$ and $q(x^n)$ including cases with sequences with very small numbers of transitions. We report values of $\tilde{r}(m)=\tilde{r}(m,x^n)$ given, as earlier, by

$$\log\frac{1}{m(x^n|x_0)}-\log\frac{1}{p(x^n|x_0,\hat{\boldsymbol{\eta}})}-\log\frac{n}{2\pi}$$

using either the Jeffreys rule or its modification q. The column heading $\tilde{r}_{\mathrm{low}}$ refers to lower bounds on regret of the procedural obtained from the upper bound on $m_J(x^n)$ in Lemma 3; the heading $\tilde{r}_{up}$ refers to upper bounds on regret obtained from the lower bound on $m_J(x^n)$.

The total sample size as earlier is $n=10^7$. Each Monte Carlo calculation is performed by the improved-precision version developed here. The objective is to render these digit accuracy on these regrets. For initial sequence of length less than 100, the Beta distributions are not so peaked and we used Monte Carlo size of 100 000.

Once all four counts $n_{0|1}$, $n_{1|0}$, $n_{1|1}$, and $n_{0|0}$ reach at least 100, we switch to the approximation formulas (12). For moderates size counts (not all at least 100) the Monte Carlo refinement to the A.F. with Monte Carlo size of 10 000.

This scheme allowed sensible precision of computation over a broader range of cases than before.

VII. CONCLUDING REMARK

We have shown that the modified Jeffreys mixtures asymptotically achieve the minimax regret for Markov models without any restriction on the sequences. The obtained regret is of the same form as that for the multinomial Bernoulli models. Then, we consider the computational aspects of the minimax strategies, and we have obtained an approximation formula of Jeffreys mixture for Markov models.

APPENDIX A
JEFFREYS POSTERIOR UPDATING

Here, we derive (10) and explain the Jeffreys posterior and its relationship to the Dirichlet posterior. Note that the Jeffreys posterior given x^n is proportional to

$$p(x^n|\boldsymbol{\eta})w_J(\boldsymbol{\eta})\propto p(x^n|\boldsymbol{\eta})\prod_{s\in L}\mu_s^{d/2}D_{(1/2)}(\boldsymbol{\eta}_s).$$

Since $p(x^n|\boldsymbol{\eta})=\prod_s\prod_x(\eta_{x|s})^{n_{x|s}}$, it is proportional to

$$\Big(\prod_{s\in L}\mu_s^{d/2}\Big)\prod_{s\in L}\prod_x(\eta_{x|s})^{n_{x|s}+1/2}$$

where $\boldsymbol{n}=(\boldsymbol{n}_s)_{s\in L}$ is a collection of counts from x^n. Since the posterior for the Dirichlet $(1/2,\ldots,1/2)$ prior, denoted by $\bar{D}_{(\mathbf{1}/2+\boldsymbol{n})}(\boldsymbol{\eta})$, is proportional to $\prod_{s\in L}\prod_x(\eta_{x|s})^{n_{x|s}+1/2}$, we have

$$w_J(\boldsymbol{\eta}|x^n)\propto\Big(\prod_{s\in L}\mu_s^{d/2}\Big)\bar{D}_{(\mathbf{1}/2+\boldsymbol{n})}(\boldsymbol{\eta}).$$

APPENDIX B
EXPRESSION OF STATIONARY PROBABILITIES OF A MARKOV MODEL

Here, we will prove Lemma 4, which gives an explicit formula of the stationary probabilities for Markov chains and describe a certain property of it.

For its proof, we utilize the following theorem given by Chaiken and Kleitman [5].

Theorem 4 (Matrix Tree Theorem): Let $M(\{x_q\})$ denote a squared matrix of order γ, whose entries are

$$M(\{x_q\})_{ij}=\begin{cases}\sum_{k\neq i}M_{ik}x_k, & i=j,1\le j\le\gamma\\ -M_{ij}x_j, & i\neq j,1\le i,j\le\gamma.\end{cases}$$

Let $f(j_1,\ldots,j_k)$ $(k\le n)$ be the determinant of the matrix obtained by omitting the j_ith row and column of $M(\{x_q\})$ for all $i:1\le i\le k$. Let S be the set of all arborescences on vertexes $v_1,v_2,\ldots,v_\gamma$ rooted at $v_{j_1},\ldots,v_{j_k}$. For each a in S, let w_a be the product of $M_{ij}x_j$ over all directed arcs $(j\to i)$ in a. Then, the identity $f(j_1,\ldots,j_k)=\sum_{a\in S}w_a$ holds.

See [5] for the proof. Here, an *arborescence* is a graph in which every vertex other than roots has in-degree one, there are no cycles, and the roots have in-degree zero. The matrix tree theorem is well known in circuit theory and graph theory and

several variations exist (see, e.g., [4], [12], and [14]). Theorem 4 is a fairly general one.

We have the following.

Lemma 4: Let A be a state transition matrix of a first-order Markov chain with alphabet $\{1,2,\ldots,\gamma\}$, i.e., A_{ij} is a conditional probability of i's generation after j's. Let μ_i be the stationary probability of the symbol i defined by the Markov chain. Let $\epsilon \stackrel{\rm def}{=} \min_{i,j} A_{ij}$, and let Δ_{ij} denote the (i,j)th cofactor of the matrix $I-A$. Then, we have the following.

1) For each j, $\Delta_{1j}=\Delta_{2j}=\cdots=\Delta_{\gamma j}$ holds.
2) Each Δ_{ij} is a sum of products of $\gamma-1$ certain components of A, in particular, not less than $\epsilon^{\gamma-1}$.
3) When $\epsilon>0$, the following equalities hold:

$$\mu_i=\frac{\Delta_{ii}}{\sum_{l=1}^{\gamma}\Delta_{ll}} \quad (i=1,2,\ldots,\gamma).$$

Proof: Let $B_{ij}=(I-A)_{ij}$. Since $\sum_{i=1}^{\gamma}A_{ij}=1$, we have $\sum_{i=1}^{\gamma}B_{ij}=0$ $(j=1,2,\cdots,\gamma)$. Hence, adding the ith line of B to the first line for $i=3,4,\ldots,\gamma$, the first line of the resultant matrix is equal to minus the second line of B. This implies $\Delta_{2j}=\Delta_{1j}$ $(j=1,2,\ldots,\gamma)$. Since this argument holds for any pair of lines by symmetry, we have item 1.

In order to show item 2, we use Theorem 4, assuming $x_j=1$ for $j=1,\ldots,\gamma$. Then, B_{ij} satisfies the property of $M(\{x_q\})$ in Theorem 4, where we have $M_{ij}=A_{ij}$ $(i\neq j)$ and $\Delta_{ii}=f(i)$. Hence, the following holds:

$$\Delta_{ii}=f(i)=\sum_{a\in S}w_a.$$

This implies that Δ_{ii} is a sum of products of $\gamma-1$ certain nondiagonal elements of A. Hence, $\Delta_{ii}\geq\epsilon^{\gamma-1}$. By item 1, this holds for every Δ_{ij}.

Now, we will show item 3. Let $\mathrm{Adj}B$ denote the matrix with (i,j) entries are Δ_{ji}. Then, we have $(I-A)\mathrm{Adj}B=B\mathrm{Adj}B=(\det B)I=0$. This implies that the vectors $(\Delta_{j1},\Delta_{j2},\ldots\Delta_{j\gamma},)^t$ $(j=1,\ldots,\gamma)$ are the eigenvector of A with eigenvalue 1. Here, note that $\Delta_{ij}>0$ when $\epsilon>0$. Then, we have obtained

$$\mu_i=\frac{\Delta_{ji}}{\sum_{l=1}^{\gamma}\Delta_{jl}}=\frac{\Delta_{ii}}{\sum_{l=1}^{\gamma}\Delta_{ll}}.$$

This completes the proof of Lemma 4.

Appendix C
Lemma for Jeffreys Prior

Lemma 5: There exists a certain integer $m_{\min}$, such that for all $\boldsymbol{\eta}'\in K$, for all $x^k\in\mathcal{X}^k$, for all $x\in\mathcal{X}'$, and for all $s\in L$

$$\frac{\partial\log w_J(\boldsymbol{\theta})}{\partial\theta_{x|s}}G(m_{min}\cdot\mathbf{1},\boldsymbol{\theta},\boldsymbol{\eta}')\cdot w_J(\boldsymbol{\theta})$$

is integrable over Θ.

Proof: Recall that

$$w_J(\boldsymbol{\theta})=\frac{1}{C_J}\prod_{s\in L}\mu_s^{d/2}D_{(1/2)}(\boldsymbol{\eta}_s).$$

We have

$$\log w_J(\boldsymbol{\theta})=\frac{d}{2}\sum_{s\in L}(\log\mu_s+\log D_{(1/2)}(\boldsymbol{\eta}_s))-\log C_J.$$

Therefore, recalling (16), it is sufficient to show that the following two are integrable for all $\boldsymbol{\eta}'\in K$ and for all $y\in\mathcal{X}'$

$$\frac{\partial\log D_{(1/2)}(\boldsymbol{\eta}_s)}{\partial\eta_{y|s}}G(m_{\min}\cdot\mathbf{1},\boldsymbol{\theta},\boldsymbol{\eta}')w_J(\boldsymbol{\theta}) \tag{61}$$

$$\frac{\partial\log\mu_s}{\partial\eta_{y|t}}G(m_{\min}\cdot\mathbf{1},\boldsymbol{\theta},\boldsymbol{\eta}')w_J(\boldsymbol{\theta}). \tag{62}$$

Now let κ be $\min_{\boldsymbol{\eta}'\in K,}\min_{s\in L}\min_{y\in\mathcal{X}}{\eta'}_{y|s}$. As for (48), note that

$$\frac{\partial\log D_{(1/2)}(\boldsymbol{\eta}_s)}{\partial\eta_{y|s}}=-\frac{1}{2}\frac{\partial\sum_{x\in\mathcal{X}}\log\eta_{x|s}}{\partial\eta_{y|s}}=-\frac{1}{2}\Big(\frac{1}{\eta_{y|s}}-\frac{1}{\eta_{0|s}}\Big).$$

Recall that $G(m_{\min}\cdot\mathbf{1},\boldsymbol{\theta},\boldsymbol{\eta}')=\prod_{s\in L,y\in\mathcal{X}}(\eta_{y|s})^{m_{\min}{\eta'}_{y|s}}$. Hence if $m_{\min}\geq 1/\kappa$, we have $G(m_{\min}\mathbf{1},\boldsymbol{\theta},\boldsymbol{\eta}')\leq\prod_{s\in L,x\in\mathcal{X}}\eta_{x|s}$ for all $\boldsymbol{\theta}\in\Theta$. Hence, for all $\theta\in\Theta$, we have

$$\Big|\frac{\partial\log D_{(1/2)}(\boldsymbol{\eta}_s)}{\partial\eta_{y|s}}G(m_{\min}\cdot\mathbf{1},\theta,\eta')\Big|\leq 1.$$

Hence, when $m_{\min}\geq 1/\kappa$, (48) is integrable.

Now, we examine (62). Let $\eta_{\min}\stackrel{\rm def}{=}\min_{s\in L,y\in\mathcal{X}}\eta_{y|s}$; then, by Proposition 3, we have

$$\Big|\frac{\partial\log\mu_s}{\partial\eta_{y|t}}\Big|\leq\frac{C_1}{{\eta_{\min}}^r}$$

where $r=k(\ell-1)$. Hence, if $m_{\min}\geq r/\kappa$, we have

$$\Big|\frac{\partial\log\mu_s}{\partial\eta_{y|t}}\Big|G(m_{\min}\cdot\mathbf{1},\boldsymbol{\theta},\boldsymbol{\eta}')\leq\frac{C_1{\eta_{\min}}^r}{{\eta_{\min}}^r}\leq C_1.$$

Hence when $m_{\min}\geq r/\kappa$, (62) is integrable. This completes the proof of Lemma 5.

Appendix D
Theoretical Value of the Minimax Regret for the Simplest Case

For the simplest case (Example 1), since $\mu_0=\eta_{0|1}/(\eta_{0|1}+\eta_{0|1})$ and $\mu_1=\eta_{1|0}/(\eta_{0|1}+\eta_{0|1})$, we have

$$\begin{aligned}C_J&=\int_H\sqrt{\det(I(\boldsymbol{\eta}))}d\boldsymbol{\eta}\\&=\int_H\prod_{s\in\{0,1\}}\Big(\frac{\sqrt{\mu_s}}{\sqrt{\eta_{0|s}(1-\eta_{0|s})}}\Big)d\boldsymbol{\eta}\\&=\int_{[0,1]^2}\frac{d\eta_{1|0}d\eta_{0|1}}{(\eta_{0|1}+\eta_{1|0})\sqrt{(1-\eta_{0|1})(1-\eta_{0|1})}}.\end{aligned}$$

The last expression equals four times the Catalan constant (see, e.g., [2]), which equals $4 \cdot 0.915965594 \cdots \approx 3.66386237$. (See, e.g., [9, p. 1036].) Hence, we have $\log C_J \approx 1.2985$.

ACKNOWLEDGMENT

The authors express their sincere gratitude to anonymous referees, in particular for teaching us about the matrix tree theorem, and to Mariko Tsurusaki for helping them to perform the numerical experiments.

REFERENCES

[1] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2104–2109, Sep. 1999.
[2] D. M. Bradley, Representations of Catalan's constant 1998 [Online]. Available: http://germain.umemat.maine.edu/faculty/bradley/papers/c1.ps
[3] L. D. Brown, *Fundamentals of Statistical Exponential Families*. Hayward, CA: Inst. Math. Statist., 1986.
[4] A. Cayley, "A theorem on trees," *Quart. J. Math.*, vol. 23, pp. 376–378, 1889.
[5] S. Chaiken and D. J. Kleitman, "Matrix tree theorems," *J. Combin. Theory, Series A*, vol. 24, pp. 377–381, May 1978.
[6] B. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.
[7] B. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Plann. Infer.*, vol. 41, pp. 37–60, 1994.
[8] M. Gotoh, T. Matsushima, and S. Hirasawa, "A generalization of B. S. Clarke and A. R. Barron's asymptotics of Bayes codes for FSMX sources," *IEICE Trans. Fund.*, vol. E81-A, no. 10, pp. 2123–2132, 1998.
[9] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. New York: Academic, 2000.
[10] F. B. Hildebrand, *Advanced Calculus for Applications*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1976.
[11] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for Markov sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, Jul. 2004.
[12] G. Kirchhoff, "Über die auflosung der gleichungen auf welche man bei der untersuchung der linearen verteilung galvanisher ströme gefuhrt wird," *Ann. Phys. Chem.*, vol. 72, pp. 497–508, 1847.
[13] T. Kawabata and F. Willems, "A context tree weighting algorithm with an incremental context set," *IEICE Trans. Fund.*, vol. E83-A, no. 10, pp. 1898–1903, 2000.
[14] J. C. Maxwell, *A Treatise on Electricity and Magnetism I*, 3rd ed. London, U.K.: Oxford Univ. Press, 1892, ch. 6.
[15] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1993.
[16] J. Rissanen, "A universal data compression system," *IEEE Trans. Inf. Theory*, vol. 29, no. 5, pp. 656–664, Sep. 1983.
[17] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40–47, Jan. 1996.
[18] Y. M. Shtar'kov, "Universal sequential coding of single messages," *Probl. Inf. Transmiss.*, vol. 23, no. 3, pp. 3–17, Jul. 1987.
[19] J. Takeuchi, "Characterization of the Bayes estimator and the MDL estimator for exponential families," *IEEE Trans. Inf. Theory*, vol. 43, no. 4, pp. 1165–1174, Jul. 1997.
[20] J. Takeuchi, "Fisher information determinant and stochastic complexity for Markov models," in *Proc. IEEE. Int. Symp. Inf. Theory*, Seoul, Korea, Jun. 2009, pp. 1894–1898.
[21] J. Takeuchi and A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," in *Proc. IEEE. Int. Symp. Inf. Theory*, Boston, MA, Aug. 1998, p. 318.
[22] J. Takeuchi and T. Kawabata, "Approximation of Bayes code for Markov sources," in *Proc. IEEE. Int. Symp. Inf. Theory*, Whistler, BC, Canada, Sep. 1995, p. 391.
[23] J. Takeuchi, T. Kawabata, and A. R. Barron, "Properties of Jeffreys mixture for Markov sources," in *Proc. 4th Workshop Inf. Based Induction Sciences (it IBIS2001)*, Tokyo, Japan, Jul. 2001, pp. 327–332.
[24] M. J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 643–652, May 1995.
[25] F. Willems, Y. Shtar'kov, and T. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.
[26] Q. Xie, *Minimax coding and prediction*, Doctoral Dissertation, Dept. of Statistics, Yale University, 1997.
[27] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 646–657, Mar. 1997.
[28] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 431–445, Mar. 2000.
[29] H. Itoh and S. Amari, "Geometry of information sources (in Japanese)," in *Proc. 11th Symp. Inf. Theory Appl.*, Ooita, Japan, Dec. 1988, pp. 57–60.
[30] T. Kawabata, "Bayes codes and context tree weighting method (in Japanese)," *Tech. Rep. IEICE,* IT93-121. Mar. 1994, pp. 7–12.
[31] J. Takeuchi, "On minimax regret with respect to families of stationary stochastic processes (in Japanese)," in *Proc. 3rd Workshop Inf. Based Induct. Sci.*, Shizuoka, Japan, Jul. 2000, pp. 63–68.
[32] J. Takeuchi and T. Kawabata, "On data compression algorithms by Bayes coding for Markov sources (in Japanese)," in *Proc. 17th Symp. Inf. Theory Appl.*, Hiroshima, Japan, Dec. 1994, pp. 513–516.

Jun'ichi Takeuchi (M'05) was born in Tokyo, Japan in 1964. He graduated from the University of Tokyo in majoring physics in 1989. He received the Dr. Eng. degree in mathematical engineering from the University of Tokyo in 1996. From 1989 to 2006, he worked for NEC Corporation, Japan. In 2006, he joined Kyushu University, Fukuoka, Japan, where he is a Professor of Mathematical Engineering. From 1996 to 1997 he was a Visiting Research Scholar at Department of Statistics, Yale University, New Haven, CT, USA. His research interest includes mathematical statistics, information geometry, information theory, and machine learning. He is a member of IEEE, IEICE, IPSJ, and JSIAM.

Tsutomu Kawabata (M'93) was born in Toyama, Japan, in 1955. He received BE, ME, and DE degrees in mathematical engineering from the University of Tokyo, in 1978, 1980, and 1993 respectively. He joined the University of Electro-Communications in 1982 and is currently a Professor at the Department of Communication Engineering and Informatics. He was a visitor at Stanford University during 1987–89 and 1996–97, and at Eindhoven University of Technology in 1995, and at INRIA in 1996. His research interests lie in information and communication theory, and include quantizations, rate-distortions, and lossless data compressions.

Andrew R. Barron (S'84–M'85–SM'00–F'12) was born in Trenton, NJ, on September 28, 1959. He received the B.S. degree in electrical engineering and mathematical sciences from Rice University, Houston, TX, in 1981, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1982 and 1985, respectively.

From 1977 to 1982, he was a consultant and summer employee of Adaptronics, Inc., McLean, VA. From 1985 until 1992, he was a faculty member of the University of Illinois at Urbana-Champaign in the Department of Statistics and the Department of Electrical and Computer Engineering. He was a Visiting Research Scholar at the Berkeley Mathematical Sciences Research Institute in the Fall of 1991 and Barron Associates, Inc., Standardsville, VA, in the Spring of 1992.

In 1992, he joined Yale University, New Haven, CT, as a Professor of Statistics, where he has served as Chair of Statistics from 1999–2006. His research interests include the study of information-theoretic properties in the topics of probability limit theory, statistical inference, high-dimensional function estimation, neural networks, model selection, communication, universal data compression, prediction, and investment theory.

Dr. Barron received (jointly with Bertrand S. Clarke) the 1991 Browder J. Thompson Prize (best paper in all IEEE TRANSACTIONS in 1990 by authors age 30 or under) for the paper "Information-Theoretic Asymptotics of Bayes Methods." Dr. Barron was an Institute of Mathematical Statistics Medallion Award recipient in 2005. He served on the Board of Governors of the IEEE Information Theory Society from 1995 to 1999, and was Secretary of the Board of Governors during 1989–1990. He has served as an Associate Editor for the IEEE TRANSACTION ON INFORMATION THEORY from 1993 to 1995, and the Annals of Statistics for 1995–1997.