Iterated Logarithmic Expansions of the Pathwise Code Lengths for Exponential Families

Lei Li

Florida State University

Bin Yu

Bell laboratories, Lucent Technologies and University of California at Berkeley

January 14, 2000

Abstract

Rissanen's Minimum Description Length (MDL) principle is a statistical modeling principle motivated by coding theory. For exponential families we obtain pathwise expansions, to the constant order, of the predictive and mixture code lengths used in MDL. The results are useful for understanding different MDL forms.

Key words: minimum description length, mixture code, predictive code, two-stage code, law of iterated logarithm

1 Introduction and background

The Minimum Description Length (MDL) principle is introduced by Rissanen as a fundamental principle to model data, see [15, 17] and the reference list in [18]. If we encode data from a source by prefix codes, the best code is the one that achieves the minimum description length among all prefix codes if there is such a one. Because of the equivalence between a prefix codelength and the negative logarithm of the corresponding probability distribution (via Kraft's inequality), this in turn gives us a modeling principle, namely, the MDL principle: choose the model which gives the minimal description of data.

When the source distribution is known, Shannon's coding theorem dictates that we use the source distribution for the optimal code, and its codelength is the negative logarithm of the probability of the data. When the source distribution is unknown, the search for a universal modeling procedure is the search for a universal coding scheme. This universal coding scheme provides us with a distribution, the negative logarithm of whose frequency function or density function is the corresponding codelength of the (ideal) optimal code (ignoring the precision issue). Many favorable properties have been shown for MDL-based statistical procedures in both parametric and nonparametric contexts (see the two recent reviews by Barron, Rissanen and Yu [1] and Hansen and Yu [13]). A crucial step in implementing a MDL procedure is to choose the form of the universal code. Three forms have been studied extensively: two-stage, predictive and mixture, although new forms are appearing such as Normalized Maximum Likelihood (NML) (Rissanen, [19]). In order to compare these forms, here we provide pathwise expansions of codelengths generated from exponential families. Finite mixtures of the corresponding conjugate distributions are used as priors for mixture codes. A related work on the mixture code is Takeuchi and Barron [20] where they get an almost sure expansion for exponential families and Jeffreys' priors (when they exist).

Given a parametric family $\{P_\theta : \theta \in \text{ an open set } \Theta \subset \mathbf{R}^d\}$, we denote the density functions by $p(x|\theta)$. Assume $X_1, \cdots, X_n$ are independent and identically distributed (iid) random variables from P_{θ_0} and θ_0 is an interior point of Θ. Then the codelength of the optimal code is given by Shannon's coding theorem as $L_0 = -\sum_{i=1}^n \log p(X_i|\theta_0)$. The codelength corresponding to a given distribution $Q(x)$ is $L_Q = -\sum_{i=1}^n \log q(X_i)$, and the redundancy is $R_Q = L_Q - L_0$. Rissanen [17] shows that for each positive number ϵ and for all $\theta_0 \in \Theta$, except in a set whose volume goes to zero as $n \longrightarrow \infty$,

$$E_{P_{\theta_0}} R_Q \geq \frac{d-\epsilon}{2} \log\, n\,. \tag{1}$$

Later Barron and Hengartner ([2]) prove that except for a set of parameter values with a volume zero,

$$\overline{\lim}_{n\to\infty}\, E_{P_{\theta_0}} \frac{R_Q}{\frac{d}{2}\log\, n} \geq 1\,. \tag{2}$$

Several coding schemes have been shown to achieve this bound. Among them are two-stage, mixture and predictive codes. In particular, under regularity conditions Clarke and Barron [5, 1] prove that the expected redundancy of the mixture code is given by

$$E_{P_{\theta_0}} R_Q = \frac{d}{2} \log\, n - \frac{d}{2} \log\, 2\pi\, e + \log \frac{(det\, I(\theta_0))^{1/2}}{w(\theta_0)} + o(1)\,, \tag{3}$$

where $I(\theta)$ is the Fisher information matrix, and $w(\theta)$ is the density of the prior. For the predictive code, the leading term $\frac{d}{2}\log\, n$ of the average regret has also been obtained, see [16, 18].

In this paper, we carry out a careful study of the pathwise codelengths of the predictive and mixture codes when the source distribution belongs to an exponential family and the prior is taken to be a finite mixture of its conjugate distribution. It is shown that the leading term of the pathwise redundancies of these two codes is $\frac{d}{2}\log n$, and the second term is of the order of iterated-logarithm with the same nonpositive bounded random coefficient. The third term is a constant for the mixture code, and a finite random variable for the predictive code which may not depend on the parameter of the underlying distribution for some sources. A similar expansion for the two-stage codelength can be derived.

The rest of the paper is organized as follows. Section 2 contains the main results and Section 3 consists of a discussion. All the proofs can be found in Section 4 (Appendix).

2 Main results

Consider a canonical exponential family of distributions $\{P_\theta : \theta \in \Theta\}$, where the natural parameter space Θ is an open set of $\mathbf{R}^d$. The density function is given by

$$p(x|\theta) = \exp\{\theta' T(x) - A(\theta)\}\,, \tag{4}$$

with respect to some measure $\mu(dx)$ on the support. The transposition of a matrix (or vector) is represented by $'$ here and throughout the paper. $T(\cdot)$ is the sufficient statistic for the parameter θ. Recall that $X_1, \cdots, X_n$ are iid random variables from P_{θ_0}, where θ_0 is an interior point of Θ. We denote the first derivative and the second derivative of $A(\theta)$ by $\dot{A}(\theta)$ and $\ddot{A}(\theta)$ respectively. The (differential) entropy of P_θ is

$$H(\theta) = A(\theta) - \theta' \dot{A}(\theta)\,. \tag{5}$$

First let us consider the mixture code. The conjugate prior of (4) has the following form.

$$u(\theta) = \exp\{\alpha' \theta - \beta A(\theta) - B(\alpha, \beta)\}, \tag{6}$$

where α is a vector in $\mathbf{R}^d$, β is a scalar, and

$$B(\alpha, \beta) = \log \int_\Theta \exp\{\alpha' \theta - \beta A(\theta)\} \, d\theta . \tag{7}$$

In this paper, we assume that the prior of θ takes the form of a finite mixture of the conjugate distributions as in

$$w(\theta) = \sum_{j=1}^{J} \lambda_j \exp\{\alpha_j' \theta - \beta_j A(\theta) - B(\alpha_j, \beta_j)\} = \sum_{j=1}^{J} \lambda_j \, u_j(\theta) . \tag{8}$$

Not only this form is technically easy to handle for the purpose of obtaining the almost sure convergence, but also has it the virtue of approximating general priors at least for the Binomial and Gaussian cases. In the first case, the conjugate prior is the Beta distribution. Notice that any prior $w(\theta)$ continuous on $[0,1]$ can be uniformly approximated by the Bernstein polynomials (See [14])

$$\sum_{j=1}^{J} \frac{J!}{j!(J-j)!} \theta^j (1-\theta)^{J-j} w(\frac{j}{J}) = \sum_{j=1}^{J} \frac{1}{J+1} w(\frac{j}{J}) \, b_{j+1, J-j+1}(\theta) , \tag{9}$$

where $b_{r,s}(x) = (r+s-1)!/[(r-1)!(s-1)!] \, x^{r-1}(1-x)^{s-1}$ is the density function of the Beta distribution with integer parameters r and s. In order to construct a legitimate distribution, we need to normalize the weights in (9). Since $\frac{1}{J+1} w(\frac{j}{J})$ sums to unity as J goes large, we see that the prior $w(\theta)$ can be well approximated by a finite mixture of Beta distributions. In the case of the Gaussian distribution with unknown location parameter and known variance parameter, the conjugate prior is still Gaussian. Ferguson [10] remarked that an arbitrary density on the real line may be closely approximated by Gaussian mixtures.

Under prior $w(\cdot)$, the marginal distribution of $(x_1, ..., x_n)$ is

$$m(x_1, ..., x_n) = \int_{\Theta} p(x|\theta) w(\theta) d\theta, \tag{10}$$

and the codelength of the mixture code is given by $L_{mixture} = -\log m(X_1, \cdots, X_n)$. Hence the redundancy of the mixture code is $R_{mixture} = L_{mixture} - L_0$. For the predictive code, after obtaining observations $x_1, \cdots, x_i$, we estimate θ by the maximum likelihood estimate $\hat{\theta}_{i-1}$, and in turn encode the next observation according to this current estimated distribution. This procedure (Rissanen [16, 17]) has intimate connections with the prequential approach to statistical inference as advocated by Dawid [7, 8]. The redundancy of this coding scheme is given by $R_{predictive} = L_{predictive} - L_0$, where $L_{predictive} = (-\sum_{i=2}^{n} \log p(X_i|\hat{\theta}_{i-1}))$, and we take L_0 excluding the first observation in this situation for simplicity. The pathwise asymptotics of $R_{predictive}$ is given by the following theorem.

Theorem 2.1

$$R_{predictive} = \frac{d}{2} \log\, n - C_n(\omega)\, (\log \log\, n) + D_n(\omega)\,, \tag{11}$$

where the sequence of nonnegative random variables $\{C_n(\omega)\}$ has the following property,

$$\overline{\lim}_{n \to \infty} C_n(\omega) \quad \leq \quad d\,, \tag{12}$$

for almost all path ω's, and the sequence of random variables $\{D_n(\omega)\}$ converge to an almost surely finite random variable $D(\omega)$.

As a matter of fact, the upper limit of $C_n(\omega)$ is bounded below by 1 when $d > 1$ (more accurate bounds are difficult to obtain in this case). For general source distributions, it is difficult to obtain the third term $\{D_n(\omega)\}$. But in some specific cases, this term can be calculated explicitly. The following is an example.

Proposition 2.1 *Suppose the source of the data follows a Gaussian $N(\mu, \Sigma)$ with a known Σ, then the distributions of the random variables $\{D_n(\omega)\}$ and $\{D(\omega)\}$ do not depend on the parameters in the source distribution. The limit mean of $D_n(\omega)$ is given by*

$$\lim_{n\to\infty} E(D_n(\omega)) = E(D(\omega)) = \frac{d(c_{euler}+1)}{2}\,, \tag{13}$$

where c_{euler} is the Euler constant defined as $\lim_{n\to\infty}(\sum_{i=1}^{n} 1/i - \log n)$. Moreover, the expected redundancy of the predictive code satisfies

$$\sup_{\mu} E(R_{predictive}) = \frac{d}{2}\log\, n + \frac{d}{2}\, c_{euler} + o(1)\,. \tag{14}$$

A simulation study of the distribution of D_n can be found in the discussion section. We can also compare the predictive codelength with $n\,H(\hat{\theta}_n)$ to get an expansion on the regret of the predictive code as

Proposition 2.2

$$\begin{aligned} L_{predictive} &= (-\sum_{i=1}^{n} \log p(X_i|\hat{\theta}_n)) + \frac{d}{2}\log\, n + \tilde{D}_n(\omega) \\ &= nH(\hat{\theta}_n) + \frac{d}{2}\log\, n + \tilde{D}_n(\omega)\,, \end{aligned} \tag{15}$$

where the sequence of random variables $\{\tilde{D}_n(\omega)\}$ *converges to an almost surely finite random variable* $\tilde{D}(\omega)$. *In addition, if the source follows a Gaussian* $N(\mu, \Sigma)$, *then* $\tilde{D}_n(\omega) = D_n(\omega) + \log p(X_1|\theta_0)$, *where* $D_n(\omega)$ *is the same as that in Proposition 2.1.*

Now we turn to the mixture code length, whose pathwise asymptotics is as follows.

Theorem 2.2

$$R_{mixture} = \frac{d}{2} \log\ n - C_n(\omega)\,(\log\log\ n) - \frac{d}{2} \log\ 2\pi + \log \frac{(det\, I(\theta_0))^{1/2}}{w(\theta_0)} + o(1)\,, \tag{16}$$

where the sequence of nonnegative random variables $\{C_n(\omega)\}$ *has the following property,*

$$\overline{\lim}_{n\to\infty}\, C_n(\omega) \quad \leq \quad d\,, \tag{17}$$

for almost all path ω*'s.*

$\{C_n(\omega)\}$ is the same as that in Theorem 2.1. Similarly to Proposition 2.2, we have the following,

Proposition 2.3

$$\begin{aligned} L_{mixture} \quad &= \quad (-\sum_{i=1}^{n} \log p(X_i|\hat{\theta}_n)) + \frac{d}{2}\log\ n - \frac{d}{2}\log\ 2\pi + \log\frac{(det\, I(\theta_0))^{1/2}}{w(\theta_0)} + o(1) \\ &= \quad nH(\hat{\theta}_n) + \frac{d}{2}\log\ n - \frac{d}{2}\log\ 2\pi + \log\frac{(det\, I(\theta_0))^{1/2}}{w(\theta_0)} + o(1)\,. \end{aligned} \tag{18}$$

3 Discussion

3.1 The two-stage code

The two-stage code introduced by Rissanen ([18]) has a code length with the same leading terms as that of the mixture code. That is,

$$\begin{aligned} L_{two-stage} &= (-\sum_{i=1}^{n} \log p(X_i|\hat{\theta}_n)) + \frac{d}{2}\log\, n + Constant \\ &= nH(\hat{\theta}_n) + \frac{d}{2}\log\, n + Constant\,. \end{aligned} \tag{19}$$

Thus the iterated logarithmic pathwise expansion for the mixture code applies to the two-stage code except for the constant term. Later for Jeffreys priors Rissanen ([19]) modified the two-stage code to match the mixture code length not only the $\log n$ term but also the constant term by incorporating the Fisher information. This modified two-stage code is called Normalized Maximum Likelihood (NML) code. For example, if the source follows a binomial distribution, then the NML turns out to be equivalent to the mixture code with respect to the Jeffreys' prior. In this case, our pathwise expansion for the mixture code applies to the NML code up to the $o(1)$ term.

3.2 Comparing the mixture and predictive code

Comparing Theorem 2.1 and 2.2, we see that the redundancies of the predictive and mixture codes have the same $\log n$ term and $\log\log n$ term. They differ in the third term, which is finite in both cases. In the case of the mixture code, this term is a constant depending on the parameter θ_0 unless the prior is selected as Jeffreys' prior (cf. [6, 20]). For some families such as the Gaussian distribution, proper Jeffreys' priors do not exist over the natural parameter spaces so modifications have to be made, and this may result in a large constant term. In the case of a predictive code, the third term is a random variable.

If the source follows a Gaussian distribution, then the distribution of this random variable does not depend on the parameter of the source distribution. We compare the average redundancies of the mixture and predictive codes using (3) and (14). The constant term for the predictive code does not depend on the source distribution; whereas that for the mixture code does depend on the source distribution unless the prior is taken to be Jeffreys' (or its modifications).

In order to find the distribution of $D(\omega)$ in Proposition 2.1, we carried out a simulation study for the case $d = 1$ under the Gaussian assumption (with variance 1). We simulated 20,000 paths. For each of these paths, $D_{5,000}(\omega)$ in Proposition 2.1 was calculated. The histogram of these 20,000 $D_{5,000}(\omega)$'s is shown in Figure 1. Since $D(\omega)$ is the pathwise limit of $D_n(\omega)$, the distribution of $D_n(\omega)$ converges to the distribution of $D(\omega)$ as n gets large. Here we use the distribution of $D_{5,000}(\omega)$ to approximate the distribution of $D(\omega)$. The sample mean and standard deviation of the 20,000 $D_{5,000}(\omega)$'s are 0.793 and 1.386 respectively, whereas the theoretical mean of $D_{5,000}(\omega)$ is well approximated by $0.5 \times (1 + c_{euler}) = 0.789$, see Proposition 2.1. The theoretical standard deviation is difficult to calculate. In the Gaussian case, if $d > 1$, then $D(\omega)$ is the sum of d independent and identically distributed copies of that when $d = 1$ as discussed above. This can be seen from the proof of Theorem 2.1 (see the term R_1 in the proof).

4 Appendix

This section contains the proofs of the results in section 2 and 3. The following basic facts about the exponential family (4) are needed (see [3]).

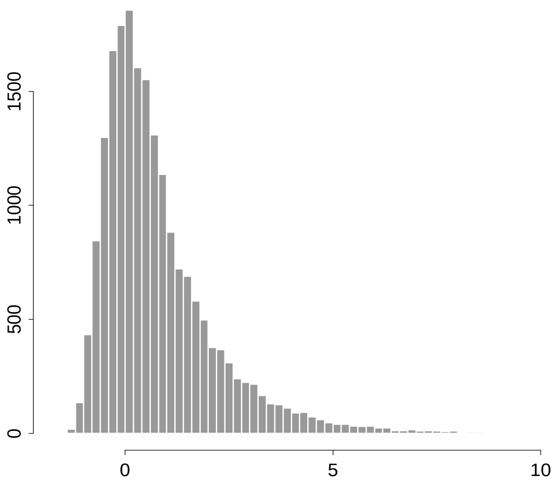

Figure 1: The histogram of $D_{5,000}(\omega)$ for a Gaussian source with $d = 1$. It is obtained from 20,000 paths.

1. $E(T(X)) = \dot{A}(\theta)$, and $Var(T(X)) = \ddot{A}(\theta)$.

2. $\dot{A}(\cdot)$ is one to one on the the natural parameter space.

3. The maximum likelihood estimate $\hat{\theta}_n$ based on $(X_1, \cdots, X_n)$ is given by

$$\hat{\theta}_n = \dot{A}^{-1}(\bar{T}_n)\,, \tag{20}$$

   where $\bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$.

4. The Fisher information matrix $I(\theta) = \ddot{A}(\theta)$.

5. The MLE $\hat{\theta}_n \longrightarrow \theta$ almost surely, since $\bar{T}_n \longrightarrow ET(X) = \dot{A}(\theta)$ almost surely by the strong law of large numbers.

Proof of Theorem 2.1. Let $G = \{\omega : \bar{T}_n(\omega) \longrightarrow \dot{A}(\theta_0)\}$. It is easy to see that $\hat{\theta}_n \longrightarrow \theta_0$ for each path in G. Since $P(G) = 1$, we only consider the paths in G from now on. The redundancy can be rearranged as

$$
\begin{aligned}
R_{predictive} &= \sum_{i=1}^{n-1}\{[A(\hat{\theta}_i) - \hat{\theta}_i' T(X_{i+1})] - [A(\theta_0) - \theta' T(X_{i+1})]\} \\
&= \sum_{i=1}^{n-1}\{-[T(X_{i+1}) - \dot{A}(\theta_0)]'(\hat{\theta}_i - \theta_0) + [A(\hat{\theta}_i) - A(\theta_0)] - \dot{A}'(\theta_0)(\hat{\theta}_i - \theta_0)\} \\
&= \sum_{i=1}^{n-1}\{-[T(X_{i+1}) - \dot{A}(\theta_0)]'(\hat{\theta}_i - \theta_0) + \frac{1}{2}(\hat{\theta}_i - \theta_0)'\ddot{A}(\theta_0)(\hat{\theta}_i - \theta_0) + r_i\,,
\end{aligned}
\tag{21}
$$

where r_i is the residual term in the Taylor expansion of $A(\hat{\theta}_i) - A(\theta_0)$ up to the second term. In the sequel, we represent the k-th component of a vector b by $b^{(k)}$. Then r_i can be expressed as

$$
r_i = \sum_{\{k,l,m\}} [\hat{\theta}_i^{(k)} - \theta_0^{(k)}][\hat{\theta}_i^{(l)} - \theta_0^{(l)}][\hat{\theta}_i^{(m)} - \theta_0^{(m)}]\, \frac{\partial^3 A(\theta)}{\partial\theta^{(k)}\,\partial\theta^{(l)}\,\partial\theta^{(m)}}\Big|_{\theta_0 + \beta(\hat{\theta}_i - \theta_0)}\,, \tag{22}
$$

where β is a number between 0 and 1, and the sum is calculated over all possible triples $\{k, l, m\}$. Furthermore, we express $\hat{\theta}_i - \theta_0$ by $\bar{T}_i - \dot{A}(\theta_0)$.

$$
\begin{aligned}
\hat{\theta}_i - \theta_0 &= \dot{A}^{-1}(\bar{T}_i) - \dot{A}^{-1}(\dot{A}(\theta_0)) \\
&= \ddot{A}^{-1}(\theta_0)(\bar{T}_i - \dot{A}(\theta_0)) + s_i\,,
\end{aligned}
\tag{23}
$$

where $s_i = (s_i^{(1)}, \cdots, s_i^{(d)})'$, and

$$
s_i^{(k)} = \sum_{l,m=1,\cdots,d} s_{i,(l,m)}^{(k)} [\bar{T}_i^{(l)} - \dot{A}(\theta_0)^{(l)}][\bar{T}_i^{(m)} - \dot{A}(\theta_0)^{(m)}]\,. \tag{24}
$$

$\{s^{(k)}_{i,(l,m)}\}$ are the third order derivatives evaluated at a point between θ_0 and $\hat{\theta}_i$. Plug this into (21), we have

$$R_{predictive} = R_1 + R_2 + R_3 \,, \tag{25}$$

where

$$\begin{aligned}
R_1 &= \frac{1}{2}\sum_{i=1}^{n-1}\{-2[T(X_{i+1}) - \dot{A}(\theta_0)]'\ddot{A}^{-1}(\theta_0)[\bar{T}_i - \dot{A}(\theta_0)] + [\bar{T}_i - \dot{A}(\theta_0)]'\ddot{A}^{-1}(\theta_0)[\bar{T}_i - \dot{A}(\theta_0)]\}\,, \\
R_2 &= \sum_{k,l,m}\sum_{i=1}^{n-1} a_{i,(k,l,m)}[\bar{T}^{(k)}_i - \dot{A}(\theta_0)^{(k)}][\bar{T}^{(l)}_i - \dot{A}(\theta_0)^{(l)}][\bar{T}^{(m)}_i - \dot{A}(\theta_0)^{(m)}]\,, \qquad (26) \\
R_3 &= \sum_{i=1}^{n-1}[T(X_{i+1}) - \dot{A}(\theta_0)]'\, s_i \\
&= \sum_{k=1}^{d}\sum_{i=1}^{n-1}[T(X_{i+1})^{(k)} - \dot{A}(\theta_0)^{(k)}]\, s^{(k)}_i \\
&= \sum_{k,l,m}\sum_{i=1}^{n-1} s^{(k)}_{i,(l,m)}[T(X_{i+1})^{(k)} - \dot{A}(\theta_0)^{(k)}][\bar{T}^{(l)}_i - \dot{A}(\theta_0)^{(l)}][\bar{T}^{(m)}_i - \dot{A}(\theta_0)^{(m)}]\,, \qquad (27)
\end{aligned}$$

where $\{a_{i,(k,l,m)}\}$ have absorbed relevant coefficients.

First let us study the main part R_1. Let

$$Y_i = \ddot{A}^{-\frac{1}{2}}(\theta_0)[T(X_i) - \dot{A}(\theta_0)] = (Y^{(1)}_i, Y^{(2)}_i, \cdots, Y^{(d)}_i)'\,. \tag{28}$$

Then we can check that $Y_1, \cdots, Y_n$ are i.i.d. with $EY_i = 0$, and $VarY_i = I_d$, where I_d is the identity matrix of order d. Let $\bar{Y}_i = \frac{1}{i}\sum_{j=1}^{i} Y_i$, $S_i = \sum_{j=1}^{i} Y_i = (S^{(1)}_i, S^{(2)}_i, \cdots, S^{(d)}_i)'$. Then

$$\begin{aligned}
R_1 &= \frac{1}{2}\sum_{i=1}^{n-1}\{\frac{-2Y'_{i+1}S_i}{i} + \frac{S'_iS_i}{i^2}\} \\
&= \frac{1}{2}\sum_{i=1}^{n-1}\{\frac{S'_iS_i}{i} - \frac{S'_{i+1}S_{i+1}}{i+1}\} + \frac{1}{2}\sum_{i=1}^{n-1}\{\frac{-2Y'_{i+1}S_i}{i} + \frac{S'_iS_i}{i^2} - \frac{S'_iS_i}{i} + \frac{S'_{i+1}S_{i+1}}{i+1}\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}(S_1'S_1 - \frac{S_n'S_n}{n}) + \frac{1}{2}\sum_{i=1}^{n-1}\{\frac{-2Y_{i+1}'S_i}{i} + \frac{S_i'S_i}{i^2} - \frac{S_i'S_i}{i} + \frac{S_i'S_i + 2Y_{i+1}'S_i + Y_{i+1}'Y_{i+1}}{i+1}\} \\
&= \frac{1}{2}(\| S_1 \|^2 - \frac{\| S_n \|^2}{n}) + \frac{1}{2}\sum_{i=1}^{n-1}\frac{\| S_i \|^2}{i^2(i+1)} - \sum_{i=1}^{n-1}\frac{Y_{i+1}'S_i}{i(i+1)} + \frac{1}{2}\sum_{i=1}^{n-1}\frac{\| Y_{i+1} \|^2}{i+1} \\
&= \frac{d}{2}\log\, n - \frac{\| S_n \|^2}{2n\log\log\, n}\log\log\, n + I_n, \qquad (29)
\end{aligned}
$$

where

$$
\begin{aligned}
I_n \;=\; & \frac{1}{2}\| S_1 \|^2 + \frac{1}{2}\sum_{i=1}^{n-1}\frac{\| S_i \|^2}{i^2(i+1)} - \sum_{i=1}^{n-1}\frac{Y_{i+1}'S_i}{i(i+1)} \\
& + \frac{1}{2}\sum_{i=1}^{n-1}(\frac{\| Y_{i+1} \|^2}{i+1} - \frac{d}{i+1}) + \frac{d}{2}(\sum_{i=1}^{n-1}\frac{1}{i+1} - \log\, n)\,. \qquad (30)
\end{aligned}
$$

Next we consider these terms separately. The first term is the leading term in (11). As for the second term, notice that

$$\frac{\| S_n \|^2}{2n\log\log\, n} = \sum_{k=1}^{d}[\frac{S_n^{(k)}}{\sqrt{2n\log\log\, n}}]^2\,.$$

Apply the law of iterative logarithm (see [9]) to each component $S_n^{(k)}$, namely,

$$
\begin{aligned}
\overline{\lim}_{n\to\infty}\frac{S_n^{(k)}}{\sqrt{2n\log\log\, n}} &= 1\,, \qquad (31) \\
\underline{\lim}_{n\to\infty}\frac{S_n^{(k)}}{\sqrt{2n\log\log\, n}} &= -1\,. \qquad (32)
\end{aligned}
$$

In fact, $C_n(\omega) = \frac{\|S_n\|^2}{2n\log\log\, n}$ in (11). Next we consider the terms in I_n separately.

1. The first term is finite almost surely.

2. Notice that $\sum_{i=1}^{n-1}\frac{\|S_i\|^2}{i^2(i+1)} = \sum_{k=1}^{d}\sum_{i=1}^{n-1}[\frac{S_i^{(k)2}}{i^2(i+1)}]$, and $E(\frac{S_i^{(k)2}}{i^2(i+1)}) = \frac{1}{i(i+1)}$. Thus

$$E(\sum_{i=1}^{+\infty}\frac{\| S_i \|^2}{i^2(i+1)}) = \sum_{i=1}^{+\infty}\frac{d}{i(i+1)} = d < +\infty\,.$$

Thus $\sum_{i=1}^{n}\frac{\|S_i\|^2}{i^2(i+1)} < +\infty$ converges almost surely and also in L^1.

3. Let $V_n = \frac{Y_n' S_{n-1}}{(n-1)n}$, then $E(V_n|Y_1,\cdots,Y_{n-1}) = 0$. Thus it is a martingale. And

$$\begin{aligned}
E(\| V_n \|^2) &= E[\frac{Y_n' S_{n-1} S_{n-1}' Y_n}{(n-1)^2 n^2}] \\
&= \frac{1}{(n-1)^2n^2}E[trace(S_{n-1}S_{n-1}'Y_nY_n')] \\
&= \frac{1}{(n-1)^2n^2}E[trace(S_{n-1}S_{n-1}'E(Y_nY_n'|Y_1,\cdots,Y_{n-1}))] \\
&= \frac{1}{(n-1)^2n^2}E[trace(S_{n-1}S_{n-1}'E(Y_nY_n'))] \\
&= \frac{d}{(n-1)^2n^2}E[trace(S_{n-1}S_{n-1}')] \\
&= \frac{d}{(n-1)^2n^2}E(\| S_{n-1} \|^2) \\
&= \frac{d^2}{(n-1)n^2}\,.
\end{aligned}$$

The expectation is evaluated by first conditioning on the σ algebra $\sigma(Y_1,\cdots,Y_{n-1})$. Since $\sum_{n=1}^{+\infty} EV_n^2 < +\infty$, $\frac{1}{2}\sum_{i=1}^{n-1}\frac{Y_{i+1}'S_i}{i(i+1)}$ converges almost surely, in L^2, and thus in L^1 (see [9] for the relevant result).

4. Rewrite the fourth term by each component of Y_i,

$$\sum_{i=1}^{n-1}(\frac{\| Y_{i+1} \|^2}{i+1} - \frac{d}{i+1}) = \sum_{i=1}^{n-1}\sum_{k=1}^{d}(\frac{Y_{i+1}^{(k)2}}{i+1} - \frac{1}{i+1}) = \sum_{k=1}^{d}\sum_{i=1}^{n-1}(\frac{Y_{i+1}^{(k)2}}{i+1} - \frac{1}{i+1})\,.$$

Notice that

$$E(\frac{Y_{i+1}^{(k)2}-1}{i+1})=0\,,\quad Var(\frac{Y_{i+1}^{(k)2}-1}{i+1})=\frac{2}{(i+1)^2}\,,$$

$$\sum_{i=1}^{+\infty}Var(\frac{Y_{i+1}^{(k)2}-1}{i+1})<+\infty\,.$$

Thus by the two-series Theorem (Page 49 and 218 in [9]), $\sum_{i=1}^{n-1}(\frac{\|Y_{i+1}\|^2}{i+1}-\frac{d}{i+1})$ converges almost surely, in L^2, and thus in L^1.

5. Finally,

$$\frac{d}{2}(\sum_{i=1}^{n-1}\frac{1}{i+1}-\log\, n)=\frac{d}{2}(c_{euler}-1)\,.$$

Now let us look at R_2. It is sufficient to consider one triple (k,l,m). According to the law of iterated logarithm, except for paths in a set with zero probability, the following is true

$$|\bar{T}_i^{(k)}-\dot{A}(\theta_0)^{(k)}|\leq 2\sqrt{\frac{2\log\log i}{i}}.$$

Also notice that $a_{i,(k,l,m)}$ is bounded as i gets large. Therefore the summand in (26) is at most of the order $(\log\log i)^{\frac{3}{2}}/i^{\frac{3}{2}}$. This means R_2 converges for almost all the paths.

As for R_3, notice that for each triple (k,l,m),

$$\sum_{i=1}^{n-1}s_{i,(l,m)}^{(k)}[T(X_{i+1})^{(k)}-\dot{A}(\theta_0)^{(k)}][\bar{T}_i^{(l)}-\dot{A}(\theta_0)^{(l)}][\bar{T}_i^{(m)}-\dot{A}(\theta_0)^{(m)}] \tag{33}$$

is a martingale with respect to $\sigma(X_1,\cdots,X_{n-1})$, and

$$E\{[s_{i,(l,m)}^{(k)}[T(X_{i+1})^{(k)}-\dot{A}(\theta_0)^{(k)}][\bar{T}_i^{(l)}-\dot{A}(\theta_0)^{(l)}][\bar{T}_i^{(m)}-\dot{A}(\theta_0)^{(m)}]]^2|X_1,\cdots,X_i\}$$

$$= \quad Var(T(X_1)^{(k)})[s_{i,(l,m)}^{(k)}[\bar{T}_i^{(l)} - \dot{A}(\theta_0)^{(l)}][\bar{T}_i^{(m)} - \dot{A}(\theta_0)^{(m)}]]^2 \,.$$

According to Theorem 2.17 in [12], (33) converges on the set

$$W = \{\omega : \sum_{i=1}^{+\infty}[s_{i,(l,m)}^{(k)}[\bar{T}_i^{(l)} - \dot{A}(\theta_0)^{(l)}][\bar{T}_i^{(m)} - \dot{A}(\theta_0)^{(m)}]]^2 < +\infty\} \,. \tag{34}$$

Again by the law of iterated logarithm and the fact that $s_{i,(l,m)}^{(k)}$ is bounded as i gets large, the summand in (34) is at most of order $(\log\log i)^2/i^2$. Therefore $P(W) = 1$. This completes the proof.

Proof of Proposition 2.1. In this case, R_2 and R_3 vanish in (25) (see (21) and (23)). Hence D_n is exactly I_n in (29). Since D_n converges to D in L^1, we have

$$E(D(\omega)) \quad = \quad \lim_{n\to\infty} E(I_n(\omega)) = \frac{d}{2}(c_{euler} + 1) \,.$$

We can obtain (14) by taking the expectation of the second term in (29) and using $E \parallel S_n \parallel^2 = n\,d$.

Proof of Proposition 2.2. Notice that

$$\begin{aligned}
& (-\sum_{i=1}^{n} \log p(X_i|\hat{\theta}_n)) - (-\sum_{i=1}^{n} \log p(X_i|\theta_0)) \\
= \quad & \sum_{i=1}^{n}\{-[T(X_n) - \dot{A}(\theta_0)]'(\hat{\theta}_n - \theta_0) + [A(\hat{\theta}_n) - A(\theta_0)] - \dot{A}(\theta_0)'(\hat{\theta}_n - \theta_0)\} \\
= \quad & -n[\bar{T}_n - \dot{A}(\theta_0)]'(\hat{\theta}_n - \theta_0) + n[A(\hat{\theta}_n) - A(\theta_0) - \dot{A}(\theta_0)'(\hat{\theta}_i - \theta_0)] \\
= \quad & -\frac{n}{2}[\bar{T}_n - \dot{A}(\theta_0)]'\ddot{A}^{-1}(\theta_0)[\bar{T}_n - \dot{A}(\theta_0)] + O(n \parallel \hat{\theta}_n - \theta_0 \parallel^3) \\
= \quad & -\frac{\parallel S_n \parallel^2}{2n} + o(1) \,,
\end{aligned} \tag{35}$$

where we use the Taylor expansions in (23). Next, we combine (25), (29) and (35). It is easy to check that the negative logarithm of the likelihood evaluated at the maximum likelihood estimate $\hat{\theta}_n$ is $nH(\hat{\theta}_n)$.

Proof of Theorem 2.2: If the prior is given by (6), then the marginal density of $(X_1, \cdots, X_n)$ is

$$m(x_1, \cdots, x_n) = \exp\{B(\sum_{i=1}^{n} T(x_i) + \alpha, n + \beta) - B(\alpha, \beta)\}\,, \tag{36}$$

according to the definition of $B(\cdot,\cdot)$. Therefore

$$\begin{aligned} R_m &= B(\alpha,\beta) - B(\sum_{i=1}^{n} T(X_i) + \alpha, n+\beta) - nA(\theta_0) + (\sum_{i=1}^{n} T(X_i))'\theta_0 \\ &= B(\alpha,\beta) - \log(\int_{\Theta} \exp\{nL_n(t,\omega)\}\,d\,t)\,, \end{aligned} \tag{37}$$

where

$$nL_n(t,\omega) = \alpha'\,t - \beta\,A(t) + [\sum_{i=1}^{n} T(X_i)]'(t-\theta_0) - n[A(t) - A(\theta_0)]\,. \tag{38}$$

Since $\bar{T}_n \longrightarrow \dot{A}(\theta_0)$ almost surely, for almost all the path, when n is sufficiently large, the unique minimum of $L_n(t,\omega)$ is achieved at

$$\tilde{\theta}_n(\omega) = \dot{A}^{-1}(\frac{\sum_{i=1}^{n} T(X_i) + \alpha}{n+\beta}) = \hat{\theta}_n(\omega) + O(\frac{1}{n}) \longrightarrow \theta_0\,.$$

Notice that

$$-\ddot{L}_n(t) = \frac{n+\beta}{n}\ddot{A}(t)\,.$$

Let

$$\Sigma = -\ddot{L}_n^{-1}(\tilde{\theta}_n(\omega)) = \frac{n}{n+\beta}\ddot{A}^{-1}(\tilde{\theta}_n(\omega))\,.$$

By expanding $L_n(t,\omega)$ at the saddle point $\tilde{\theta}_n(\omega)$ and applying the Laplace method (see [4, 21]), we have

$$\begin{aligned} \log(\int_\Theta \exp\{nL_n(t,\omega)\}\,d\,t) &= -\frac{d}{2}\log\,n + \frac{d}{2}\log\,2\pi + \frac{1}{2}\log(det\,\Sigma) + nL_n(\tilde{\theta}_n(\omega),\omega) + O(\frac{1}{n}) \\ &\longrightarrow -\frac{d}{2}\log\,n + \frac{d}{2}\log\,2\pi - \frac{1}{2}\log(det\,I(\theta_0)) + nL_n(\tilde{\theta}_n(\omega),\omega)\,. \end{aligned} \tag{39}$$

Next, we further expand $nL_n(\tilde{\theta}_n(\omega),\omega)$.

$$\begin{aligned} nL_n(\tilde{\theta}_n(\omega),\omega) &= \alpha'\,\tilde{\theta}_n(\omega) - \beta\,A(\tilde{\theta}_n(\omega)) + [\sum_{i=1}^n T(X_i)]'(\tilde{\theta}_n(\omega)-\theta_0) - n[A(\tilde{\theta}_n(\omega)) - A(\theta_0)] \\ &= \alpha'\,\theta_0 - \beta\,A(\theta_0) + n[\bar{T}_n'(\tilde{\theta}_n(\omega)-\theta_0) - \dot{A}(\theta_0)'(\tilde{\theta}_n(\omega)-\theta_0) \\ &\quad -\frac{1}{2}(\tilde{\theta}_n(\omega)-\theta_0)'\ddot{A}(\theta_0)(\tilde{\theta}_n(\omega)-\theta_0)] + o(1) \\ &= \alpha'\,\theta_0 - \beta\,A(\theta_0) + n[(\bar{T}_n - \dot{A}(\theta_0))'(\tilde{\theta}_n(\omega)-\theta_0) \\ &\quad -\frac{1}{2}(\tilde{\theta}_n(\omega)-\theta_0)'\ddot{A}(\theta_0)(\tilde{\theta}_n(\omega)-\theta_0)] + o(1) \\ &= \alpha'\,\theta_0 - \beta\,A(\theta_0) + \frac{n}{2}[(\bar{T}_n - \dot{A}(\theta_0))'\ddot{A}^{-1}(\theta_0)(\bar{T}_n - \dot{A}(\theta_0))] + o(1)\,. \end{aligned} \tag{40}$$

In the last step, we use the expansion

$$\tilde{\theta}_n(\omega) - \theta_0 = \ddot{A}^{-1}(\theta_0)(\bar{T}_n - \dot{A}(\theta_0)) + O(\|\,\bar{T}_n - \dot{A}(\theta_0)\,\|^2)\,.$$

Now we put (37), (39), and (40) together,

$$
\begin{aligned}
R_m &= \frac{d}{2}\log\, n - \frac{n}{2}[(\bar{T}_n - \dot{A}(\theta_0))'\ddot{A}^{-1}(\theta_0)(\bar{T}_n - \dot{A}(\theta_0))] - \frac{d}{2}\log\, 2\pi + \frac{1}{2}\log(det\, I(\theta_0)) \\
&\quad -\{\alpha'\,\theta_0 - \beta\, A(\theta_0) - B(\alpha,\beta)\} + o(1) \\
&= \frac{d}{2}\log\, n - \frac{\| S_n \|^2}{2n} - \frac{d}{2}\log\, 2\pi + \frac{1}{2}\log(det\, I(\theta_0)) - \log w(\theta_0) + o(1)\,, \qquad (41)
\end{aligned}
$$

where we use the notations in (28).

In the more general case when the prior is a finite mixture of conjugate distribution as in (8), it is easy to see from (10) and (36) that the marginal density is given by

$$
m(x_1,\cdots,x_n) = \sum_{j=1}^{J}\lambda_j\, \exp\{B(\sum_{i=1}^{n}T(x_i)+\alpha_j, n+\beta_j) - B(\alpha_j,\beta_j)\}\,. \qquad (42)
$$

According to (41), each term in this summation has the following almost sure asymptotics,

$$
\begin{aligned}
& \exp\{B(\sum_{i=1}^{n}T(x_i)+\alpha_j, n+\beta_j)\} \\
=\; & \exp\{-L_0 - \frac{d}{2}\log\, n + \frac{\| S_n \|^2}{2n} + \frac{d}{2}\log\, 2\pi - \frac{1}{2}\log(det\, I(\theta_0)) + \log u_j(\theta_0) + o(1)\} \\
=\; & [u_j(\theta_0)+o(1)]exp\{-L_0 - \frac{d}{2}\log\, n + \frac{\| S_n \|^2}{2n} + \frac{d}{2}\log\, 2\pi - \frac{1}{2}\log(det\, I(\theta_0))\}\,.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& m(x_1,\cdots,x_n) \\
=\; & [\sum_{j=1}^{J}\lambda_j\, u_j(\theta_0)+o(1)]\exp\{-L_0 - \frac{d}{2}\log\, n + \frac{\| S_n \|^2}{2n} + \frac{d}{2}\log\, 2\pi - \frac{1}{2}\log(det\, I(\theta_0))\}
\end{aligned}
$$

$$
\begin{aligned}
&= [w(\theta_0)+o(1)]\exp\{-L_0-\frac{d}{2}\log\, n+\frac{\parallel S_n\parallel^2}{2n}+\frac{d}{2}\log\, 2\pi-\frac{1}{2}\log(det\, I(\theta_0))\} \\
&= \exp\{-L_0-\frac{d}{2}\log\, n+\frac{\parallel S_n\parallel^2}{2n}+\frac{d}{2}\log\, 2\pi-\frac{1}{2}\log(det\, I(\theta_0))+\log w(\theta_0)+o(1)\}\,.
\end{aligned}
$$

Hence, (41) is valid even if the prior takes the more general form of a finite mixture of conjugates. Next, those lines regarding the term $\frac{\|S_n\|^2}{2n}$ in the proof of Proposition 2.2 can be used to finish the proof.

References

[1] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE. Trans. Inform. Theory.*, pages 2743–2760, 1998.

[2] A. R. Barron and N. Hengartner. Information theory and superefficiency. *Annals of Statistics*, page to appear, 1998.

[3] L. D. Brown. *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics: California,, 1987.

[4] N. G. De Bruijn. *Asymptotic Methods in Analysis.* North-Holland: Amsterdam, 1958.

[5] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayesian methods. *IEEE Transactions on Information Theory*, 36:453–471, 1990.

[6] B. S. Clarke and A. R. Barron. Jeffrey' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–64, 1994.

[7] A. P. Dawid. Present position and potential developments: some personal views, statistical theory, the prequential approach. *J. R. Stat. Soc. Ser. B*, 147, 1984.

[8] A. P. Dawid. Prequential analysis, stochastic complexity and Bayesian inference. In *Fourth Valencia International Meeting on Bayesian Statistics*, pages 15–20. 1992.

[9] R. Durrett. *Probability: Theory and Examples.* Wadsworth & Brooks/Cole, 1991.

[10] T. S. Ferguson. Bayesian density estimation via mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302. Academic Press, New York, 1983.

[11] Y. Freund. Predicting a binary sequence almost as well as the optimal biased coin. *Unpublished manuscript.*

[12] P. Hall and C. C. Heyde. *Martingale Limit Theorem and Its Application.* Academic Press: New York, London, 1980.

[13] M. Hansen and B. Yu. Model selection and minimum description length principle. *JASA*, 1998. submitted.

[14] M. J. D. Powell. *Approximation theory and methods.* Cambridge University Press: Cambridge, 1981.

[15] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[16] J. Rissanen. A predictive least squares principle. *IMA Journal of Mathematical Control and Information*, 3:211–222, 1986.

[17] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

[18] J. Rissanen. *Stochastic Complexity and Statistical Inquiry.* World Scientific: Singapore, 1989.

[19] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory*, 42:40–47, 1996.

[20] J. Takeuchi and A. R. Barron. Asymptotically minimax regret by Bayesian mixtures. In *Proceedings IEEE 1998 International Symposium on Information Theory.*

[21] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.