ON THE STOCHASTIC COMPLEXITY FOR ORDER-1 MARKOV CHAINS AND THE CATALAN CONSTANT

Ciprian Doru Giurcăneanu[1], Dumitru Mihalache[2], Moubinool Omarjee[3] and Marian Tetiva[2]

[1]*Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland, ciprian.giurcaneanu@tut.fi,*
[2]*"Gheorghe Roşca Codreanu" National Collegium, 11 Nicolae Bălcescu street, 731183 Bârlad, Romania,*
[3]*Lycée Jean-Lurçat, 48 avenue des Gobelins, 75013 Paris, France*

Abstract: The stochastic complexity (SC) selects from a given family of parametric models the one that yields the shortest code length for the available measurements. Theoretical developments have made possible the evolution of the SC from the "two-part code" formula to the most recent expression based on the so-called Normalized Maximum Likelihood (NML) distribution. The application of the NML criterion is recommended especially in the case of small sample size, but high computational burdens prevent its general use. During recent years increasing interest has been growing in obtaining an approximation of the NML-based SC by use of the Fisher information matrix. We show in this note that the most important step in working out such an approximate expression for order-1 Markov chains is the calculation of an integral that leads to the Catalan constant. We evaluate the accuracy of the approximation for small, moderate, and large samples, and we illustrate the use of the formula in model selection.

Keywords: stochastic complexity, model selection, order-1 Markov chains, Catalan constant, generalized Fisher information matrix, computational complexity.

1. INTRODUCTION

Model selection is an important paradigm that has received a considerable attention in the statistics community over the years. A principled method to select a particular model from a class M of models is based on the evaluation of the stochastic complexity (SC). The very first formula for SC was introduced in (Rissanen, 1978) as the celebrated "two-part code", and it was further refined in (Barron et al., 1998; Rissanen, 1978, 1996, 2000). The method is rooted in information theory (Cover and Thomas, 1991), and it relies on a coding scenario for transmitting the measurements $x^n = x_1 \ldots x_n$ from a hypothesized encoder to a decoder. We conventionally employ the notation $X^n = X_1 \ldots X_n$ for the corresponding stochastic process, and $x^n = x_1 \ldots x_n$ for a specific realization. The selection procedure chooses that model from M family that allows the data to be encoded with the shortest code length, or equivalently to minimize SC.

We focus on the model selection for strings with entries from the alphabet $A = \{0,1\}$. Our choice is motivated by the fundamental role of the binary alphabet in information theoretic criteria.

The calculation of SC relies on the Normalized Maximum Likelihood (NML) distribution (Barron et al., 1998), which for a model Ξ with parameters θ is given by

$$\hat{f}(x^n;\Xi)=\frac{f_\Xi(x^n;\hat{\theta}_{ML}(x^n))}{\sum_{y^n\in A^n} f_\Xi(y^n;\hat{\theta}_{ML}(y^n))}, \qquad (1)$$

where $\hat{\theta}_{ML}(\cdot)$ denotes the maximum likelihood estimate and $f_\Xi(y^n;\theta)$ the likelihood function for an arbitrary binary string y^n assumed to have been generated according to model Ξ. The stochastic complexity is defined as $L(x^n;\Xi)=-\ln \hat{f}(x^n;\Xi)$, where $\ln(\cdot)$ is the natural logarithm. The SC is then expressed in *nats*. The binary logarithm can be used equally well, and then SC will be expressed in *bits*.

If the maximum likelihood estimates satisfy the Central Limit Theorem and some weak smoothness conditions are verified, then SC can be approximated with the formula (Rissanen, 1996):

$$L(x^n;\Xi)=-\ln f_\Xi(x^n;\hat{\theta}_{ML}(x^n))+ \frac{k}{2}\ln\frac{n}{2\pi}+\ln\int_\Theta |\mathbf{J}(\theta)|^{1/2}\,\mathrm{d}\theta+o(1), \qquad (2)$$

where k is the number of parameters in the model Ξ, Θ denotes the entire parameter space, and $\mathbf{J}(\theta)$ is the Fisher information matrix with entries

$$J_{ij}(\theta)=-\lim_{n\to\infty}\frac{1}{n}E\left[\frac{\partial^2 \ln f(x^n;\theta)}{\partial\theta_i\partial\theta_j}\right]$$

$1\le i,j\le k$. Both practical and theoretical considerations explain the recent interest in the expression of SC given by (2), but it was already noticed that deriving closed-form formulae from (2) is not an easy task (Hanson and Fu, 2005).

The aim of this note is to elaborate on formula (2) for order-1 Markov processes and to investigate its connections with the Catalan constant G. The definition of the constant is given by

$$G=\sum_{j=0}^{\infty}\frac{(-1)^j}{(2j+1)^2},$$

and it is approximated as $G\approx 0.915965594177\ldots$. We mention that in various studies the Catalan constant is denoted by C or K.

The crucial step in working out the expression of SC for order-1 Markov chains is the evaluation of the integral

$$I=\int_0^1\int_0^1\frac{\mathrm{d}\alpha\mathrm{d}\beta}{(\alpha+\beta)\sqrt{1-\alpha}\sqrt{1-\beta}}, \qquad (3)$$

which has the value $I=4G$. Investigating integrals and series associated with G is a time honoured research topic (Adamchik, 2002; Bradley, 2001). The interested reader can find at *http://www.cs.cmu.edu/~adamchik/articles/catalan/catalan.htm* an impressive list of such results that have been verified by using Mathematica©. A proof for the identity $I=4G$ can be found in (Bradley, 2001).

The rest of the paper is organized as follows. In the next Section we briefly revisit the stochastic complexity for Bernoulli and Markov models. Note that the main results are given for the Markov models, and the results on Bernoulli models are included only for pedagogical reasons. For completeness, we give in Section 3 three proofs for the identity $I=4G$. In the last Section we compare the asymptotic approximation (2) with another SC approximation that is grounded in analysis of algorithms (Jacquet and Szpankowski, 2004). Section 4 also contains a short discussion on SC for large order Markov models.

2. A CASE STUDY ON MODEL SELECTION

For concreteness we assume that the model selection problem reduces to the decision whether the samples $x_1,\ldots,x_n$ are outcomes from a Bernoulli distribution or from an order-1 Markov process. We use the notation $Be(\delta)$ for the Bernoulli distribution whose parameter is $\delta=P(X_t=1)$, where t is an arbitrary time moment between 1 and n. For circumventing some computational difficulties, we accept as the working hypothesis that $\delta\in(0,1)$, so that the Bernoulli parameter cannot be neither zero nor one. A more detailed discussion on this hypothesis can be found in (Rissanen, 1996).

We assume that the order-1 Markov chain is time invariant:
$P(X_{t+1}=j\mid X_t=i)=P(X_2=j\mid X_1=i)$ for all $i,j\in A$ and $1<t<n$. With the convention that both the rows and the columns of the probability transition matrix $\mathbf{Q}$ are indexed from zero, the entries of $\mathbf{Q}$ are given by $Q_{ij}=P(X_{t+1}=j\mid X_t=i)$, where $i,j\in A$. Without loss of generality we further write $\mathbf{Q}=\begin{bmatrix}1-\alpha & \alpha\\ \beta & 1-\beta\end{bmatrix}$, where $\alpha,\beta\in(0,1)$ and

$\alpha \neq \beta$. We assume that $\alpha + \beta \neq 1$, because otherwise the model reduces to a memoryless source. For sake of simplicity we use the notation $Ma_1(\alpha, \beta)$ for the Markov chain introduced above. Note that $Ma_1(\alpha, \beta)$ is ergodic since the transition probabilities are chosen to be strictly positive (Cover and Thomas, 1991). Therefore, in our case study, the considered model class is $M = \{Be(\delta), Ma_1(\alpha, \beta)\}$. It is straightforward to apply the SC formula (2) for $Be(\delta)$. With the notations $n_1 = \sum_{i=1}^{n} x_i$ and $n_0 = n - n_1$, $f_{Be}(x^n; \delta) = (1-\delta)^{n_0} \delta^{n_1}$, and the Fisher information matrix reduces to the scalar $\frac{1}{\delta(1-\delta)}$ (Rissanen, 1996). Using the well-known expression of the ML estimate for the Bernoulli model and performing elementary calculations, it is easy to show that

$$L(x^n; Be(\hat{\delta}_{ML}(x^N))) = -\sum_{i=0}^{1} n_i \ln \frac{n_i}{n} + \frac{1}{2} \ln n + \frac{1}{2} \ln \frac{\pi}{2} + o(1) \quad (4)$$

The Markov models satisfy all conditions (Rissanen, 1996) required for the approximate formula (2), and we derive next the SC expression for $Ma_1(\alpha, \beta)$. The likelihood function is given by $f_{Ma_1}(x^n; \alpha, \beta) = P(X_1 = x_1) \prod_{i=0}^{1} \prod_{j=0}^{1} Q_{ij}^{n_{ij}}$, where n_{ij} denotes the number of times the symbol j occurs immediately after symbol i in the string x^n. Since no straightforward relationship exists between the stationary probability $P(X_1 = x_1)$ and the parameters α and β, we consider the probability of the string x^n, conditioned on the very first symbol, namely $f_{Ma_1}(x^n; \alpha, \beta, x_1) = \prod_{i=0}^{1} \prod_{j=0}^{1} Q_{ij}^{n_{ij}}$ (Atteson, 1999).

Let $\theta = (\alpha, \beta)$. With slight abuse of notation we define $n_0 = n_{00} + n_{01}$ and $n_1 = n_{10} + n_{11}$. Notice that the identity $n_0 + n_1 = n$ is verified for the Bernoulli model, whereas $n_0 + n_1 = n - 1$ for the order-1 Markov model. We calculate the entries of the Fisher information matrix $\mathbf{J}(\theta)$: $\frac{\partial^2 \ln f_{Ma_1}(x^n; \theta, x_1)}{\partial \theta_\ell \partial \theta_m} = \sum_{i=0}^{1} \sum_{j=0}^{1} n_{ij} \frac{\partial^2 \ln Q_{ij}}{\partial \theta_\ell \partial \theta_m}$. It is elementary to prove the identities

$$J_{11}(\alpha, \beta) = \lim_{n \to \infty} \frac{1}{n} E\left[n_{00} \frac{1}{(1-\alpha)^2} + n_{01} \frac{1}{\alpha^2} \right],$$

$J_{12}(\alpha, \beta) = J_{21}(\alpha, \beta) = 0$, and

$$J_{22}(\alpha, \beta) = \lim_{n \to \infty} \frac{1}{n} E\left[n_{10} \frac{1}{\beta^2} + n_{11} \frac{1}{(1-\beta)^2} \right].$$

Moreover, $\lim_{n \to \infty} \frac{1}{n} E[n_{11}] = (1-\beta) P_s(1)$, where $P_s(1)$ is the stationary probability for state 1 (Atteson, 1999). The two states converge to their stationary distribution $\mathbf{P_s} = [P_s(0) \quad P_s(1)]^{\mathrm{T}}$, and the column vector $\mathbf{P_s}$ verifies $\mathbf{P_s}^{\mathrm{T}} \mathbf{Q} = \mathbf{P_s}^{\mathrm{T}}$. Since the entries of $\mathbf{P_s}$ are constrained to sum to one we obtain $P_s(0) = \frac{\beta}{\alpha + \beta}$ and $P_s(1) = \frac{\alpha}{\alpha + \beta}$ (Cover and Thomas, 1991). Consequently, $\lim_{n \to \infty} \frac{1}{n} \mathrm{E}[n_{11}] = \frac{\alpha(1-\beta)}{\alpha + \beta}$.

It is easy to check that $\mathbf{J}(\alpha, \beta) = \frac{1}{\alpha + \beta} \begin{bmatrix} \frac{\beta}{\alpha(1-\alpha)} & 0 \\ 0 & \frac{\alpha}{\beta(1-\beta)} \end{bmatrix}$. Therefore the calculation of the integral term in (2) reduces to the evaluation of (3). The identity $I = 4G$ leads to

$$L(x^n; Ma_1(\hat{\theta}_{ML}(x^n))) = -\sum_{i=0}^{1} \sum_{j=0}^{1} n_{ij} \ln \frac{n_{ij}}{n_i} + \ln n + \ln \frac{2G}{\pi} + o(1),$$

where we have used the well-known result $f_{Ma_1}(x^n; \hat{\theta}_{ML}(x^n), x_1) = \prod_{i=0}^{1} \prod_{j=0}^{1} \left(\frac{n_{ij}}{n_i} \right)^{n_{ij}}$ (Finesso et al., 1996). Since the cost of transmitting the symbol x_1 from the encoder to the decoder is one bit, or equivalently $\ln 2$ nats, we obtain the following expression for the SC of order-1 Markov chains:

$$L(x^n; Ma_1(\hat{\theta}_{ML}(x^n))) = -\sum_{i=0}^{1} \sum_{j=0}^{1} n_{ij} \ln \frac{n_{ij}}{n_i} + \ln n + \ln \frac{4G}{\pi} + o(1). \quad (5)$$

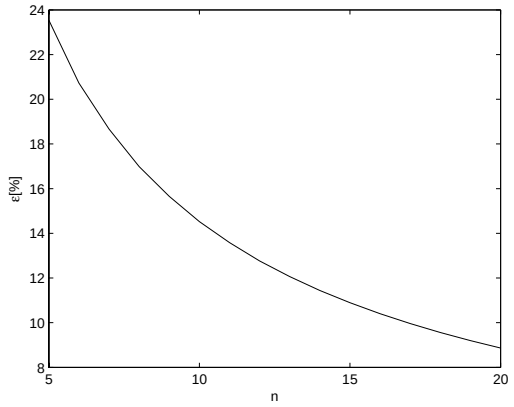

Fig. I. The percentage error ε versus the sample size n.

Table 1 After observing x^n, we compute the SC with formulae (4) and (5), and choose the model associated with the minimum SC. The empirical probability of selecting the correct model, $\widetilde{P}_c$, is evaluated in two different experiments for various values of the sample size n. Each reported $\widetilde{P}_c$ is calculated based on 10^6 trials.

Experiment 1: the entries of x^n are outcomes from $Be(0.8)$				
n	50	100	200	400
$\tilde{P}_c$	0.821	0.897	0.940	0.962
Experiment 2: the entries of x^n are outcomes from $Ma_1(0.1, 0.2)$				
n	50	100	200	400
$\tilde{P}_c$	0.973	0.999	1.000	1.000

Notice that the formula (5) is obtained for asymmetric Markov models. The Markov model is symmetric if the two conditions hold simultaneously: $\alpha = \beta$ and $\alpha \neq 1/2$. In this case, the Fisher information matrix reduces to the scalar $\frac{1}{\alpha(1-\alpha)}$, and the term $\frac{k}{2}\ln\frac{n}{2\pi}$ in (2) becomes $\frac{1}{2}\ln\frac{n}{2\pi}$. If we ignore the cost for transmitting the symbol x_1, the SC criterion for deciding between symmetric Markov and Bernoulli models is equivalent with the ML criterion.

Equation (5) is an asymptotic approximation for the exact NML distribution given by (1). More precisely, the logarithm of the normalization factor from (1), $\ln C_n = \ln \sum_{y^n \in A^n} f_{Ma_1}(y^n; \hat{\theta}_{ML}(y^n))$, is approximated by $\ln \widetilde{C}_n = \ln n + \ln \frac{4G}{\pi}$. To gain more insight we calculate the percentage error $\varepsilon = 100 \frac{\left|\ln \widetilde{C}_n - \ln C_n\right|}{\ln C_n}$ for small sample sizes, and we plot in Fig. I the value of ε when n varies between five and twenty. Notice in Figure 1 that ε decreases monotonically with increasing n, and the percentage error is as small as 9% when the sample size is twenty.

2. INTEGRALS RELATED TO *G*

We begin this Section with a well-known result:

Lemma 1. The following expressions of G

$$\int_0^{\frac{\pi}{4}} \ln(\tan x)\mathrm{d}x = \frac{1}{2}\int_0^{\frac{\pi}{2}} \ln(\tan\frac{x}{2})\mathrm{d}x = -G. \qquad (6)$$

are true.

Proof. Integrating by parts we have

$$\int_0^{\frac{\pi}{4}} \ln(\tan x)\,\mathrm{d}\,x = \left[x\ln(\tan x)\right]_0^{\frac{\pi}{4}}$$

$$-\int_0^{\frac{\pi}{4}} \frac{x}{\sin x\cos x}\,\mathrm{d}\,x = -\int_0^{\frac{\pi}{4}} \frac{2x}{\sin 2x}\,\mathrm{d}\,x$$

because

$$\lim_{x\to 0} x\ln(\tan x) = \lim_{x\to 0} \frac{x}{\tan x}(\tan x \ln(\tan x)) = 0$$

Then we put $\tan x = u \Leftrightarrow x = \arctan u$ ($\sin 2x = \frac{2u}{1+u^2}$, $\mathrm{d}x = \frac{1}{1+u^2}\mathrm{d}u$), which implies $\int_0^{\frac{\pi}{4}} \frac{2x}{\sin 2x}\mathrm{d}x = \int_0^1 \frac{\arctan u}{u}\mathrm{d}u$. One can calculate this integral with the series development $\arctan u = \sum_{j=0}^{\infty} \frac{(-1)^j u^{2j+1}}{2j+1}$. Note that the series is uniformly convergent and can be integrated term by term on the interval $[0,1]$. Consequently we obtain

$$-\int_0^{\frac{\pi}{4}} \ln(\tan x)\,\mathrm{d}\,x = \int_0^1 \frac{\arctan u}{u}\,\mathrm{d}\,u$$

$$= \sum_{j=0}^{\infty} \int_0^1 \frac{(-1)^j u^{2j}}{2j+1}\,\mathrm{d}\,u = \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)^2} = G.$$

The change of variable $x = y/2$ in the first integral from (6) leads immediately to the second integral and this ends the proof.

Proposition 1. The double integral

$$I=\int_0^1\int_0^1\frac{\mathrm{d}\alpha\mathrm{d}\beta}{(\alpha+\beta)\sqrt{1-\alpha}\sqrt{1-\beta}}$$

equals $4G$.

Proof. First method. We change the variables $\sqrt{1-\alpha}=u,\sqrt{1-\beta}=v$; u and v go along the interval $[0,1]$ and we get $\alpha=1-u^2$, $\beta=1-v^2$; consequently $\mathrm{d}\alpha=-2u\mathrm{d}u$ and $\mathrm{d}\beta=-2v\mathrm{d}v$. Our integral becomes $I=4\int_0^1\int_0^1\frac{1}{2-u^2-v^2}\mathrm{d}u\mathrm{d}v=4J$, and then we split the J integral into two separate integrals: $J=\iint_{S_1}\frac{1}{2-u^2-v^2}\mathrm{d}u\mathrm{d}v+\iint_{S_2}\frac{1}{2-u^2-v^2}\mathrm{d}u\mathrm{d}v$. Here S_1 denotes the quarter of the unit disc inside the unit square, $S_1=\{(u,v)\in[0,1]\times[0,1]:u^2+v^2\le 1\}$, and $S_2=[0,1]\times[0,1]-S_1$. One can easily compute the first integral (denoting $u=r\cos t, v=r\sin t$):

$$\iint_{S_1}\frac{1}{2-u^2-v^2}\mathrm{d}u\,\mathrm{d}v=\int_0^1\frac{r}{2-r^2}\mathrm{d}r\int_0^{\frac{\pi}{2}}\mathrm{d}t$$
$$=\frac{\pi}{4}\ln 2.$$

Applying the same $u=r\cos t$ and $v=r\sin t$ substitutions, the second integral turns into

$$K=\int_1^{\sqrt{2}}\Big(\int_{\arccos\frac{1}{r}}^{\frac{\pi}{2}-\arccos\frac{1}{r}}\mathrm{d}t\Big)\frac{r}{2-r^2}\mathrm{d}r$$
$$=\int_1^{\sqrt{2}}\frac{r}{2-r^2}\left(\frac{\pi}{2}-2\arccos\frac{1}{r}\right)\mathrm{d}r,$$

which is not an improper integral because the limit of the integrated function in $\sqrt{2}$ is finite. Further we obtain by integration by parts

$$K=-\int_1^{\sqrt{2}}\left(\ln(2-r^2)\right)'\left(\frac{\pi}{4}-\arccos\frac{1}{r}\right)\mathrm{d}r$$
$$=-\left[\ln(2-r^2)\left(\frac{\pi}{4}-\arccos\frac{1}{r}\right)\right]_1^{\sqrt{2}}$$
$$-\int_1^{\sqrt{2}}\frac{\ln(2-r^2)}{r\sqrt{r^2-1}}\mathrm{d}r=-\int_1^{\sqrt{2}}\frac{\ln(2-r^2)}{r\sqrt{r^2-1}}\mathrm{d}r.$$

We used for the limits in $\sqrt{2}$: $\lim_{r\to\sqrt{2}}(\sqrt{2}-r)\ln(2-r^2)=0$ and $\lim_{r\to\sqrt{2}}\frac{\frac{\pi}{4}-\arccos\frac{1}{r}}{\sqrt{2}-r}=\frac{1}{\sqrt{2}}$.

Now we make the change of variable $\arccos\frac{1}{r}=p\Leftrightarrow r=\frac{1}{\cos p}$, which leads to

$$-K=\int_1^{\sqrt{2}}\frac{\ln(2-r^2)}{r\sqrt{r^2-1}}\mathrm{d}r$$
$$=\int_0^{\frac{\pi}{4}}\frac{\ln\left(2-\frac{1}{\cos^2 p}\right)}{\frac{1}{\cos p}\frac{\sin p}{\cos p}}\frac{\sin p}{\cos^2 p}\mathrm{d}p$$
$$=\int_0^{\frac{\pi}{4}}\ln(1-\tan^2 p)\,\mathrm{d}p$$

giving

$$-K=\int_0^{\frac{\pi}{4}}\ln(1-\tan p)\mathrm{d}p+\int_0^{\frac{\pi}{4}}\ln(1+\tan p)\mathrm{d}p.$$

We also use the substitution $p=\frac{\pi}{4}-x$ in the first integral, which yields

$$-K=\int_0^{\frac{\pi}{4}}\ln(2\tan x)\mathrm{d}x=\frac{\pi}{4}\ln 2+\int_0^{\frac{\pi}{4}}\ln(\tan x)\mathrm{d}x$$

So we have $I=-4\int_0^{\frac{\pi}{4}}\ln(\tan x)\mathrm{d}x=4G,$ by (6).

In the second method (a variant of the first one) we simplify the calculation of the integral of the function $\frac{1}{2-u^2-v^2}$ on the surface S_2 by changing the integration order. Based on the symmetry of S_2 we write $K=2\int_0^{\frac{\pi}{4}}\left(\int_1^{\frac{1}{\cos t}}\frac{r}{2-r^2}\mathrm{d}r\right)\mathrm{d}t$.

Since $\int_1^{\frac{1}{\cos t}}\frac{r}{2-r^2}\mathrm{d}r=-\frac{1}{2}\ln(1-\tan^2 t)$, we obtain again $K=-\int_0^{\frac{\pi}{4}}\ln(1-\tan^2 t)\mathrm{d}t$ and, of course, we continue as before.

The third method also starts with $I=4J$, where $J=\int_0^1\left(\int_0^1\frac{1}{2-v^2-u^2}\mathrm{d}u\right)\mathrm{d}v,$ but now we compute first the inner integral:

$$\int_0^1 \frac{1}{2-v^2-u^2}\,\mathrm{d}u = \frac{1}{2\sqrt{2-v^2}}\left[\ln\frac{\sqrt{2-v^2}+u}{\sqrt{2-v^2}-u}\right]_0^1$$

$$= \frac{1}{2\sqrt{2-v^2}}\ln\frac{\sqrt{2-v^2}+1}{\sqrt{2-v^2}-1}$$

Therefore we have

$J = \int_0^1 \frac{1}{2\sqrt{2-v^2}}\ln\frac{\sqrt{2-v^2}+1}{\sqrt{2-v^2}-1}\,\mathrm{d}v$ and

$I = -2\int_0^1 \frac{1}{\sqrt{2-v^2}}\ln\frac{\sqrt{2-v^2}-1}{\sqrt{2-v^2}+1}\,\mathrm{d}v$. Here we change variables with $v = \sqrt{2}\sin s$, $\mathrm{d}v = \sqrt{2}\cos s\,\mathrm{d}s$ to get

$$I = -2\int_0^{\frac{\pi}{4}} \frac{1}{\sqrt{2}\cos s}\ln\frac{\sqrt{2}\cos s - 1}{\sqrt{2}\cos s + 1}\sqrt{2}\cos s\,\mathrm{d}s$$

$$= -2\int_0^{\frac{\pi}{4}} \ln\frac{\cos s - \cos\frac{\pi}{4}}{\cos s + \cos\frac{\pi}{4}}\,\mathrm{d}s$$

Then we transform in products the cosines sum and the difference, and obtain

$$I = -2\int_0^{\frac{\pi}{4}} \ln\tan\left(\frac{\pi}{8}-\frac{s}{2}\right)\mathrm{d}s - 2\int_0^{\frac{\pi}{4}} \ln\tan\left(\frac{\pi}{8}+\frac{s}{2}\right)\mathrm{d}s.$$

Simple changes of variable ($\frac{\pi}{8}-\frac{s}{2} = p$ and $\frac{\pi}{8}+\frac{s}{2} = q$) transform the two integrals into $\int_0^{\frac{\pi}{4}} \ln\tan\left(\frac{\pi}{8}-\frac{s}{2}\right)\mathrm{d}s = 2\int_0^{\frac{\pi}{8}} \ln\tan p\,\mathrm{d}p$ and $\int_0^{\frac{\pi}{4}} \ln\tan\left(\frac{\pi}{8}+\frac{s}{2}\right)\mathrm{d}s = 2\int_{\frac{\pi}{8}}^{\frac{\pi}{4}} \ln\tan q\,\mathrm{d}q$.

Therefore we have $I = -4\int_0^{\frac{\pi}{4}} \ln\tan x\,\mathrm{d}x$, which concludes the proof.

We mention in the end of this Section that problem 2793, proposed by P. Deiermann in 2002, in the *Crux Mathematicorum* Magazine, requires the calculation of the area of the surface obtained from the quarter of the unit disc contained in the first quadrant by the mapping $\zeta = \cosh^{-1}(z)$. Along the solution that was published in the same magazine in 2003, pages 522-524, integrals akin to those from this note are calculated.

4. FINAL REMARKS

In (Jacquet and Szpankowski, 2004), combinatorial methods of analysis of algorithms are applied to find an asymptotic approximation for $\log_2 C_n = \log_2 \sum_{y^n \in A^n} f_{Ma_1}\left(y^n; \hat{\theta}_{ML}\left(y^n\right)\right)$, and the following result is obtained when $A = \{0,1\}$: $\log_2 C_n = \log_2 n + \log_2 \frac{8G}{\pi} + o(1)$. The approximation is further refined by subtracting the term $\frac{\ln\ln 2}{\ln 2}$, which is due to the constraint that SC be integer-valued.

It is interesting to observe that the Catalan constant appears also in formula from (Jacquet and Szpankowski, 2004), even if their asymptotic approximation is not based on the generalized Fisher information matrix.

A natural question arises: is it possible to work out the formula (2) for Markov models whose order is $r > 1$? First we give a result proven in (Atteson, 1999): the square root of the determinant of the Fisher information matrix is

$$\left|\mathbf{J}(\theta)\right|^{1/2} = \prod_{z^r \in A^r}\left(\frac{P_s(z^r)}{\prod_{i=0}^{1} P\left(X_{t+1} = i \mid X_{t-r}^{t-1} = z^r\right)}\right)^{1/2} \quad (7)$$

where $P_s(\cdot)$ denotes the stationary probabilities. We illustrate the difficulties that occur in calculations by applying this formula for $r = 2$. To fix the ideas, let us assume that the probability transition matrix for the order-2 Markov binary chain is

$$\mathbf{Q} = \begin{bmatrix} 1-\alpha & \alpha & 0 & 0 \\ 0 & 0 & 1-\beta & \beta \\ \gamma & 1-\gamma & 0 & 0 \\ 0 & 0 & \delta & 1-\delta \end{bmatrix}$$ (Good,1963).

We denote $\theta = (\alpha, \beta, \gamma, \delta)$ and then we employ in (7) the expressions of the stationary probabilities derived in (Good, 1963) under the usual regularity conditions. We further obtain

$$\left|\mathbf{J}(\theta)\right|^{1/2} = \frac{1}{\left(\alpha\beta + \gamma\delta + 2\alpha\delta\right)^2} \times \frac{\alpha\delta}{\left[(1-\alpha)(1-\beta)(1-\gamma)(1-\delta)\right]^{1/2}},$$

which is not trivial to integrate over the parameter space.

We conclude that the evaluation of (2) for Markov chains with arbitrarily large orders is still an open question. It might be helpful for practitioners to notice that a recursive formula is given in (Finesso et al., 1996) for another form of SC that relies on the method of mixtures (Rissanen, 1989).

Acknowledgements

C.D. Giurcăneanu is indebted to Prof. Bin Yu and Prof. Jorma Rissanen for many useful discussions on stochastic complexity. The work of C.D. Giurcăneanu was supported by Academy of Finland, project no. 113572 and 213462.

REFERENCES

Adamchik, V. (2002). Certain series associated with Catalan's constant. *Zeitschrift fuer Analysis und ihre Anwendungen,* 21:1-10.

Atteson, K. (1999). The asymptotic redundancy of Bayes rules for Markov chains. *IEEE Trans. on Information Theory*, 45(6):2104-2109.

Barron, A., J. Rissanen and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Trans. on Information Theory*, 44:2743-2760.

Bradley, D.M. (2001). Representations of Catalan's constant, Technical Report Dep. of Mathematics and Statistics, Univ. of Maine, USA, http://germain.umemat.maine.edu/faculty/bradley/papers/c1.ps.

Cover, T.M. and J.A. Thomas (1991). Elements of information theory. John Wiley & Sons.

Finesso, L., C.-C. Liu and P. Narayan (1996). The optimal error exponent for Markov order estimation. *IEEE Trans. on Information Theory*, 42(5):1488-1497.

Good, I.J. (1963). Quadratics in Markov-chain frequencies, and the binary chain of order 2. *J. Royal Statistical Society. Series B (Methodological)*, 25(2):383-391.

Hanson, A.J. and P.C.-W. Fu (2005). Aplications of MDL to selected families of models. In P.D. Grünwald, I.J. Myung, and M.A. Pitt, editors, *Advances in Minimum Description Length: theory and applications*, Chapter 5, pages 125-150. MIT Press.

Jacquet, P. and W. Szpankowski (2004). Markov types and minimax redundancy for Markov sources. *IEEE Trans. on Information Theory*, 50(7):1393-1402.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465-471, 1978.

Rissanen, J. (1989). Stochastic complexity in statistical inquiry. World Scientific.

Rissanen, J. (1996). Fisher information matrix and stochastic complexity. *IEEE Trans. on Information Theory*, 42(1):40-47.

Rissanen, J. (2000). MDL denoising. *IEEE Trans. on Information Theory*, 46(7):2537-2543.