Monte Carlo Estimation of Minimax Regret with an Application to MDL Model Selection

Teemu Roos
Department of Computer Science
University of Helsinki, Finland
teemu.roos@cs.helsinki.fi

***Abstract*— Minimum description length (MDL) model selection, in its modern NML formulation, involves a model complexity term which is equivalent to minimax/maximin regret. When the data are discrete-valued, the complexity term is a logarithm of a sum of maximized likelihoods over all possible data-sets. Because the sum has an exponential number of terms, its evaluation is in many cases intractable. In the continuous case, the sum is replaced by an integral for which a closed form is available in only a few cases. We present an approach based on Monte Carlo sampling, which works for all model classes, and gives strongly consistent estimators of the minimax regret. The estimates convergence almost surely to the correct value with increasing number of iterations. For the important class of Markov models, one of the presented estimators is particularly efficient: in empirical experiments, accuracy that is sufficient for model selection is usually achieved already on the first iteration, even for long sequences.**

I. Introduction

The stochastic complexity of a sequence under a given model class is a central concept in the Minimum Description Length (MDL) principle [1], [2], [3], [4]. Its interpretation as the length of the shortest achievable encoding makes it a yardstick for the comparison of different model classes. In recent formulations of MDL, stochastic complexity is defined using the so called normalized maximum likelihood (NML) model, originally introduced by Shtarkov [5] for data compression; for the role of NML in MDL model selection, see [3], [4], [6], [7], [8].

Since the introduction of the NML universal model in the context of MDL, there has been significant interest in the evaluation of NML stochastic complexity for different practically relevant model classes, both exactly and asymptotically. For discrete models, exact evaluation is often intractable since it involves a normalizing coefficient which is a sum over all possible data-sets. For continuous cases, the normalizing coefficient is an integral which can be solved in only a few cases. Under certain conditions on the model class, different versions of stochastic complexity (which include two-part, mixture, and NML forms) have the same asymptotic form — the so called Fisher information approximation, see e.g. [4], [6], [7] However, for small data-sets and for model classes that do not satisfy the necessary conditions, the asymptotic form is not accurate [9].

Exact and computationally tractable formulas are rare: results for multinomial models are given in [10], and for Bayesian networks with structural restrictions in [11], [12], [13]; more references can be found in [3] and [4]. Similarly to the present work, in the context of structural equation models, Preacher et al. [14] estimate the normalizing coefficient by sampling random data-sets from a uniform distribution using Markov chain Monte Carlo (MCMC) methods. Navarro [15] and Giurcăneanu and Rissanen [16] use Monte Carlo methods to estimate a Fisher information integral involved in the asymptotic expansion of NML stochastic complexity (see Sec. II). Kim et al. [17] use MCMC to quantify the complexity of connectionist models in terms of the number of different data patterns they can reproduce. Lafferty and Wasserman [18] estimate minimax *risk* (instead of the regret) using an MCMC version of the Blahut-Arimoto algorithm.

The contribution of this paper is to present estimators of the normalizing coefficient in the NML model, which is equivalent to the minimax and maximin regret (see Sec. II below). The estimators are Monte Carlo estimators based on sampling of random data-sets from suitable distributions. They are shown to be (strongly) consistent in the sense that they can be made arbitrarily accurate by sampling more and more random data-sets. Our work differs from earlier work by giving provably consistent recipes for estimating the NML normalizing coefficient for general model classes directly, and not through asymptotic expansions.

In Sec. II we review some basic properties of the well-known NML model. Three general Monte Carlo methods applicable to the estimation of complex sums and integrals are briefly introduced in Sec. III. In order to demonstrate the applicability of our methods, we apply them to the class of Markov models (Sec. IV), and present some experimental results (Sec. V).

II. Normalized Maximum Likelihood (NML)

Let $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, $n \in \mathbb{N}$ be a sequence. We consider a model class $\mathcal{M} = \{p(\cdot\,;\,\theta) \,:\, \theta \in \Theta\}$, i.e., a set of probability mass or density functions over sequences in $\mathcal{X}^n$. We denote the maximum likelihood parameters by $\hat{\theta}(x^n)$. The ML parameters do not have to be unique – in fact the model does not even have to be parametric – since we will only use the maximized likelihood $p(x^n\,;\,\hat{\theta}(x^n))$.

The celebrated *normalized maximum likelihood* (NML) uni-

versal model [5], [6] is given by

$$p_{\mathrm{NML}}(x^n) = \frac{p(x^n\ ;\ \hat{\theta}(x^n))}{C_n}\ ,\ C_n = \sum_{x^n \in \mathcal{X}^n} p(x^n\ ;\ \hat{\theta}(x^n))\ ,$$

where the sum is over all possible sequences of length n, and C_n is a normalizing constant ensuring that the result is indeed a probability mass function. In the continuous case, the sum is replaced by the corresponding integral. The NML model is the unique minimax optimal universal model in the sense that it minimizes the worst-case regret

$$\max_{x^n} \mathcal{R}(p, x^n) = \max_{x^n} \log \frac{p(x^n\ ;\ \hat{\theta}(x^n))}{p(x^n)}\ .$$

In fact, it directly follows from the definition that the regret of NML is a constant dependent only on the model class and the sample size n:

$$\mathcal{R}(p_{\mathrm{NML}}, x^n) = \log C_n = \min_p \max_{x^n} \mathcal{R}(p, x^n)\ .$$

For some model classes, the normalizing coefficient is finite only if the range of the data is restricted, see e.g. [6], [19], [20]. The logarithm of the normalizing coefficient, $\log C_n$, is equal to both the minimax and maximin regret under log-loss, see e.g. [19], [21], which makes the quantity interesting in its own right.

The practical problem arising in applications of the NML universal model is the evaluation of the normalizing constant. For continuous models the integral can be solved in closed form for only a few specific models. For discrete models, the time complexity of the naive solution, i.e., summing over all possible data-sets, grows exponentially in the sample size, and quickly becomes intractable. Even the second-most naive solution, summing over equivalence classes (or *types*), sharing the same likelihood value, is usually intractable even though often polynomial in n.

On the other hand, we have the usual Fisher information approximation [6]

$$\log C_n = \frac{d}{2} \log \frac{n}{2\pi} + \log \int_\Theta \sqrt{\det I(\theta)}\, d\theta + o(1)\ , \quad (1)$$

where d is the dimension of the parameter space. It is also non-trivial to apply due to the integral involving the Fisher information $I(\theta)$. Using only the leading term (without 2π), i.e., the BIC criterion [22], gives a rough approximation. Even if the Fisher information approximation can be used, there are practical circumstances where it gives a very poor approximation [9]. Consequently, analytic approximations perform worse in model selection tasks than more refined approximations, or ideally, the exact solution, see e.g. [3, Chap. 9].

III. Monte Carlo Approximations

Monte Carlo methods provide a family of estimators of integrals or sums (or averages) that are either analytically or computationally hard to evaluate, see e.g. [23]. We give a very brief review of the topic, with three examples of estimators.

The simplest way to estimate a sum $S = \sum_{z \in \mathcal{Z}} f(z)$ is to draw a sample $z^{(1)}, \ldots, z^{(m)}$, $m \in \mathbb{N}$, uniformly from the domain $\mathcal{Z}$, and evaluate the sample average of $|\mathcal{Z}|\, f(z^{(t)}), t \in \{1, \ldots, m\}$, where $|\mathcal{Z}|$ denotes the number of elements[1] in $\mathcal{Z}$. This obviously requires that the domain $\mathcal{Z}$ is finite. The expected value of $|\mathcal{Z}|\, f(z)$ is given by

$$\mathbb{E}_{z \sim \mathrm{Uni}(\mathcal{Z})}[|\mathcal{Z}|\, f(z)] = \sum_{z \in \mathcal{Z}} \frac{1}{|\mathcal{Z}|} |\mathcal{Z}|\, f(z) = \sum_{z \in \mathcal{Z}} f(z) = S\ ,$$

where the notation $z \sim \mathrm{Uni}(\mathcal{Z})$ implies that z follows the uniform distribution on $\mathcal{Z}$.

The law of large numbers now guarantees almost sure convergence of the sample average to the mean:

$$\frac{1}{m} \sum_{t=1}^{m} |\mathcal{Z}|\, f(z^{(t)}) \xrightarrow{a.s.} S\ , \quad z^{(t)} \sim \mathrm{Uni}(\mathcal{Z}) \quad (2)$$

as the number of samples m goes to infinity. In the case of the NML normalization term, let $\mathcal{Z} = \mathcal{X}^n$ and $f(x^n) = p(x^n\ ;\ \hat{\theta}(x^n))$, so that we have $S = C_n$.

The rate of convergence of the simple estimator (2) can be very slow if the value $f(z)$ depends strongly on z since this increases the variance

$$\sum_{z \in \mathcal{Z}} \frac{1}{|\mathcal{Z}|} (|\mathcal{Z}|\, f(z) - S)^2 = |\mathcal{Z}| \sum_{z \in \mathcal{Z}} f(z)^2 - S^2\ .$$

In the discrete case where $f(z)$ are probability values, we have $\sum_{z \in \mathcal{Z}} f(z)^2 \leq \sum_{z \in \mathcal{Z}} f(z) = S$. Hence, the variance is finite whenever the sum S is finite.

A generalization of the simple estimator is obtained by replacing the uniform sampling distribution by an arbitrary distribution q on $\mathcal{Z}$, and using the *importance-sampling estimator*:

$$\frac{1}{m} \sum_{t=1}^{m} \frac{f(z^{(t)})}{q(z^{(t)})}\ , \quad z^{(t)} \sim q\ . \quad (3)$$

The simple estimator is a special case with $q(z) = 1/|\mathcal{Z}|$ for all $z \in \mathcal{Z}$. Assuming that $q(z) > 0$ whenever $f(z) > 0$, the expected value of each term is equal to the sum:

$$\mathbb{E}_{z \sim q}\left[\frac{f(z)}{q(z)}\right] = \sum_{z \in \mathcal{Z}} q(z) \frac{f(z)}{q(z)} = \sum_{z \in \mathcal{Z}} f(z) = S\ ,$$

which is again sufficient to guarantee almost sure convergence to the desired sum.

If the shape of q is similar to the shape of f in the sense that the variance

$$\sum_{z \in \mathcal{Z}} q(z) \left(\frac{f(z)}{q(z)} - S\right)^2 = \sum_{z \in \mathcal{Z}} \frac{f(z)^2}{q(z)} - S^2\ ,$$

is small, then the rate of convergence of the importance-sampling estimator is fast. Under the assumption $f(z) > 0 \Rightarrow q(z) > 0$, the variance is guaranteed to be finite for all *finite* domains $\mathcal{Z}$. Finding a distribution q which is as close as possible to the NML model has to be done on a case-by-case basis. Typically, good candidates can be found among universal mixture models, see the following sections.

[1] When estimating an integral, $|\mathcal{Z}|$ is the *volume* of $\mathcal{Z}$, which has to be finite so that the uniform distribution is well-defined.

Ideally, if we could choose $q(z) = f(z)/S$ for all $z \in \mathcal{Z}$, then the variance of the importance-sampling estimator becomes zero, and a single term in the sum (3) gives S exactly. Unfortunately, this cannot be done when estimating the normalizing constant C_n since it requires that (i) we can draw i.i.d. samples from the NML distribution p_{NML}, and (ii) the probability $p_{\mathrm{NML}}(z)$ can be evaluated. Obviously, if this was possible, there would be nothing to estimate.

However, as far as point (i) above is concerned, it is possible to draw *non*-i.i.d. samples from distribution f/S without knowing the normalizing constant by using MCMC techniques, see [23], [24]. For instance, the Metropolis-Hastings algorithm starts with an arbitrary value $z^{(0)}$, and draws a potential new value z' from some *proposal distribution* q, which is accepted as the new value in a randomized fashion:

$$z^{(t+1)} = \begin{cases} z', & \text{with prob. } \dfrac{p(z')q(z^{(t)})}{p(z^{(t)})q(z')}, \\ z^{(t)}, & \text{otherwise.} \end{cases}$$

Under mild conditions on the proposal distribution, the chain is ergodic and has as its stationary distribution p; in the long run, averages computed from a sample drawn from the chain converge to the population average under p. In practice, certain design issues (most importantly, the choice of the proposal distribution) must be dealt with carefully in order to avoid convergence problems.

One way to estimate the sum S based on an MCMC sample from the normalized distribution f/S is the *harmonic mean estimator* [25]:

$$\frac{1}{m}\sum_{t=1}^{m}\frac{1}{f(z^{(t)})} \ , \quad z^{(t)} \sim \frac{f}{S} \ . \tag{4}$$

We easily get the expectation

$$\mathbb{E}_{z\sim f/S}\left[\frac{1}{f(z)}\right] = \sum_{z\in\mathcal{Z}}\frac{f(z)}{S}\frac{1}{f(z)} = \sum_{z\in\mathcal{Z}}\frac{1}{S} = |\mathcal{Z}|\,S^{-1} \ ,$$

which is seen to be finite for all finite domains. The ergodic theorem (see e.g. [26]) then guarantees almost sure convergence of the estimator (4) to its expectation. This implies that that by dividing (4) by $|\mathcal{Z}|$ and taking the inverse gives a (strongly) consistent estimator of the normalizing constant:

$$\left(\frac{1}{m|\mathcal{Z}|}\sum_{t=1}^{m}\frac{1}{f(z^{(t)})}\right)^{-1} \xrightarrow{a.s.} S \ . \tag{5}$$

The variance of each term of the harmonic mean estimator (4) is given by

$$\sum_{z\in\mathcal{Z}}\frac{f(z)}{S}\left(\frac{1}{f(z)} - \frac{|\mathcal{Z}|}{S}\right)^2 = \frac{1}{S}\left(\sum_{z\in\mathcal{Z}}\frac{1}{f(z)} - \frac{|\mathcal{Z}|^2}{S}\right), \tag{6}$$

where the sum is taken over z with $f(z) > 0$. If $\mathcal{Z}$ is finite, the variance is finite. Note, however, that the variance of (5) may be significantly larger than (6) since the former involves the inverse, and hence, (6) is not comparable to the variances of the other two estimators. In addition to higher variance of the inverse, the the convergence of the harmonic mean estimator may be slowed down due to the fact that the MCMC sample, while ergodic, is not i.i.d.

IV. Markov Models

The class of Markov models is widely studied in the context of universal prediction and compression, see e.g. [27], [28], [29], [30]. Under a kth order Markov model, the joint probability over sequences $x^n \in \mathcal{X}^n$ is given by

$$p_k(x^n) = p^0(x^k)\prod_{t=k+1}^{n} p(x_t \mid x_{t-k}^{t-1}) \ , \tag{7}$$

where the first k observations follow the initial distribution p^0, and the distribution of $x_{k+1}, \ldots, x_n$ is given by the *transition probabilities* $p(x_t \mid x_{t-k}^{t-1})$. We study the discrete case, where each observation x_t takes on values in a finite set $\mathcal{X}$, and both the initial and the transition probabilities are multinomial distributions.

As is well known (see e.g. [29]), in the multinomial case the maximized likelihoods are given by empirical frequencies:

$$p^0(s\ ;\ \hat{\theta}(x^n)) = I_{\{x_1^k = s\}} \ ,$$

$$p(u \mid s\ ;\ \hat{\theta}(x^n)) = \frac{\sum_{t=k+1}^{n} I_{\{x_{t-k}^{t-1}=s, x_t=u\}}}{\sum_{t=k+1}^{n} I_{\{x_{t-k}^{t-1}=s\}}} = \frac{N_{su}(x^n)}{N_s(x^{n-1})} \ ,$$

for $u \in \mathcal{X}$ and $s \in \mathcal{X}^k$, where $I_{\{\cdot\}}$ is the indicator function, and $N_s(y)$ denotes the number of occurrences of subsequence s in sequence y; the transition probabilities are undefined whenever the sum in the denominator equals zero.

For fixed x^n, the maximized likelihood under a kth order Markov model is given by

$$\begin{aligned} p(x^n\ ;\ \hat{\theta}(x^n)) &= \prod_{t=k+1}^{n}\frac{N_{x_{t-k}^t}(x^n)}{N_{x_{t-k}^{t-1}}(x^{n-1})} \\ &= \prod_{a^{k+1}\in\mathcal{X}^{k+1}}\left(\frac{N_{a^{k+1}}(x^n)}{N_{a^k}(x^{n-1})}\right)^{N_{a^{k+1}}(x^n)} \ . \end{aligned} \tag{8}$$

In the latter product, terms with $N_{a^k}(x^{n-1}) = 0$ are omitted in line with the convention $(\frac{0}{0})^0 = 1$.

The time complexity of evaluating (8) is $\mathcal{O}(kn\log n)$: we can scan the sequence from left to right and keep a list of states that occur at least once, in a balanced binary tree (such as AVL or red-black, see e.g. [31]) where search and addition operations can be performed in logarithmic time. In each search operation we need $\mathcal{O}(k)$ operations to compare two states, and hence, constructing the tree takes time $\mathcal{O}(kn\log n)$. By also storing the counts N_{a^k} and $N_{a^{k+1}}$ together with the states a^k we can evaluate (8) in the end by visiting all the nodes in the tree. Since the number of states with non-zero count is bounded by n, the complexity of this phase is $\mathcal{O}(n)$.[2]

[2]Assuming that the size of the alphabet $K = |\mathcal{X}|$ is at most of order $k\log n$, the effect of the size of the alphabet to the total time complexity is negligible: in the first phase, the symbols following each state can be kept in a balanced binary tree in time $\mathcal{O}(n\log K)$, and in the second phase, visiting each state–symbol pair takes time $\mathcal{O}(nK)$. For moderate alphabet sizes, this is dominated by $\mathcal{O}(kn\log n)$.

On the other hand, if both the alphabet size $K = |\mathcal{X}|$ and the model order k are small, it may be faster to simply tabulate the state–symbol pairs. The space and time complexity of the table initialization and traversal is then $\mathcal{O}(K^{k+1})$, while processing the actual sequence can be done in linear time, $\mathcal{O}(nk)$ (for each symbol the table index depends on the previous k symbols).

The normalizing constant of the NML distribution for the kth order Markov model, $C_{k,n}$ is obtained by normalizing (8) over $\mathcal{X}^n$:

$$C_{k,n} = \sum_{y^n \in \mathcal{X}^n} \prod_{a^{k+1} \in \mathcal{X}^{k+1}} \left(\frac{N_{a^{k+1}}(y^n)}{N_{a^k}(y^{n-1})} \right)^{N_{a^{k+1}}(y^n)} .$$

In principle, the normalizing constant in the NML distribution can be computed by summing over all *types* (possible sets of counts) of sequences, which are polynomial in number with respect to n, but exponential in both the size of the alphabet $|\mathcal{X}|$ and k. This approach leads to the intricate combinatorial problem of computing the number of sequences of each type. Jacquet and Szpankowski give asymptotic expressions for the normalizing constant by this method [27]. Unfortunately, their asymptotic expressions are given in closed form only for orders $k = 0$ and $k = 1$.

V. Experiments

We estimate the normalizing constants $C_{k,n}$ for Markov models using Monte Carlo methods, as discussed in Sec. III. In the importance-sampling estimator (3), we use as the sampling distribution a mixture model where all the parameters were drawn from Dirichlet distributions with parameters $(1/2, 1/2, \ldots, 1/2)$. This yields the well-known Krichevsky-Trofimov[3] universal model [32]. The resulting estimator was observed to be superior compared to both the simple estimator (2) using uniformly random sequences, and the harmonic mean estimator (4) where the same Dirichlet-mixture model was used as the proposal distribution of the MCMC sampler. For this reason, we focus here on the importance-sampling estimator.

Figures 1–3 illustrate the behavior of the importance-sampling estimates of $\log C_{k,n}$ for various combinations of the sequence length n, model order k, and the number of iterations. In all figures, the size of the alphabet was $K = |\mathcal{X}| = 4$, and the experiment was repeated 30 times in order to get an idea of the variability of the estimates. In Fig. 1, it can be seen that the 30 estimates of $\log C_{5,15625}$, each based on a single iteration (a single sampled sequence), vary within the range $[5450, 5650]$, but the range is much narrower for estimates based on more iterations.

Most importantly, even the accuracy of the estimates based on a single iteration is sufficient for model selection, where the *relative* code-lengths of models with different order is important: Fig. 2 shows estimates of $\log C_{k,15625}$ for $k \in \{0, 1, \ldots, 15\}$. These are based on a single iteration, but still the variation over 30 repetitions is negligible compared to the relative differences between different order models.

[3]We found that the Krichevsky-Trofimov model gives significantly faster convergence to the correct value than, for instance, the Laplace predictor, i.e., a mixture with uniform prior. This is natural: for the multinomial model, the K-T model is asymptotically equivalent to NML except at the boundaries of the parameter simplex, see e.g. [21].

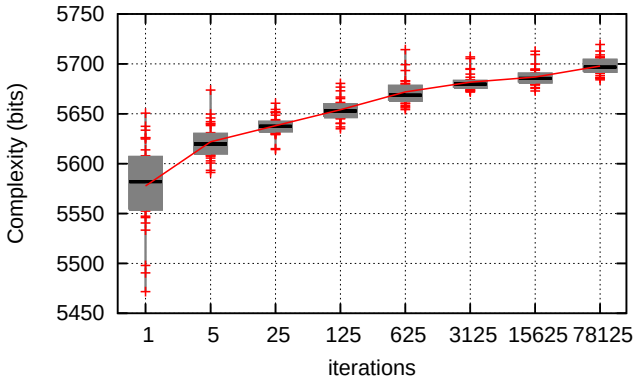

Fig. 1. Convergence of the importance-sampling estimator of $\log C_{k,n}$ with $k = 5, n = 15625$. The gray box shows the first to third quartile range, the black bar shows the *median*, and the red line shows the *mean* over all 30 repetitions. Values outside the first to third median range are marked by red crosses. Note the log-scale for the number of iterations.

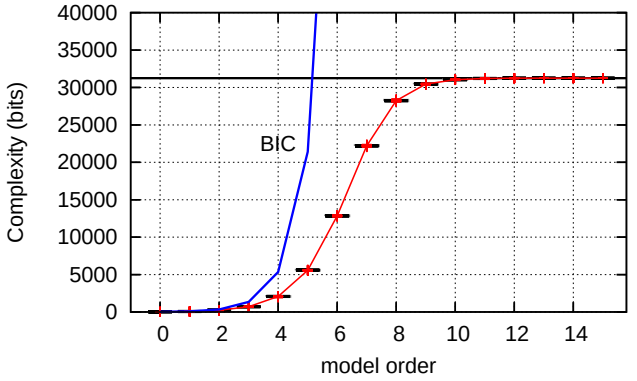

Fig. 2. Estimates of $\log C_{k,n}$ for $n = 15625$ and $k \in \{0, 1, \ldots, 15\}$ based on a *single iteration*. The style is the same as in Fig. 1. The BIC complexity term $\frac{1}{2}K^k(K-1)\log n$ is plotted with a blue line, and the black horizontal bar shows the upper bound $n \log K$.

Another observation in Fig. 2 is the "saturation" of the NML complexity term $\log C_{k,n}$ as the model order k increases. This can be explained by noting that the sum of maximized likelihoods over all data-sets cannot be greater than the number of data-sets, so that we have $\log C_n \leq \log K^n = n \log K$. This gives a uniform upper bound on the NML complexity term which holds for all (discrete) models. In contrast, the BIC complexity term $\frac{1}{2}K^k(K-1)\log n$ overshoots both the NML complexity term and the upper bound $n \log K$ by a large margin for large k. In fact the inconsistency of NML (and Bayesian) model selection [29], which occurs when the data is uniformly random, can be traced to the fact that as $k \to n$, the maximized likelihood approaches one uniformly for all sequences, and consequently, the NML distribution approaches the uniform distribution. The over-penalization for complexity by BIC, seen in Fig. 2, makes BIC consistent in all cases,

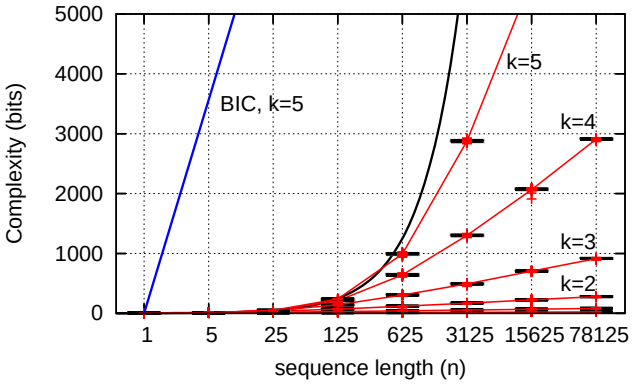

Fig. 3. Estimates of $\log C_{k,n}$ for $k \in \{0, 1, 2, 3, 4, 5\}$ as a function of the sequence length n. The blue line shows the BIC complexity term for the case $k = 5$, and the black curve shows the upper bound $n \log K$.

including the uniformly random one.

Figure 3 shows the growth of $\log C_{k,n}$ as a function of sequence length n for model orders $k \in \{0, 1, 2, 3, 4, 5\}$. Again, even though the estimates are based on a single iteration, the variation over 30 repetitions is negligible compared to the relative differences between different order models. Asymptotically, as n grows, the BIC complexity term and $\log C_{k,n}$ differ only by the constant involving the Fisher information, recall Eq. (1), which can be seen in the figure from the fact that the slope of the $k = 5$ curve approaches the (constant) slope of the "BIC, $k = 5$" curve; note that while both grow logarithmically in n, the log-scale on the vertical axis makes the curves appear asymptotically linear. For small n, the NML complexity terms are tightly bounded by $n \log K$, while BIC clearly overshoots both.

Acknowledgments

The author thanks Mikko Koivisto, Tommi Mononen, and Tomi Silander for useful discussions. This research has been funded in part by the Finnish Funding Agency for Technology and Innovation under the project KUKOT, and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

[1] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 445–471, 1978.

[2] ——, "Stochastic complexity," *Journal of the Royal Statistical Society*, vol. 49, no. 3, pp. 223–239 and 252–265, 1987.

[3] ——, *Information and Complexity in Statistical Modeling*. Springer, 2007.

[4] P. Grünwald, *The Minimum Description Length Principle*. MIT Press, 2007.

[5] Y. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3–17, 1987.

[6] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[7] ——, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1712–1717, July 2001.

[8] J. Myung, D. Navarro, and M. Pitt, "Model selection by normalized maximum likelihood," *Journal of Mathematical Psychology*, vol. 50, pp. 167–179, 2006.

[9] D. Navarro, "A note on the applied use of MDL approximations," *Neural Computation*, vol. 16, no. 9, pp. 1763–1768, 2004.

[10] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, vol. 103, no. 6, pp. 227–233, 2007.

[11] P. Kontkanen, H. Wettig, and P. Myllymäki, "NML computation algorithms for tree-structured multinomial Bayesian networks," *EURASIP Journal on Bioinformatics and Systems Biology*, 2007, article ID 90947.

[12] H. Wettig, P. Kontkanen, and P. Myllymäki, "Calculating the normalized maximum likelihood distribution for Bayesian forests," in *Proc. IADIS International Conference on Intelligent Systems and Agents*, Lisbon, Portugal, July 2007.

[13] T. Mononen and P. Myllymäki, "Fast NML computation for naive Bayes models," in *Proc. 10th International Conference on Discovery Science*, V. Corruble, M. Takeda, and E. Suzuki, Eds. Springer, 2007, pp. 151–160.

[14] K. Preacher, L. Cai, and R. MacCallum, "Alternatives to traditional model comparison strategies for covariance structure models," in *Modeling Contextual Effects in Longitudinal Studies*. Mahwah, NJ: Lawrence Erlbaum Associates, 2007, pp. 33–62.

[15] D. Navarro, "Calculating geometric complexity for three categorization models: PRT, GCM and GCM-gamma," Ohio State University, Tech. Rep., 2004.

[16] D. Giurcăneanu and J. Rissanen, "Estimation of AR and ARMA models by stochastic complexity," in *Time Series and Related Topics: In Memory of Ching-Zong Wei*, ser. IMS Lecture Notes–Monograph Series. Institute of Mathematical Statistics, 2006, vol. 52, pp. 48–59.

[17] W. Kim, D. Navarro, M. Pitt, and I. Myung, "An MCMC-based method of comparing connectionist models in cognitive science," in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds., vol. 16. Cambridge, MA: MIT Press, 2004, pp. 937–944.

[18] J. Lafferty and L. Wasserman, "Iterative Markov chain Monte Carlo computation of reference priors and minimax risk," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 2001, pp. 293–300.

[19] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[20] S. de Rooij and P. Grünwald, "An empirical study of minimum description length model selection with infinite parametric complexity," *Journal of Mathematical Psychology*, vol. 50, no. 2, pp. 180–192, 2006.

[21] Q. Xie and A. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, March 2000.

[22] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[23] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2004.

[24] B. Carlin and S. Chib, "Bayesian model choice via Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 3, pp. 473–484, 1995.

[25] A. Raftery, M. Newton, J. Satagopan, and P. Krivitsky, "Estimating the integrated likelihood via posterior simulation using the harmonic mean identity," in *Bayesian Statistics*, vol. 8, 2007, pp. 1–45.

[26] R. Gray, *Probability, Random Processes, and Ergodic Properties*. Springer, 1988.

[27] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for Markov sources," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1393–1402, 2004.

[28] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Transactions on Information Theory*, vol. 35, no. 5, pp. 1014–1019, 1989.

[29] I. Csiszár and P. Shields, "The consistency of the BIC Markov order estimator," *Annals of Statistics*, vol. 28, no. 6, pp. 1601–1619, 2000.

[30] A. Dhulipala and A. Orlitsky, "Universal compression of Markov and related sources over arbitrary alphabets," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4182–4190, 2006.

[31] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms*. MIT Press, 1990.

[32] R. Krichevsky and V. Trofimov, "The performance of universal coding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.