Journal of Statistical Planning and Inference 39 (1994) 353–372
North-Holland

Predictive stochastic complexity and model estimation for finite-state processes*

Marcelo J. Weinberger**

IBM-Almaden Research Center, San Jose, Ca 95120-6099, USA

Meir Feder

Department of Electrical Engineering-Systems, Tel Aviv University, Tel Aviv 69978, Israel

Received 30 November 1992; revised manuscript received 27 April 1993

Abstract

It is shown that the predictive and nonpredictive stochastic complexities relative to the class of finite-state models are asymptotically equivalent in a probabilistic sense. To this end, a universal, sequential, noiseless coding scheme attaining the minimum description length (MDL) of the data with respect to this class is presented and investigated. It relies on an MDL-based estimator of the model structure, which is proved to be strongly consistent. An interpretation of this result is that a process 'close' to every process in the class, regardless of the model structure, can be constructed. This universal process can be employed in the solution of sequential decision problems like coding, prediction, and gambling, in an asymptotically optimal manner.

AMS Subject Classification: Primary 62B10, 62A99, 94A29, 62F12; secondary 60J10.

Key words: Stochastic complexity; finite-state processes; model estimation; universal coding.

1. Introduction

In this work, we consider the parametrization of discrete random processes in terms of finite-state (FS) models. The broad class of FS models, which includes Markov models as a special case, is flexible and rich, and so it has extensively been used for modeling the data in various applications; one such application, that will draw special attention in this paper, is data compression (Davisson, 1983; Rissanen, 1986b; Weinberger et al., 1992). In many cases it seems plausible that an observed data sequence over a discrete alphabet can be 'explained' as a sample of some FS process, and for this

Correspondence to: M.J. Weinberger, IBM-Almaden Research Center, San Jose, Ca 95120-6099. USA.

*This paper has been partially presented at the 1993 IEEE ISIT, San Antonio, Texas, January 1993.

**This research was started while this author was with the Department of Electrical Engineering, Technion-Israel Institute of Technology.

SSDI 0378-3758(93)E0046-J

matter it is interesting to estimate the model in the class that fits best this data. This model estimation problem is one of the concerns of this paper. Roughly speaking (a formal definition of this class is given in Section 2), it involves the estimation of a model structure (and, particularly, a model order), namely an FS machine, and of the model parameters.

Estimating the model parameters when the supporting FS machine F is known is straightforward. Classical statistical methods can be used; the maximum-likelihood (ML) estimates of the parameters are the empirical probabilities, and the estimation error in this case can be evaluated by, say, the Cramer–Rao lower bound. However, the estimation of the structure leads to conceptual problems. The ML criterion for estimating it is clearly unacceptable, since the likelihood of the data can only increase as the number of states in F increases. Similar phenomena of model order estimation have been observed and considered in many modeling problems, e.g., Gaussian AR and ARMA models. A common technique for model order estimation is to optimize an expression composed of the likelihood and a 'penalty' term for the number of parameters, to compensate over-parametrization. Now, some of the proposed criteria and penalty terms are ad hoc, and some, like Akaike's criterion (Akaike, 1974) are asymptotically inconsistent. Other approaches for estimating F are based on decision theory (e.g. Whittle, 1952; Anderson, 1963); one such recent decision theoretic approach (Merhav et al., 1989; Ziv and Merhav, 1992) might be, again, inconsistent.

A particularly interesting approach for estimating the model structure is the minimum description length (MDL) principle (Rissanen, 1983). In its basic form it states that the chosen model should minimize the number of bits needed to describe both the model and the data in terms of this model. It can be shown that, asymptotically, this amounts to minimizing an expression composed of the normalized likelihood and a penalty term of the form $0.5Kn^{-1}\log n$, where K is the number of free parameters in the model, and n is the data length (a formally similar criterion was independently proposed by Schwartz (1978)). Although the idea behind this principle is fitting probabilistic models to individual sequences, rather than estimating an underlying (unknown) distribution, it can be applied in a probabilistic environment thus raising the issue of consistency. Moreover, one cannot argue that a particular way of describing the data is optimal, except in a probabilistic setting, where a lower bound on the expected code length is available (Rissanen, 1984). Now, while the consistency of the MDL estimator was shown for AR (Hannan and Quinn, 1973) and ARMA models, and the rate at which the corresponding error probability tends to zero was investigated, no such results are available for the class of FS models. One of the results of this paper is a strong consistency proof for an MDL-based estimator and further properties of its error probability for this class of models. We note that, recently, Kieffer (1993) proved strong consistency results for an MDL-based model class selection rule which apply to a broad class of models, including FS ones, provided a 'nesting' property is satisfied (i.e. each class is contained in the subsequent classes).

This rule uses a penalty term *larger* than the one given by the MDL. Furthermore, FS models satisfy the nesting requirements only if we consider all models with the same cardinality as a single model class. Therefore, unlike our results, this procedure can be used to estimate only the order of the model, rather than its full structure. We also note that the strong consistency of an estimator for the structure of a Markov chain with a bounded number of states is reported in Rudich (1985).

However, the scope of this paper is broader than the consistency of the MDL estimator for the class of FS models. The MDL principle is closely related to the notion of stochastic complexity of a data string, defined as the shortest description of this string in terms of the models in a given class (Rissanen, 1983; Rissanen, 1986a). A nonpredictive notion of stochastic complexity results from a description using a 'two-part' code, in which the parameters are explicitly encoded, and then followed by a codeword describing the data in terms of these parameters. This complexity is the minimal value attained by the goal function used in the definition of the MDL criterion. Rissanen (1986a) has introduced additional notions of stochastic complexity, namely semi-predictive and (fully) predictive complexities, according to increasing degrees of sequentiality in the scheme used to encode the data. In the semi-predictive complexity, the model parameters used to encode the data are determined sequentially, but a prescan step is still needed since the supporting machine F is determined from the whole data. In the fully predictive complexity both the machine and the parameters are determined sequentially so that the number of parameters in the fitted models is penalized without any explicitly added terms, which have only an asymptotic meaning (see Section 2). This property makes the predictive approach especially suited for modeling finite sequences. It was conjectured by Rissanen (1986a) that all these definitions of the stochastic complexity are asymptotically equivalent, up to terms of $O(n^{-1})$, at least in a probabilistic sense. This property is essential in order to consider a unique concept of 'minimum description length'. One step in the corroboration of this conjecture was taken by Rissanen (1986b), who proved that for the class of FS models the semi-predictive and nonpredictive stochastic complexities are equivalent. The main contribution of this paper is in proving this conjecture in full for the FS class of models in a probabilistic setting, where the data is assumed to be a sample of an FS process. In this setting, it turns out that all the definitions of the stochastic complexity tend, up to $O(n^{-1})$ terms, to the optimal expected code length achievable by any algorithm used to encode the data.

The proof of the main result, in Section 3, is based on the observation that any model estimator obtained from the so far processed sequence x_1^i, whose probability of error in determining the underlying FS model tends to zero fast enough, can be used for sequential coding of x_{i+1}, and the resulting expected code length will be minimum (i.e. the stochastic complexity) up to an $O(n^{-1})$ term. Now, since an MDL-based estimator is used in this fully sequential scheme, the consistency of this estimator is shown, and further properties of its error probability are analyzed, as a step in the course of proving this main result.

In many applications sequentiality is, of course, a desirable property of a data compression scheme. But this is not the only appealing property of the predictive nature of this approach to stochastic complexity. The main result of this paper can be interpreted in an alternative way: it shows the existence of a *universal* process that is 'close' to any process generated by an FS model (and even specifies it). Specifically, the predictive stochastic complexity corresponds to a fully sequential universal scheme for coding the data. As such, it induces a probability distribution for the next outcome given the past string. Suppose we define a stochastic process such that the probability it assigns to each string is the product, along the time indices, of the conditional probabilities induced by the coding scheme used in the definition of the predictive complexity. Our results imply that the maximal log-likelihood assigned to that string by any FS model is only better by an $O(\log n)$ term (on the average and a.s.) than the log-likelihood induced by the coding process of the predictive complexity. Since this term represents the minimal possible deviation (given by the lower bound (Rissanen, 1984)), this can be considered as a universal process in the class of FS models. While the idea of universal probability goes back to Solomonoff (1964) and Kolmogorov (1965), the specific notion and its relation to the stochastic complexity has been presented by Weinberger et al. (1993), who provided such a universal probability for a more restricted class of tree sources. Such universal processes can be used not only for universal coding and prediction (Rissanen, 1984; Weinberger et al., 1993), but also for other *sequential decision problems* like gambling (Feder, 1991) as well. These ideas are further discussed in Section 4.

2. Predictive stochastic complexity for finite-state models

A unifilar FS probabilistic source $\mathscr{X}$ over a discrete alphabet A of α letters, is defined by an FS machine F on a state space S of finite cardinality k, and an $(\alpha-1)k$-vector of parameters θ given by the k probability measures $p(a|z)$, $a\in A$, $z\in S$. The transitions between states of S are determined by a 'next-state' function f that maps $S\times A$ into S, together with a given initial state z_0. The probability that a string $x_1^n = x_1\, x_2 \cdots x_n$, $x_i\in A$, $1\leqslant i\leqslant n$, be emitted by $\mathscr{X}$ is given by

$$P(x_1^n;\mathscr{X}) = \prod_{i=1}^{n} p(x_i \mid z_{i-1}), \quad z_i \triangleq f(z_{i-1}, x_i), \quad 1\leqslant i\leqslant n.$$

In this work, $P(X_1^n;\mathscr{X})$ is called an FS process, and the pair $\mathscr{X}=(F,\theta)$ is referred to as the model. An FS machine can be illustrated as a directed graph with k vertices corresponding to the states and with edges corresponding to the allowable state transitions dictated by f. Thus, we assume α outgoing edges from each vertex. We further assume that the model is irreducible and aperiodic (i.e. the process is ergodic) (Cox and Miller, 1967, p. 101). The machine F is also denoted by the quadruple (S,k,f,z_0), where $k=|S|$. The per-symbol entropy of $P(X_1^n;\mathscr{X})$ is denoted by $H_n(\mathscr{X})$.

In this section we review the various definitions of the stochastic complexity for FS models, and describe the model estimation problem which is inherent in the definition of the predictive complexity. Consider an FS model $\mathscr{X}=(F,\theta)$ that generates a process $P(X_1^n;\mathscr{X})$, with $F=(S,k,f,z_0)$. Hereafter, we assume that F is a minimal machine, in the sense that no model with fewer states generates the same process. Since the definition of the predictive complexity involves algorithms that estimate sequentially an FS model from the data, it is necessary to establish the uniqueness of F. We begin with a preliminary discussion of this issue.

Specifically, we show in Lemma 1 and its Corollary below, the uniqueness (up to a permutation) of a minimal machine that supports an FS model. Note that this claim does not hold for nonunifilar sources (see Blackwell and Koopmans, 1957). To state Lemma 1 we define, following Feder et al. (1992) a refinement F' of F as another machine (S',k',f',z_0') satisfying $z_i=g(z_i')$ for every emitted sequence x_1^i and every $i\geqslant 0$, where $\{z_i'\}$ is the sequence of states defined by f' and z_0', and $g(\cdot)$ is some function $S'\to S$.

Lemma 1. *Given two FS models defining the same process, let F and F' denote the corresponding finite-state machines and assume that F is minimal. Then F' is a refinement of F.*

Proof. Call two states in a model equivalent if the two models obtained by starting emission at these states define the same process. Two distinct states s and z in the supporting set S of F cannot be equivalent, for otherwise deleting state s from S and redirecting all its incident edges to z, we obtain a model that defines the same process, thus contradicting the minimality of F. Consider two paths over F' leading from its initial state z_0' to the same state z'. The corresponding paths over F starting at its initial state z_0, which exist since the two models define the same process, lead to the same final state defined to be $g(z')$, for otherwise S would contain distinct equivalent states. In particular, if z_0' is reachable from itself, we have $g(z_0')=z_0$, for otherwise S would contain two equivalent states $g(z_0')$ and z_0. If it is not reachable, define $g(z_0')\triangleq z_0$. Then, for any emitted sequence x_1^i, $i\geqslant 0$, by the definition of $g(\cdot)$ we have $z_i=g(z_i')$, where $\{z_i\}$ and $\{z_i'\}$ are the sequences of states defined by F and F', respectively. Therefore, $g(\cdot)$ defines a refinement F' of F. □

Clearly, if F is a minimal machine then every state in its supporting set S is reachable from the initial state z_0 (except, possibly z_0 itself). Thus, if the supporting set of F' has the same cardinality k, then the mapping $g(\cdot)$ in the proof of Lemma 1 is one-to-one. That proof also implies a one-to-one correspondence between the edges of F and F'. Hence, we obtain the following corollary.

Corollary. *The minimal machine generating an FS process is unique up to a permutation.*

We proceed now and review the various notions of stochastic complexity for FS models. The corresponding definitions involve empirical probability measures and

empirical entropies derived from α-ary sequences, for which we introduce the following notation. Given an α-ary sequence x_1^n, let z_0^n denote the sequence of states generated by F. For every $z \in S$ and $a \in A$, let

$$\mu_j(za) \triangleq \sum_{i=1}^{j} \delta(z_{i-1}, z; x_i, a),$$

where

$$\delta(z_{i-1}, z; x_i, a) \triangleq \begin{cases} 1 & \text{if } z_{i-1} = z \text{ and } x_i = a, \\ 0 & \text{otherwise,} \end{cases}$$

denote the number of times that an a occurred at state z. An empirical measure over $S \times A$ is given by

$$\hat{P}_n(za) \triangleq \frac{\mu_n(za)}{n},$$

and the corresponding measure on A, conditioned on $z \in S$, relative to x_1^n is

$$\hat{P}_n(a \mid z) \triangleq \begin{cases} 0 & \text{if } \sum_{a \in A} \hat{P}_n(za) = 0, \\ \dfrac{\hat{P}_n(za)}{\sum_{a \in A} \hat{P}_n(za)} & \text{otherwise.} \end{cases}$$

This measure defines the maximum-likelihood model supported by F, i.e.

$$\min_{\theta} [-\log P(x_1^n; \mathscr{X})] = -\sum_{i=1}^{n} \hat{P}_n(x_i \mid z_{i-1}), \tag{1}$$

where $\mathscr{X} = (F, \theta)$ and hereafter the logarithm base is 2. I it is well known that this minimal value is $n \hat{H}(x_1^n \mid F)$, where $\hat{H}(x_1^n \mid F)$ is the conditional entropy with respect to F of this measure, namely

$$\hat{H}(x_1^n \mid F) \triangleq -\sum_{a \in A} \sum_{z \in S} \log \hat{P}_n(za) \log \hat{P}_n(a \mid z).$$

Consequently, the probability assigned by the empirical measure to the entire sequence is $2^{-n\hat{H}(x_1^n \mid F)}$, and this is the maximal probability that can be assigned to that sequence with a supporting machine F.

The nonpredictive stochastic complexity (Rissanen, 1983) of an α-ary sequence x_1^n in the class of FS models, defined as the shortest description length of x_1^n in terms of the models in the class, is given, up to O(1) terms, by

$$\begin{aligned} l_{\mathrm{NP}}(x_1^n) &\triangleq \min_{k, F, \theta} \left\{ -\log P(x_1^n; \mathscr{X}) + \frac{(\alpha - 1)k}{2} \log(n+1) \right\} \\ &= \min_{k, F} \left\{ n \hat{H}(x_1^n \mid F) + \frac{(\alpha - 1)k}{2} \log(n+1) \right\}, \end{aligned} \tag{2}$$

where the first minimum is taken over all FS models $\mathscr{X}=(F, \theta)$ on a set of k states for all k, and the equality follows from (1). This formula results from a particular way of encoding the data, with a 'batch' universal encoder that sends as a header the empirical counts in each state of the model, and then assigns to the data a code matched to the empirical probabilities. The first term in the right-hand side of (2) represents the cost of encoding the data given the counts, the second term represents the cost of encoding the counts, and the cost of encoding the machine is independent of n. Nevertheless, it can be shown (Rissanen, 1986b, Theorem 1) that (2) represents (on the average) the minimum description length of the sequence x_1^n, assumed to be a sample of an FS process, using any noiseless code.

While the above definition of stochastic complexity involves (possibly) batch codes, a partial step in the sequential formulation of the problem is considering the description length of the sequence x_1^n when the parameters are estimated sequentially but the machine is still estimated from the entire data. Formally, for a given FS model, let $\hat{\theta}(i)$ be an estimator of the vector of parameters θ based on a sample of length i. The (semi-) predictive stochastic complexity of an α-ary sequence x_1^n relative to the class of FS models and to $\hat{\theta}(\cdot)$ is defined (Rissanen, 1986a, b) as

$$l_{\mathrm{SP}}(x_1^n) \triangleq \min_{k, F}\left\{-\sum_{i=0}^{n-1} \log \hat{\theta}_i(x_{i+1} \mid z_i)+\log^* j+c\right\}, \tag{3}$$

where

- the minimum is taken over all finite-state machines F on a set of k states, and all k;
- z_0 is the initial state for F and z_0^{n-1} is the sequence of states associated with x_1^n and F;
- $\hat{\theta}_i(\cdot \mid \cdot)$ is the transition probability induced by $\hat{\theta}(i)$;
- j denotes the index of F when all FS machines are ordered in such a way that a machine with fewer states precedes another with more states;
- $\log^* k \triangleq \log k+\log\log k+\cdots$, the sum including all the positive iterates; and
- c is the constant $\log \sum_{j=1}^{\infty} 2^{-\log^* j}$, which forces the code lengths for the indices j to satisfy the Kraft inequality (with equality), and thus to correspond to a uniquely decodable code.

In (3), $\hat{\theta}(i)$ is computed in a predictive way, but the optimizing machine F is not. The term $\log^* j+c$ can be shown to be the length of the header needed to encode the machine, while $-\log \hat{\theta}_i(x_{i+1} \mid z_i)$ is the ideal code length assigned to x_{i+1} based on the estimate of the parameters at time i, given the machine.

Now, for a given F, let $\hat{\theta}(\cdot)$ be the Laplace estimator, namely

$$\hat{\theta}_i(a \mid z)=P_i(a \mid z) \triangleq \frac{\mu_i(za)+1}{\sum_{a \in A} \mu_i(za)+\alpha}.$$

Then it is shown in (Rissanen, 1986b, Theorem 2) that for every FS model $\mathscr{X}=(F,\theta)$ such that the components of θ are bounded away from 0 and 1, we have

$$\frac{-1}{n}E^n_{\mathscr{X}}[l_{\mathrm{sp}}(x_1^n)]\leqslant H_n(\mathscr{X})+\frac{(\alpha-1)k}{2n}\log n+\mathrm{O}(n^{-1}),$$

where $E^n_{\mathscr{X}}[\cdot]$ denotes expectation relative to the process $P(X_1^n;\mathscr{X})$, and $H_n(\mathscr{X})$ denotes the corresponding per-symbol binary entropy of strings of length n. Moreover, it was shown by Krichevsky and Trofimov (1981) that with the modified estimator

$$\hat{\theta}_i(a\,|\,z)=P_i'(a\,|\,z)\triangleq\frac{\mu_i(za)+1/2}{\sum_{a\in A}\mu_i(za)+\alpha/2} \tag{4}$$

we obtain, for *every* individual sequence x_1^n,

$$\frac{-1}{n}l_{\mathrm{SP}}(x_1^n)\leqslant\hat{H}(x_1^n\,|\,F)+\frac{(\alpha-1)k}{2n}\log n+\mathrm{O}(n^{-1}). \tag{5}$$

Hence, with this new choice of $\hat{\theta}(\cdot)$, the (semi-) predictive stochastic complexity deviates from the non-predictive complexity by no more than $\mathrm{O}(n^{-1})$ per symbol. Note that $P_i'(\cdot\,|\,\cdot)$ differs both from the maximum-likelihood measure $\hat{P}_i(\cdot\,|\,\cdot)$ and from Laplace measure $P_i(\cdot\,|\,\cdot)$. Like the latter, it results from averaging θ over the whole parameter space, but instead of using a uniform prior, it is based on one that emphasizes the values situated along the boundary.

The most interesting case is a fully predictive one where, in addition to the parameters, the model is also estimated sequentially, and thus a third, fully predictive, measure of complexity $l_{\mathrm{p}}(x_1^n)$ suggested by Rissanen (1986a) can be defined. Let $(\hat{F}(i),\hat{\theta}(i))$ be an FS model estimator based on a sample of length i, and let $\hat{f}_i$, $\hat{k}(i)$, and $\hat{z}_0(i)$, denote the corresponding next-state function, cardinality, and initial state, respectively. The predictive stochastic complexity of x_1^n relative to the class of FS models and to the model estimator $(\hat{F}(\cdot),\hat{\theta}(\cdot))$, is defined as

$$l_p(x_1^n)\triangleq-\sum_{i=0}^{n-1}\log\hat{\theta}_i(x_{i+1}\,|\,\hat{z}_i), \tag{6}$$

where $\hat{z}_i$ is the state at time i generated by the estimated machine, that evolves recursively according to $\hat{f}_i(\hat{z}_{i-1},x_i)$. Thus, $\hat{\theta}_i(x_{i+1}\,|\,\hat{z}_i)$ is a probability that depends only on the past outcomes, and hence it defines an FS-generated process whose universality with respect to the FS processes is discussed in Section 4.

The various notions of complexity defined above apply to individual strings of data and can be formulated without assuming the existence of an underlying probability distribution. In the following, however, the data is assumed to be a sample of an (unknown) FS process. In this setting, we show in Section 3 the asymptotic equivalence between these complexities. Specifically, it is shown that all the definitions of the stochastic complexity tend to $H_n(\mathscr{X})+[(\alpha-1)k\log n]/2n$, up to $\mathrm{O}(n^{-1})$ terms, both

on the average and almost surely. Note that the various definitions of the stochastic complexity correspond to various coding algorithms, and following the lower bound obtained by Rissanen (1984), $[(\alpha-1)k\log n]/2n$ is the optimal rate of convergence of the expected code length of any algorithm to the entropy. It turns out that this asymptotic equivalence is related to the consistency of the estimator $\hat{F}$: intuitively, if this estimator is strongly consistent, it eventually provides, with high probability, the true underlying FS machine. From that instant there is no additional cost for coding using an incorrect model. Furthermore, if its error probability vanishes fast enough, the extra coding cost for the period until the true model is estimated correctly, is also small. We provide a suitable estimator, possessing these properties, by slightly modifying the MDL estimator, which is given by the minimizing FS machines of (2) or (3). It is an open problem whether such results can be proved using the MDL estimator itself.

In the sequel, we assume that $\hat{\theta}_i(\cdot\,|\,\cdot)$ equals either Laplace's $P_i(\cdot\,|\,\cdot)$ or Krichevsky and Trofimov's $P_i'(\cdot\,|\,\cdot)$ estimators. Since our main result holds in an average sense, these estimators (as well as any other estimator obtained using reasonable priors) are equivalent. The extension of this result to almost sure convergence requires restricting $\hat{\theta}_i$ to be P_i'.

3. The main result

In this section, we compare between the *average values* of the predictive complexity (6) and the (semi-)predictive complexity (3), and establish their asymptotic equivalence. It is conjectured by Rissanen (1986b) that if $\hat{F}(i)$ is taken as the minimizing FS machine that defines $l_{\mathrm{SP}}(x_1^i)$ in (2) or (3) (MDL estimator), then these complexities differ, on the average, by no more than $\mathrm{O}(n^{-1})$ per symbol. The main result of this paper, stated in Theorem 1 below, is that when the initial state for every FS machine F is assumed to be a unique, known state (i.e., when the next-state functions of two distinct models in the class with equal parameter vectors differ by more than a permutation), this conjecture holds with a slight modification of the estimator used.

Theorem 1. *Let the estimator*

$$\hat{F}(i) \triangleq \arg\min_{M} \left\{ \hat{H}(x_1^i \,|\, M) + \frac{2C\alpha\,|\,Z\,|\log(i+1)}{i} \right\}, \tag{7}$$

where the minimum is taken over all FS machines M on a set of states Z and $C>1+0.5\alpha^{-1}$, be used for defining the predictive stochastic complexity $l_p(x_1^n)$ of a sequence x_1^n. Then, for every FS model $\mathscr{X}=(F,\theta)$ with k states, we have

$$\frac{1}{n} E_{\mathscr{X}}^n [l_{\mathrm{P}}(x_1^n) - l_{\mathrm{SP}}(x_1^n)] \leqslant \mathrm{O}(n^{-1}) \tag{8}$$

and, consequently,

$$\frac{-1}{n} E_{\mathscr{X}}^{n}[l_{\mathrm{P}}(x_1^n)] \leqslant H_n(\mathscr{X}) + \frac{(\alpha-1)k}{2n} \log n + \mathrm{O}(n^{-1}).$$

Note that the difference between the 'penalty term' in (7) and the one used in the asymptotic version of the MDL criterion is only in the multiplying constant (which is $0.5\,(\alpha-1)|Z|$ for MDL), and not in its functional behavior. In particular, unlike the criterion proposed by Kieffer (1993), there is an explicit linear penalty for the number of parameters.

Theorem 1 follows from Lemmas 2 and 3 below. Lemma 2 states that in order to establish the asymptotic equivalence (8) it is sufficient that the probability of error in estimating the machine F tend to zero fast enough, while Lemma 3 shows that the estimator (7) has this desired property. To state these lemmas, define the probability of error in determining the model, based on the observations x_1^i up to time i, as

$$P_{\mathrm{error}}(i) \triangleq \sum_{x_1^i \in B} P(x_1^i; \mathscr{X}),$$

where

$$B \triangleq \{x_1^i \in A^i : \hat{F}(i) \neq F\}$$

is the error event, and $P(\cdot\,; \mathscr{X})$ is the probability measure defined by the *true underlying* F and θ over A^n. By the corollary to Lemma 1, B is a well-defined set.

Lemma 2 [Weinberger et al. (1992), Theorem 4(a)]. *A sufficient condition for* (8) *is*

$$\sum_{i=1}^{\infty} P_{\mathrm{error}}(i) \log i < \infty. \tag{9}$$

Although Theorem 4 of Weinberger et al. (1992) was proved for the more restricted class of FSMX models, the proof is equally valid for the class of FS models. In the FSMX case, $P_{\mathrm{error}}(i)$ is the probability of making an error in the estimation of the *specific state* z_i, rather than in the estimation of the whole structure of the finite-state machine F. It follows from Lemma 2 that we need to investigate not only the asymptotic consistency of $\hat{F}(\cdot)$ but also the rate at which $P_{\mathrm{error}}(i)$ approaches zero. Lemma 3 states that the model estimator (7) satisfies (9).

Lemma 3. *If the model estimator* $\hat{F}(i)$ *is defined by* (7), *then* (9) *holds for every FS model* $\mathscr{X} = (F, \theta)$.

Remark. By Lemma 3, the sequence $P_{\mathrm{error}}(i)$ is summable. Therefore, by the Borel–Cantelli lemma, $\hat{F}(i)$ is a strongly consistent estimator for F. Moreover, with similar arguments one can easily prove that when $\hat{\theta}_i = P'_i$, equation (8) holds also in an almost sure sense, i.e. the semi-predictive and the predictive complexities are equal

with probability one. It seems plausible (Finesso, 1993) that using a law of iterated logarithm, this strong consistency can be proved for any estimator with a penalty term that vanishes at a rate slower than $O((\log\log n)/n)$. This would match the result proved by Hannan and Qinn (1979) for AR processes. However, deriving this law for the class of FS processes requires further investigation. Moreover, this method would not provide a bound similar to Lemma 3 on the rate at which the probability of error vanishes and, therefore, it would not be sufficient to derive (8) in Theorem 1.

Proof of Lemma 3. Let $P_M(i)$ denote the probability that the estimate $\hat{F}(i)$ of $F=(S, k, f, z_0)$ take a *specific* value $M=(Z, q, m, z_0(M))$, namely

$$P_M(i) \triangleq \sum_{x_1^i \in B_M} P(x_1^i; \mathscr{X}),$$

where

$$B_M \triangleq \{x_1^i \in A^i: \hat{F}(i) = M\}.$$

The main idea in this proof is to distinguish between the cases where the estimated machine M is a refinement of the true machine F or it is not. The former is an overparametrization case and is handled by the penalty term. The latter yields a model that does not fit the data and, therefore, is handled using large deviations techniques. However, the infiniteness of the number of refined and nonrefined machines M poses a problem when summing over all the cases. We cope with this problem by using a rough technique to rule out, first, machines with a very large number of states (namely $q \geq 4\alpha k$), so that only finitely many machines remain. Thus, let

$$P_1(i) \triangleq \sum_{M, q \geq 4\alpha k} P_M(i)$$

and

$$P_2(i) \triangleq \sum_{q < 4\alpha k,\ M \neq F} P_M(i).$$

Clearly, $P_{\text{error}}(i) = P_1(i) + P_2(i)$. Hence, it suffices to prove that

$$\sum_{i=1}^{\infty} P_1(i) \log i < \infty \tag{10}$$

and

$$\sum_{i=1}^{\infty} P_2(i) \log i < \infty. \tag{11}$$

First, assume $q \geq 4\alpha k$. Under criterion (7), we have

$$B_M \subseteq \left\{ x_1^i \in A^i: \hat{H}(x_1^i \mid F) - \hat{H}(x_1^i \mid M) \geq \frac{2C\alpha(q-k)\log(i+1)}{i} \right\} \triangleq B'_M. \tag{12}$$

Therefore, the criterion involves a comparison between the maximum likelihoods under F and M, with a penalty term that tends to zero as $O(i^{-1}\log i)$. With the convention $0\log 0 \triangleq 0$, we have

$$\log P(x_1^i;\mathscr{X}) = i\sum_{z\in S}\sum_{a\in A}\hat{P}_i(za)\log p(a\,|\,z)$$

$$= i\sum_{z\in S}\left[\sum_{a\in A}\hat{P}_i(za)\sum_{a\in A}\hat{P}_i(a\,|\,z)\log p(a\,|\,z)\right]$$

$$\leqslant i\sum_{z\in S}\left[\sum_{a\in A}\hat{P}_i(za)\sum_{a\in A}\hat{P}_i(a\,|\,z)\log P_i(a|z)\right] = -i\hat{H}(x_1^i\,|\,F), \quad (13)$$

where we have used Gibbs' inequality. By (12) and (13) it follows that

$$P_M(i) \leqslant \sum_{x_1^i\in B_M'} 2^{-i\hat{H}(x_1^i\,|\,F)} < \sum_{x_1^i\in A^i} 2^{-i\hat{H}(x_1^i\,|\,M)-2C\alpha(q-k)\log(i+1)}$$

$$< (i+1)^{-2C\alpha(q-k)}\sum_{x_1^i\in A^i} 2^{-i\hat{H}(x_1^i\,|\,M)}. \quad (14)$$

Next, we use the method of types to show that the sum on the right-hand side of (14) grows polynomially fast with i. Let $\boldsymbol{\Phi}(x_1^i)$ denote the $q\times\alpha$ matrix whose entries are $\mu_i(za)$, $z\in Z$, $a\in A$. The set $T(x_1^i)$ of all sequences in A^i having the same matrix $\boldsymbol{\Phi}(\cdot)$ as x_1^i is referred to as the FS-type of x_1^i relative to M. Let τ_M denote the set of distinct FS-types relative to M. For a type $T\in\tau_M$, let $|T|$ denote its cardinality, let $\hat{H}(T\,|\,M)$ denote the empirical conditional entropy with respect to M of the sequences in T (which depends on the sequence only through its type), and let $Q_T(\cdot)$ denote the empirical probability measure implied on A^i by the counts associated with T. Consequently,

$$\sum_{x_1^i\in A^i} 2^{-i\hat{H}(x_1^i\,|\,M)} = \sum_{T\in\tau_M}|T|\,2^{-i\hat{H}(T|M)}. \quad (15)$$

By (13), for every $y_1^i\in T$ we have $Q_T(y_1^i) = 2^{-i\hat{H}(T|M)}$. Thus, for any $T\in\tau_M$,

$$1 \geqslant \sum_{y_1^i\in T} Q_T(y_1^i) = |T|2^{-i\hat{H}(T|M)},$$

implying, by (15),

$$\sum_{x_1^i\in A^i} 2^{-i\hat{H}(x_1^i\,|\,M)} < |\tau_M| < (i+1)^{\alpha q}.$$

Therefore, (14) takes the form

$$P_M(i) < (i+1)^{-\alpha(2Cq-2Ck-q)}. \quad (16)$$

In addition, it is easy to see that $P_M(i)=0$ for $q>i$. Since there are no more than $(q+1)^{\alpha q}$ distinct machines of cardinality q (in fact, this number includes permuted machines and reducible chains), it follows from (16) that

$$\begin{aligned} P_1(i) &< \sum_{q=4\alpha k}^{i} (q+1)^{\alpha q}(i+1)^{-\alpha(2Cq-2Ck-q)} < \sum_{q=4\alpha k}^{i} (i+1)^{-2\alpha[(C-1)q-Ck]} \\ &< i\cdot(i+1)^{-2\alpha[(C-1)4\alpha k-Ck]} < (i+1)^{-2k\alpha[(C-1)4\alpha-C]+1} \\ &< (i+1)^{-2k\alpha(1-0.5\alpha^{-1})+1} < (i+1)^{-(2\alpha-1)+1} < (i+1)^{-2}, \end{aligned} \tag{17}$$

where we have used the inequalities $C>1+0.5\alpha^{-1}$, $k\geq 1$ and $\alpha\geq 2$. Clearly, (17) implies (10).

As for (11), since the number of distinct FS machines with less than $4\alpha k$ states is bounded and independent of i, it suffices to show that

$$\sum_{i=1}^{\infty} P_M(i)\log i < \infty \tag{18}$$

for every machine M of cardinality $q<4\alpha k$, $M\neq F$. First, assume that M is not a refinement of F. We have

$$B_M \subseteq \left\{ x_1^i \in A^i : \hat{H}(x_1^i \mid M) - \hat{H}(x_1^i \mid F) \leq \frac{2C\alpha(k-q)\log(i+1)}{i} \right\} \triangleq B_M''.$$

Since the right-hand side of the inequality defining B_M'' tends to zero as i tends to infinity, it suffices to prove (18) for the probability of the set

$$\{x_1^i \in A^i : \hat{H}(x_1^i \mid M) - \hat{H}(x_1^i \mid F) \leq \varepsilon\}$$

for some $\varepsilon>0$ depending on F and M but not on i. This is stated by the following lemma.

Lemma 4. *Given an ergodic FS model $\mathscr{X}=(F,\theta)$, consider another ergodic FS model $\mathscr{X}'$, whose FS machine M is not a refinement of F and is characterized by a next-state function that is not a permutation of the next-state function of F. Then, there exists a constant $\delta>0$, that depends only on $\mathscr{X}$ and $\mathscr{X}'$, such that for evry $\varepsilon<\delta$*

$$\limsup_{i\to\infty} \frac{1}{i}\log P\{x_1^i \in A^i : \hat{H}(x_1^i \mid M) - \hat{H}(x_1^i \mid F) \leq \varepsilon; \mathscr{X}\} < 0. \tag{19}$$

The proof of Lemma 4 is given in the appendix and it uses large deviations techniques.

Next, assume that M is a (proper) refinement of F, given by a function d: $Z\to S$. Thus, $q>k$ and the error set B_M satisfies (12). For every $z\in S$, let $Z_S(z)$ denote the subset of Z of cardinality $q(z)$ whose states s satisfy $d(s)=z$. Finally, let $S(Z)$ denote the

set of the l states z of S for which $q(z)>1$. Clearly,

$$\sum_{z\in S(Z)} q(z)=q-k+l. \tag{20}$$

For every $z\in S$ define

$$\Delta_i(z)\triangleq\sum_{a\in A}\sum_{s\in Z_S(z)}\hat{P}_i(sa)\log\frac{\hat{P}_i(a\,|\,s)}{\hat{P}_i(a\,|\,z)}.$$

By the refinement property, we have

$$\hat{H}(x_1^i\,|\,F)-\hat{H}(x_1^i\,|\,M)=\sum_{z\in S(Z)}\Delta_i(z),$$

implying, by (20),

$$B_M\subseteq\left\{x_1^i\in A^i\colon \Delta_i(z)\geqslant\frac{2C\alpha(q-k)q(z)\log(i+1)}{(q-k+l)i}\text{for some } z\in S(Z)\right\}.$$

Hence,

$$P_M(i)<\sum_{z\in S(Z)}P\left\{\Delta_i(z)\geqslant\frac{2C\alpha(q-k)q(z)\log(i+1)}{(q-k+l)i};\mathscr{X}\right\}\triangleq\sum_{z\in S(Z)}P_z(i). \tag{21}$$

In order to upper-bound $P_z(i)$, $z\in S(Z)$, we define a machine $M'(z)$ on a set $Z'(z)$ of $q-q(z)+1$ states, obtained from M by deleting all the states of $Z_S(z)$ except an arbitrary one, which is also denoted by z, and by redirecting to z all the edges incident at the deleted states. The initial state for $M'(z)$ is $z_0(M)$, in case it has not been deleted, or z otherwise. Since $M'(z)$ is still a refinement of F (for which the newly defined z is the only state whose image under d is $z\in S$), an FS model defining the same process as $\mathscr{X}$ is obtained by assigning to every state of $Z'(z)$ the same transition probabilities as those at the corresponding state of S. These probabilities are also denoted by $p(\cdot\,|s)$, $s\in Z'(z)$. Using this new model to define the process, we can proceed as in (13) to obtain for every $x_1^i\in A^i$

$$\begin{aligned}\log P(x_1^i;\mathscr{X})&=i\sum_{s\in Z'(z)}\sum_{a\in A}\hat{P}_i(sa)\log p(a|s)\\&\leqslant i\sum_{s\in Z'(z)-\{z\}}\sum_{a\in A}\hat{P}_i(sa)\log p(a|s)+i\sum_{a\in A}\hat{P}_i(za)\log\hat{P}_i(a|z).\end{aligned} \tag{22}$$

The computation of $\hat{P}_i(sa)$, $s\neq z$, in (22), involves the number of occurrences of s in the sequence of states generated by x_1^i and $M'(z)$. By the definition of the merging process that leads to $M'(z)$, this number may differ from the number of occurrences of the corresponding state with M. However, $p(a\,|\,s)$, $a\in A$, depends only on the state of S determined by the refinement function, and the number of occurrences of this state is the same in both cases. Furthermore, the number of occurrences of z with the machines F and $M'(z)$ is the same. Therefore, by the definition of $\Delta_i(\cdot)$, it can be

readily seen that (22) takes the form

$$\log P(x_1^i;\mathscr{X}) \leqslant i \sum_{s\in Z-Z_S(z)} \sum_{a\in A} \hat{P}_i(sa)\log p(a|s) + i \sum_{s\in Z_S(z)} \sum_{a\in A} \hat{P}_i(sa)\log \hat{P}_i(a|s) - i\,\Delta_i(z). \tag{23}$$

Now, given x_1^i and z, consider a 'semi-empirical' FS model defined by the machine M, in which the empirical transition probabilities gathered from x_1^i are assigned to the states of $Z_S(z)$, and the true (unknown) transition probabilities extended from S as above are used for the other states of Z. Denote by $Q_{x_1^i}(\cdot)$ the probability measure induced over A^i by this model. As a result, in this semi-empirical model the transition probabilities are $p(\cdot|s)$ for every $s\in Z-Z_S(z)$ and $\hat{P}_i(\cdot|s)$ for every $s\in Z_S(z)$. Consequently, we can rewrite (23) as

$$\log P(x_1^i;\mathscr{X}) \leqslant \log Q_{x_1^i}(x_1^i) - i\,\Delta_i(z).$$

Hence, proceeding as in (14),

$$P_z(i) \leqslant 2^{-2C\alpha(q-k)q(z)\log(i+1)/(q-k+l)} \sum_{x_1^i\in A^i} Q_{x_1^i}(x_1^i) = (i+1)^{-2C\alpha(q-k)q(z)/(q-k+l)} \sum_{x_1^i\in A^i} Q_{x_1^i}(x_1^i). \tag{24}$$

To upper-bound the sum on the right-hand side of (24) using the method of types, we define the $q(z)\times\alpha$ matrix $W(x_1^i)$ whose entries are $\mu_i(sa)$, $s\in Z_S(z)$, $a\in A$. The set $T_{M,z}(x_1^i)$ of all sequences in A^i having the same matrix $W(\cdot)$ as that of x_1^i is referred to as the (M,z)-type of x_1^i. Note that $T(x_1^i)\subseteq T_{M,z}(x_1^i)$, where $T(x_1^i)$ denotes the FS-type of x_1^i relative to M, defined previously. Note further that $Q_{y_1^i}(x_1^i)=Q_{x_1^i}(x_1^i)$ for every $y_1^i\in T_{M,z}(x_1^i)$. Let $\tau_{M,z}$ denote the set of distinct (M,z)-types. Proceeding as in (15) and (16), (24) implies

$$P_z(i) < (i+1)^{(-2C\alpha(q-k)q(z))/(q-k+l)}\,|\tau_{M,z}| < (i+1)^{(-2C\alpha(q-k)q(z))/(q-k+l)+\alpha q(z)}.$$

Hence, by (21),

$$P_M(i) < \sum_{z\in S(Z)} (i+1)^{(-2C\alpha(q-k)q(z))/(q-k+l)+\alpha q(z)} = \sum_{z\in S(Z)} (i+1)^{-\alpha q(z)[2C/(1+l(q-k)^{-1})-1]}. \tag{25}$$

Now, since each of the l states of $S(Z)$ corresponds to at least two states of Z after the refinement, we have $q\geqslant k+l$. In addition, by the definition of $S(Z)$, $q(z)\geqslant 2$ for every $z\in S(Z)$. Consequently, the left-hand side of (25) can be further upper-bounded by

$$P_M(i) < k\cdot(i+1)^{-2\alpha(C-1)}.$$

Finally, we have $C = 1 + 0.5\alpha^{-1}(1+\varepsilon)$ for some $\varepsilon > 0$. Hence,

$$P_M(i) < k \cdot (i+1)^{-(1+\varepsilon)},$$

which implies (18) and completes the proof. □

4. Discussion

In Sections 2 and 3 the definition of a measure of complexity in a predictive way by means of (4), (6) and (7), induces a sequence of probability distributions $\hat{\theta}_i(x_{i+1}|\hat{z}_i)$, where $\hat{\theta}_i(\cdot|\cdot)$ is given by (4), and $\hat{z}_i$ is the state at time i, that evolves recursively according to the machine $\hat{F}(i)$ defined in (7). Each distribution depends only on the past outcomes. Thus, this sequence defines a random process $U(X_1^n)$ by assigning

$$U(x_1^n) = 2^{-l_p(x_1^n)}, \quad n = 1, 2, \ldots,$$

which satisfies the marginality condition

$$\sum_{a \in A} U(x_1^n a) = U(x_1^n).$$

By Theorem 1, for *any* FS process $P(X_1^n; \mathscr{X})$ defined by an FS model $\mathscr{X}$ with k states, we have

$$\frac{1}{n} E_{\mathscr{X}}^n \left[\log \frac{U(X_1^n)}{P(X_1^n; \mathscr{X})} \right] \leqslant \frac{(\alpha-1)k}{2n} \log n + O(n^{-1}). \tag{26}$$

On the other hand, (Rissanen, 1986b, Theorem 1), the right-hand side is, for every process $U(\cdot)$ and almost every FS process $P(\cdot)$, also a lower bound (up to an $O(n^{-1})$ term) on the per-symbol average given by the left-hand side. Therefore, $U(\cdot)$ is a *universal* process for the class of FS processes, in the sense that it is as 'close' as possible to all the processes in the class. An upper bound similar to (26) also holds in the almost sure sense. This can be shown by using (5), Lemma 3, the Borel–Cantelli lemma, and the asymptotic equipartition property.

Note that although $U(\cdot)$ is not an FS process itself, it is generated by a sequence of FS distributions $\hat{\theta}_i(\cdot|\cdot)$. Hence, we call it a *universal finite-state-generated* process. Although 'pointwise' universal processes (in the sense that (26) holds for *every* individual sequence x_1^n rather than only on the average) can be obtained (Weinberger et al., 1993a), these are not FS-generated. In other words, the 'plug-in' approach of estimating the best FS model fitting the data string, and then using it to assign a conditional probability to the next symbol, fails in the pointwise case.

The idea of a universal process goes back to Kolmogorov (1965) and Solomonoff (1964), who proved the existence of such process for the class of all recursively enumerable measures. However, their universal measure turns out to be noncomputable and, hence, cannot be used in the solution of sequential decision problems. By

constraining the class of models to be finite state (which represents a huge class by itself), we obtained a universal computable process $U(\cdot)$. But since the complexity of the minimizing step (7) needed to compute $U(x_1^n)$ is enormous due to the size of the class, we still remain at the same level of only theoretically appealing results. The situation changes when we further constrain the models to possess a finite-memory property. In this case, a relatively simple algorithm (the so-called Context algorithm) is shown to generate a universal process (Weinberger et al., 1993b). An immediate application is a universal data compression system, but it can also be used for sequential prediction and, more generally, to make universal sequential decisions on the future outcomes of the observed sequence, as in the gambling (Feder, 1991) problem.

Appendix

Proof of Lemma 4. Let $R=(V, v, r, z_0(R))$ be a common refinement (Feder et al., 1992) of $F=(S, k, f, z_0)$ and $M=(Z, q, m, z_0(M))$. Denote by $g(\cdot)$ and $d(\cdot)$ the functions defining F and M, respectively, from R. It can be readily seen that there exists such a refinement (think of the machine defined by the Cartesian product of F and M). Consequently, $k<v\leqslant qk$. Clearly,

$$\begin{aligned} &P\{x_1^i\in A^i\colon \hat{H}(x_1^i|M)-\hat{H}(x_1^i|F)\leqslant\varepsilon;\mathscr{X}\} \\ &\quad\leqslant P\{x_1^i\in A^i\colon \hat{H}(x_1^i|F)-\hat{H}(x_1^i|R)\geqslant\varepsilon;\mathscr{X}\} \\ &\qquad + P\{x_1^i\in A^i\colon \hat{H}(x_1^i|M)-\hat{H}(x_1^i|R)\leqslant 2\varepsilon;\mathscr{X}\}\triangleq P(E_1)+P(E_2). \end{aligned}$$

Hence, it suffices to show (19) for $P(E_1)$ and $P(E_2)$ separately. Note that the graph corresponding to the FS machine R is not necessarily irreducible, for $z_0(R)$ might be a transient state. However, the arguments used in the proof of the overestimation bound in the first part of Theorem 1 are equally valid for R. Proceeding as in (14)–(16), we obtain

$$P(E_1)<2^{-i\varepsilon}|\tau_R|<(i+1)^{\alpha qk}2^{-i\varepsilon}.$$

Consequently, (19) holds for $P(E_1)$ for every $\varepsilon>0$. As for $P(E_2)$, we proceed as follows.

For every $z\in V$, let $V_M(z)$ and $V_F(z)$ denote the sets of states $z'\in V$ such that $d(z')=d(z)$ and $g(z')=g(z)$, respectively, and define a parameter vector θ^* for R by $p(a|z)=p(a|g(z))$. In other words, $p(\cdot|z)$ is extended to R by making it constant over $V_F(z)$, so that $\mathscr{X}$ and $(R, \theta^*)\triangleq\mathscr{R}$ define the same process. To upper-bound the probability of E_2, we consider $\mathscr{R}$ as the actual model, and we use techniques from the theory of large deviations as applied to Markov chains; in particular, a well-known lemma due to Csiszar et al. (1987), Lemma 2(a). This lemma requires that the error event be given in terms of a set of probability distributions such that it includes

a certain empirical distribution derived from $x_1^i \in E_2$. Consider the two-dimensional probability distribution $P_i(\cdot,\cdot)$ over $V \times V$ defined by

$$P_i(z,s) \triangleq \begin{cases} \hat{P}_i(za) & \text{if } s \triangleq r(z,a), \text{ some } a \in A, \\ 0 & \text{otherwise.} \end{cases}$$

In general, the marginals of this distribution are not equal. For a distribution $Q(\cdot,\cdot)$ over $V \times V$, let $\bar{Q}(\cdot)$ denote its left marginal, and define

$$q(s|z) \triangleq \frac{Q(z,s)}{\bar{Q}(z)}, \quad s,z \in R, \quad \bar{Q}(z) \neq 0.$$

Further, let

$$q(s|d(z)) \triangleq \frac{\sum_{z' \in V_M(z)} Q(z',s)}{\sum_{z' \in V_M(z)} \bar{Q}(z')}, \quad s,z \in V, \quad \sum_{z' \in V_M(z)} \bar{Q}(z') \neq 0.$$

Note that $q(\cdot|d(z))$ is constant over $V_M(z)$. Finally, let Γ denote the set of distributions over $V \times V$ defined by

$$Q(\cdot,\cdot) \in \Gamma \quad \text{iff} \quad \varepsilon(Q) \triangleq \sum_{s,z \in V} Q(z,s) \log \frac{q(s|z)}{q(s|d(z))} \leqslant 2\varepsilon.$$

Clearly, $\varepsilon(Q) \geqslant 0$ for every distribution Q. By the definitions of E_2 and $\hat{H}(x_1^i|\cdot)$, it follows that $x_1^i \in E_2$ if and only if $P_i(\cdot,\cdot) \in \Gamma$ or, equivalently,

$$P(E_2) = P\{P_i(\cdot,\cdot) \in \Gamma; \mathscr{R}\}.$$

Denote by Γ_0 the set of distributions belonging to the closure of Γ (relative to the set of all distributions over $V \times V$) and for which the two marginals are identical. By the above-mentioned large deviations lemma we then obtain

$$\limsup_{i \to \infty} \frac{1}{i} \log P\{P_i(\cdot,\cdot) \in \Gamma; \mathscr{R}\} \leqslant -D,$$

where

$$D \triangleq \min_{Q \in \Gamma_0} D(Q \| P)$$

and

$$D(Q \| P) \triangleq \sum_{s,z \in R} Q(s,z) \log \frac{Q(s,z)}{\bar{Q}(s)\, p(z|s)}.$$

(We use the conventions $0 \log 0 \triangleq 0 \log 0/0 \triangleq 0$ and $\log h/0 \triangleq \infty$ if $h > 0$.) Clearly, $p(z|s)$ is unambiguously determined by $p(a|s)$ for $a \in A$. Note that, by the definition of Γ, D is independent of i, so in order to complete the proof of Lemma 3 it suffices to show that there exists $\delta > 0$ such that if $\varepsilon < \delta$ then $D \neq 0$.

Let $Q^* \in \Gamma_0$ be such that $\varepsilon(Q^*)$ is minimum under the constraint $D(Q^* \| P) = 0$ (clearly, this minimum is attained). Define $\delta \triangleq \varepsilon(Q^*)/2$. Taking $\varepsilon < \delta$ we obtain $D \neq 0$.

Hence, it suffices to show that $\varepsilon(Q^*)\neq 0$. Clearly, $D(Q^*\|P)=0$ implies that $q^*(s|z)=p(s|z)$ so

$$\varepsilon(Q^*)=\sum_{s,z\in V,\overline{Q^*}(z)\neq 0} Q^*(z,s)\log\frac{p(s|z)}{q^*(s|d(z))}.$$

Consequently, $\varepsilon(Q^*)=0$ if and only if $p(s|z)=q^*(s|d(z))$ for every $s,z\in V$ such that $\overline{Q^*}(z)\neq 0$. Thus, in this case $p(\cdot|z)$ must be constant not only over $V_F(z)$ (which holds by definition) but also for those states z' of $V_M(z)$ for which $\overline{Q^*}(z')\neq 0$. Next, this property is used to define a vector of transition probabilities for M. Note that, in particular, if $\mathscr{R}$ is an ergodic FS model with stationary distribution $P^0(\cdot)$, then for every $s,z\in V$ we have $Q^*(z,s)=P^0(z)p(s|z)$ and $\overline{Q^*}(z)=P^0(z)\neq 0$. However, in the general case we might have $\overline{Q^*}(z)=0$ for some $z\in V$, and we overcome this difficulty by using a well-known result in the theory of Markov chains (Cox and Miller, 1967, pp. 99–100) stating that there exists a closed (irreducible) subset V' of V such that $\overline{Q^*}(z)\neq 0$ for every $z\in V'$. By the ergodicity of F and M, for every $u\in S$ and every $w\in Z$ there exist $z,z'\in V'$ such that $u=g(z)$ and $w=d(z')$. Hence, if $\varepsilon(Q^*)=0$ we can define a vector of transition probabilities $t(a|w)$, $a\in A$, $w\in Z$, given by $t(a|w)=p(a|z)$, where $w=d(z)$, $z\in V'$. Now, let $z\in V'$ be such that $g(z)$ is the actual initial state z_0 of the source (in case $\overline{Q^*}(z_0(R))\neq 0$ we may take $z=z_0(R)$), and consider a machine M' identical to M, except that the initial state is $d(z)$ (if $z=z_0(R)$ then $M=M'$). It follows that M', together with the vector of probabilities $t(\cdot|\cdot)$, defines the same process as $\mathscr{R}$ and, consequently, the same as $\mathscr{X}$. In addition, since M is not a refinement of F, and their next-state functions are assumed to differ by more than a permutation, M' is neither a refinement nor a permutation of F. This contradicts either the minimality of F or Lemma 1, and the proof is complete. □

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control.* **AC-19**, 716–723.

Anderson, T.W. (1963). Determination of the order of dependence in normally distributed time series. In: M. Rosenblatt Ed., *Proc. Symp. on Time Series Analysis.* Wiley, New York, 425–446.

Blackwell, D. and L. Koopmans (1957). On the identifiability problem for functions of finite Markov chains, *Ann. Math. Statist.* **28**, 1011–1015.

Cox, D.R. and H.D. Miller, (1967). *The Theory of Stochastic Processes.* Methuen, London.

Csiszar I., T.M. Cover and B. Choi (1987). Conditional limit theorems under Markov conditioning. *IEEE Trans. Inform. Theory* **IT-33**, 788–801.

Davisson, L.D. (1983). Minimax noiseless universal coding for Markov sources. *IEEE Trans. Inform. Theory* **IT-29**, 211–215.

Feder, M. (1991). Gambling using a finite state machine. *IEEE Trans. Inform. Theory* **IT-37**, 1459–1465.

Feder, M., N. Merhav and M. Gutman (1992). Universal prediction of individual sequences. *IEEE Trans. Inform. Theory* **IT-38**, 1258–1270.

L. Finesso (1993). private communication.

Hannan, E.J. and B.G. Quinn, (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B*, **41**, 190–195.

Kieffer, J.C. (1992). Strongly consistent MDL-based selection of a model class for a finite-alphabet source (preprint).

Kolmogorov, A.N. (1965). Three approaches to the quantitative definition of information, *Problems Inform. Transmission* **1**, 4–7.

Krichevsky, R.E. and V.K. Trofimov (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory* **IT-27**, 199–207.

Merhav, N., M. Gutman and J. Ziv, (1989). On the estimation of the order of a Markov chain and universal data compression. *IEEE Trans. Inform. Theory* **IT-35**, 1014–1019.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416–431.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* **IT-30**, 629–636.

Rissanen, J. (1986a). Stochastic complexity and modeling, *Ann. Statist.* **14**, 1080–1100.

Rissanen, J. (1986b). Complexity of strings in the class of Markov sources, *IEEE Trans. Inform. Theory* **IT-32**, 526–532.

Rudich, S. (1985). Inferring the structure of a Markov chain from its output. In: *Proc. 26th IEEE Symp. on Foundations of Computer Science*, 321–326.

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Solomonoff, R.J. (1964). A formal theory of inductive inference I, II. *Inform. Control* **7**, 1–22, 224–254.

Weinberger, M.J., N. Merhav and M. Feder (1993a). Optimal sequential probability assignment for individual sequences. *IEEE Trans. Inform. Theory* (accepted for publication).

Weinberger, M.J., A. Lempel and J. Ziv (1992). A sequential algorithm for the universal of coding finite-memory sources. *IEEE Trans. Inform. Theory* **IT-38**, 1002–1014.

Weinberger, M.J., J. Rissanen and M. Feder (1993b). A universal finite memory source. *IEEE Trans. Inform. Theory* (submitted).

Whittle, P. (1952). Tests of fit in time series. *Biometrika* **39**, 309–318.

Ziv, J. and N. Merhav (1992). Estimating the number of states of a finite-state source. *IEEE Trans. Inform. Theory* **IT-38**, 61–65.