Optimal Sequential Probability Assignment for Individual Sequences

Marcelo J. Weinberger, *Member, IEEE,* Neri Merhav, *Senior Member, IEEE,* and Meir Feder, *Senior Member, IEEE*

***Abstract*— The problem of sequential probability assignment for individual sequences is investigated. We compare the probabilities assigned by any sequential scheme to the performance of the best "batch" scheme (model) in some class. For the class of finite-state schemes and other related families, we derive a deterministic performance bound, analogous to the classical (probabilistic) Minimum Description Length (MDL) bound. It holds for "most" sequences, similarly to the probabilistic setting, where the bound holds for "most" sources in a class. It is shown that the bound can be attained both pointwise and sequentially for any model family in the reference class and without any prior knowledge of its order. This is achieved by a universal scheme based on a mixing approach. The bound and its sequential achievability establish a completely deterministic significance to the concept of predictive MDL.**

***Index Terms*—Universal coding, sequential schemes, minimum description length, finite-state machines, prediction, gambling.**

I. Introduction

It is widely recognized, following Solomonoff [28] and more recently Rissanen [21], [23] and Dawid [9], that an important goal in inductive inference is learning a conditional probability distribution of future data based on the past. Imagine a situation where data is observed sequentially, i.e., at each time instant i and after having observed past data $x_1^i = x_1 x_2 \cdots x_i$ one wishes to make inferences on the next outcome x_{i+1} by assigning a conditional probability distribution $p(\cdot|x_1^i)$ to it. In the long run, the goal is to maximize the assigned probability of the entire sequence

$$P(x_1^n) = \prod_{i=0}^{n-1} p(x_{i+1}|x_1^i). \tag{1}$$

This probability assignment problem finds its applications in coding [24], gambling [10], and prediction [26]. In noiseless coding, for example, $-\log P(x_1^n)$ is the code length of a Shannon code, based on the above probability assignment, which can be implemented sequentially by arithmetic coding [24]. Clearly, a good inference procedure that induces a high probability $P(x_1^n)$ also yields a short code to the given input. As for gambling, if the $\{x_i\}$ are binary, then $2^n P(x_1^n)$ is the capital gain obtained by investing, at each time instant, a fraction of the current fortune that is proportional to the conditional probability $p(\cdot|x_1^i)$ of the next outcome. Again, the larger the probability assigned to x_1^n, the larger is the final fortune.

In spite of being concerned with a sequential probability assignment problem, we do not assume the data to be generated by a probabilistic source but, rather, we regard x_1^n as an individual sequence over a finite alphabet A. Suppose that we have a class of machines with limited resources that assign probabilities to sequences. In many cases, the probabilities induced by these machines correspond to parametric probabilistic models, e.g., the class of finite-state (FS) machines (or FSM's) is associated with probabilistic FS sources, block codes correspond to blockwise memoryless sources, and so on. Let each machine (or model) in the class be indexed by a parameter vector $\boldsymbol{\theta}$ that takes values in a set Θ whose dimension k expresses the amount of resources. Let $P_{\boldsymbol{\theta}}(x_1^i)$ be the probability assigned to a sequence x_1^i by the machine $\boldsymbol{\theta}$. Observe that any machine $\boldsymbol{\theta}$ can be viewed as a sequential probability assignment scheme which at time $i+1$, after having observed x_1^i, assigns to x_{i+1} a conditional probability

$$p_{\boldsymbol{\theta}}(x_{i+1}|x_1^i) = \frac{P_{\boldsymbol{\theta}}(x_1^{i+1})}{P_{\boldsymbol{\theta}}(x_1^i)}, \tag{2}$$

provided that the *marginality condition*

$$\sum_{a \in A} P_{\boldsymbol{\theta}}(x_1^i a) = P_{\boldsymbol{\theta}}(x_1^i) \tag{3}$$

is satisfied for all $i \geq 1$. Now, given x_1^n, we are interested in the logarithm of the assigned probabilities, which gives the code length, or the capital growth rate, induced by the assignment. If the whole sequence was available in advance, then the highest log-likelihood in the family, $\log \max_{\boldsymbol{\theta} \in \Theta} P_{\boldsymbol{\theta}}(x_1^n)$, would be attained by a machine $\boldsymbol{\theta}(x_1^n)$. However, since $\boldsymbol{\theta}(x_1^n)$ depends on the entire sequence, it cannot be anticipated in a sequential regime. Thus, the maximum likelihood probability cannot be assigned to all sequences by a single machine. But we wish to design a single *universal* sequential machine, that implements a probability assignment by reading the components of x_1^n serially and assigning dynamically conditional probabilities $p(x_{i+1}|x_1^i)$ on-line. Here the attributes "sequential" and "universal" are highly interrelated as they both mean that the conditional distribution $p(\cdot|x_1^i)$ depends neither on x_{i+1} nor on future outcomes. In addition, by universality we also mean closeness to optimality and this

Manuscript received April 30, 1993; revised May 25, 1993. The work of M. Feder was supported in part by the Wolfson Research Awards, administered by the Israel Academy of Sciences and Humanities.

M. J. Weinberger is with Hewlett-Packard Laboratories, Palo Alto, CA 94303, USA. This work was done while he was with IBM-Almaden Research Center, San Jose, CA USA.

N. Merhav is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel.

M. Feder is with the Department of Electrical Engineering–Systems, Tel Aviv University, Tel Aviv 69978, Israel.

IEEE Log Number 9216794.

raises the question: How does the accumulated log-likelihood $\sum_{i\geq 0} \log p(x_{i+1}|x_1^i)$ assigned by any universal scheme compare to the optimum "batch" performance $\log \max_{\boldsymbol{\theta}\in\Theta} P_{\boldsymbol{\theta}}(x_1^n)$?

An upper bound on the accumulated log-likelihood of any universal scheme, that holds for "most" sequences (with a suitable definition of this term), and which is related to the concept of competitive optimality of codes [4], [11], is applied to show our first result, stating that for FSM's and other related families, the above log-likelihood must be asymptotically far away from the maximum by a quantity at least as large as $0.5k \log n$, for "most" sequences. As for noiseless coding, this means that for any uniquely decipherable encoder that assigns to x_1^n a code whose length is $L(x_1^n)$, for any $\varepsilon > 0$, all large n, and "most" sequences,

$$L(x_1^n) \geq -\log \max_{\boldsymbol{\theta}\in\Theta} P_{\boldsymbol{\theta}}(x_1^n) + \left(\frac{k}{2} - \varepsilon\right)\log n. \quad (4)$$

This result can be viewed as a deterministic counterpart to the classical probabilistic lower bound [22], which states that for *any* (universal) code, and for "practically all" sources in the class $\{P_{\boldsymbol{\theta}}(\cdot),\ \boldsymbol{\theta} \in \Theta\}$, the expected code length essentially cannot be less than $-E_{\boldsymbol{\theta}} \log P_{\boldsymbol{\theta}}(x_1^n) + 0.5k \log n$, where $E_{\boldsymbol{\theta}}$ denotes expectation with respect to (w.r.t.) $P_{\boldsymbol{\theta}}$. The bound in [22] has strengthened earlier minimax bounds for universal codes developed in [6]–[8] and [13]. Similarly, (4) strengthens the asymptotic version of a pointwise minimax bound due to Shtar'kov [27, Theorem 1] (where a maximum over all n-sequences is taken). Thus, while the minimax results state that there *exist* "situations" where the lower bound holds (where a situation means a source or a sequence, depending on the framework), in [22] and here in (4) the bound holds for *most* situations, where the term "most" will be defined later. It must be said, however, that while the minimax bounds hold for any value of n (see [7] and [27]), the corresponding stronger bounds are merely asymptotic.

Another aspect of (4) is a justification of the concept of Minimum Description Length (MDL) for individual sequences. The quantity $-\log \max_{\boldsymbol{\theta}\in\Theta} P_{\boldsymbol{\theta}}(x_1^n) + 0.5k \log n$ in (4) is recognized as the asymptotic stochastic complexity of x_1^n w.r.t. the model family Θ, [21], and is the basic ingredient in the MDL principle [18], [19]. This principle is applied when we wish to select a model that "explains best" x_1^n, among all models in a sequence of parametric families $\{P_{m,\boldsymbol{\theta}}(\cdot), \boldsymbol{\theta} \in \Theta_m\}$, $m = 1, 2, \cdots$, where Θ_m is a set of parameter vectors whose dimension k_m is nondecreasing with m. Specifically, the MDL principle suggests choosing the model that minimizes the code length for x_1^n. In its original formulation the code length was computed with a "two-part" code, in which the parameters are optimally quantized and encoded, and then the data is encoded with an ideal code length [24], given the quantized parameters. Asymptotically, this results in choosing the model that attains

$$\min_{m,\boldsymbol{\theta}} \left[-\log P_{m,\boldsymbol{\theta}}(x_1^n) + \frac{k_m}{2}\log n\right]. \quad (5)$$

Assuming that the parametric families are nested, the first term in (5) is nonincreasing with k_m, while the second term is increasing. Thus, their sum is normally minimized by some finite m.

The justification of referring to (5) as the MDL of x_1^n has been originally provided in [18] and [19] merely on the basis of the above *particular* coding scheme. Nevertheless, a profound justification exists when x_1^n is treated as a member of a probabilistic ensemble and expectation is taken. In this probabilistic setting, the above mentioned lower bound on the expected length [22] is a *converse* to the universal coding theorem. Thus, the codes traditionally used to express the MDL (two-part, enumerative, mixture, predictive [20]–[22], [29]) are justified by achieving a lower bound "on the average." The fundamental idea of the MDL principle, however, resides in fitting a probabilistic model to a *deterministic* sequence, and thus a stronger justification of the definition of (5) as "the information in x_1^n" [22] results from the bound (4) or its immediate corollary

$$L(x_1^n) \geq \min_{m,\boldsymbol{\theta}} \left[-\log P_{m,\boldsymbol{\theta}}(x_1^n) + \left(\frac{k_m}{2} - \varepsilon\right)\log n\right], \quad (6)$$

which again holds for any encoder and "most" sequences.

Our second main result is that for many useful model classes m can be assumed as unknown, yet a sequential scheme for optimal probability assignment exists. Moreover, this scheme is *strongly* sequential, i.e., the target length n of the sequence does not have to be prespecified. Here, optimality means that our strongly sequential scheme is guaranteed to assign a likelihood that is essentially not less than $\max_{m,\boldsymbol{\theta}}[\log P_{m,\boldsymbol{\theta}}(x_1^n) - 0.5k_m \log n]$, namely, to attain the bound for any value of m *simultaneously*. Thus, the scheme is "*twice-universal*" [25]. In terms of noiseless coding, this means a strongly sequential code whose length never exceeds (6), where the minimum is taken w.r.t. any finite subset of indexes $\{m\}$ that is not necessarily known *a priori*.

The idea in constructing this doubly-universal scheme is to define the universal probability measure $P(\cdot)$ as a two-level mixture of all models (machines) in the family. The first level consists of a continuous mixture in each Θ_m. This is easy to implement (in the FS case) sequentially by accumulating a product of the current (biased) relative frequency estimates of the conditional probabilities at each time instant (a variation of Laplace's estimator [13], [27]). The second level of mixing assumes a certain prior on the integers $m = 1, 2, \cdots$, which if chosen appropriately, creates an overall mixture $P(x_1^n)$ satisfying (3), thus inducing conditional probabilities $\{p(x_{i+1}|x_1^i)\}_{i=0}^{n-1}$ that depend on i and x_1^i but not on n. Hence, it can be implemented in a strongly sequential fashion. Moreover, a key observation is that for the considered model families, the countable mixture over m, in fact, degenerates to a finite mixture, because the contribution of all first-level mixtures over Θ_m is the same for all values of m that exceed a certain threshold depending on n. A similar approach has been proposed in [26] for *batch* coding of Markovian probabilistic sources. Here we widen the scope both to the sequential deterministic setting and to a more general framework that allows several other families of machines. Examples are Markovian (finite-memory) machines, block encoders, and a family of models that allows an abrupt

switch from one FSM to another. For the sake of concreteness, we present the main results first for FSM's, but we show later how similar ideas are applicable to the other model families. In the FSM case, the double universality implies that both the cardinality and the connectivity of the graph supporting the FSM are unspecified. The achievability of the bound further strengthens the deterministic significance of the (predictive) MDL concept.

Note that the Lempel–Ziv algorithm [33] also induces conditional probability distributions [24, Theorem 1], [14] such that the log-likelihood of any individual sequence is asymptotically as large as that assigned by any fixed FS model. This follows since the induced probability measure can be interpreted as a Markovian measure of slowly growing order, which eventually assigns a probability higher than that of any FS scheme. However, the Markovian order is growing indefinitely, and hence, unlike our results, the "redundancy" is large as compared to the lower bound (6).

It is interesting to point out that in the probabilistic setting where x_1^n is governed by some FS source, it has been shown [29] that the MDL is attainable by using a "plug-in" approach, namely, by a sequential encoder that at each time instant re-estimates a model in the family (the number of states, the structure of the machine, and the parameter $\boldsymbol{\theta}$), and arithmetic coding [24] w.r.t. this estimate is employed to encode the next symbol. On first glance, it seems natural to apply this approach when double universality in a deterministic setting is required. However, we show that for a wide class of reasonable model estimators, it does not attain the MDL for every sequence.

The outline of the paper is as follows. In Section II we establish the performance bound and discuss its significance. In Section III, we discuss the achievability of the bound in light of previous work dealing with sequential or doubly-universal codes in the FS class ([20], [25], [27], [29]), and rule out the "plug-in" approach. In Section IV, we present the mixture approach and show that it attains the bound while degenerating to a finite (calculable) summation. Finally, in Section V, we show how the proposed method is applied to handle the issue of both sequentially and doubly universality for other model classes, where we have a countable set of model families and a strongly sequential universal scheme for each family.

II. Statement of the Problem and Performance Bound

Consider the problem of designing a machine $\mathcal{M}$ that, when fed with a finite-alphabet sequence x_1^n, sequentially assigns a conditional probability distribution for the next outcome given the past. What are the fundamental limitations on the highest attainable probability? To answer this question we start by establishing a simple general upper bound on the probabilities $P_{\mathcal{M}}(x_1^n)$ allocated by $\mathcal{M}$ to "most" sequences x_1^n.

Hereafter, we assume that the input alphabet A consists of α letters and we denote the set of all n-sequences by A^n. A *partition* $\{T_j\}_{j=1}^N$ of A^n is a collection of N disjoint subsets of A^n whose union is exactly A^n. The following lemma states that the set of sequences that violate the above mentioned upper bound is small in some sense. Later on we shall apply this lemma to compare the probabilities assigned by $\mathcal{M}$ to those generated by FSM's, which induce a natural partitioning of A^n into types.

Lemma 1: Let $P_{\mathcal{M}}(x_1^n)$ denote the probability assigned by a scheme $\mathcal{M}$ to a sequence x_1^n. Given $\varepsilon > 0$ and a partition $\{T_j\}_{j=1}^N$ of A^n, let $B_{\mathcal{M}}(\varepsilon)$ denote the set of sequences over A^n that do *not* satisfy the upper bound

$$P_{\mathcal{M}}(x_1^n) \le \frac{1}{|T(x_1^n)|N^{1-\varepsilon}} \tag{7}$$

where $T(x_1^n)$ denotes the class containing x_1^n. Then, the fractions ρ_j of sequences in T_j belonging to $B_{\mathcal{M}}(\varepsilon)$, defined by

$$\rho_j \triangleq \frac{|B_{\mathcal{M}}(\varepsilon) \cap T_j|}{|T_j|}, \tag{8}$$

satisfy

$$N^{-1}\sum_{j=1}^{N}\rho_j < N^{-\varepsilon}. \tag{9}$$

Lemma 1 implies that if N grows with n, then the average fraction of sequences in each class to which $\mathcal{M}$ assigns probabilities that violate (7), vanishes with n. In particular, if $\mathcal{M}$ is constrained to assign a fixed probability within each class, then the fraction of classes whose sequences have "large" probabilities vanishes.

Proof of Lemma 1: We have

$$1 = \sum_{x_1^n \in A^n} P_{\mathcal{M}}(x_1^n) > \sum_{x_1^n \in B_{\mathcal{M}}(\varepsilon)} P_{\mathcal{M}}(x_1^n) > \sum_{j=1}^{N} \frac{1}{|T_j|N^{1-\varepsilon}} \cdot \rho_j |T_j|, \tag{10}$$

which yields (9). Q.E.D.

Applying Chebyshev's inequality using a uniform distribution over the classes, it is easy to see that the fraction of classes containing a significant fraction of sequences with "large" probabilities, vanishes, as stated in Corollary 1.

Corollary 1: Let $N_{\mathcal{M}}(\varepsilon)$ denote the number of classes T for which

$$\frac{|B_{\mathcal{M}}(\varepsilon) \cap T|}{|T|} > N^{-\varepsilon/2}. \tag{11}$$

Then,

$$\frac{N_{\mathcal{M}}(\varepsilon)}{N} < N^{-\varepsilon/2}. \tag{12}$$

Lemma 1 can be interpreted in terms of competitive optimality of codes [4], [11]. Consider a probability distribution $P_T(\cdot)$ over A^n that assigns a uniform probability over the classes T_j, $1 \le j \le N$, and a uniform probability within each class. A prefix code matched to P_T would assign to x_1^n a code length $L_T(x_1^n) = \log T(x_1^n) + \log N$. The probability under P_T that a code matched to some $P_{\mathcal{M}}(\cdot)$ would assign to x_1^n less than $L_T(x_1^n) - \varepsilon \log N$ bits is upper bounded by $N^{-\varepsilon}$ (see, e.g., [11, Theorem 1]). Now, clearly, this probability is $N^{-1}\sum_{j=1}^{N}\rho_j$, which is exactly (9).

Hereafter, we assess the performance of $\mathcal{M}$ relative to some competing family F of probability assignment schemes, by

comparing each assigned probability $P_{\mathcal{M}}(x_1^n)$ to the maximum probability $P_{\mathbf{F}}(x_1^n)$ assigned by a scheme $\mathcal{F}(x_1^n)$ in $\mathbf{F}$, matched to the specific sequence. Thus, while $\mathcal{M}$ is a fixed *sequential* scheme, independent of x_1^n, a "batch" procedure in which the sequence is prescanned is allowed for chosing the best scheme in $\mathbf{F}$. Note that, in general, $\mathcal{M} \notin \mathbf{F}$. For a given family $\mathbf{F}$, let the partition of Lemma 1 be defined by the equivalence relation

$$x_1^n \sim y_1^n \text{ iff } P_{\mathcal{F}}(x_1^n) = P_{\mathcal{F}}(y_1^n) \text{ for every } \mathcal{F} \in \mathbf{F}, \quad (13)$$

i.e., two sequences are in the same class if and only if they are assigned the same probability for every $\mathcal{F} \in \mathbf{F}$. Then, Lemma 1 can be used to measure the performance of $\mathcal{M}$ relative to $\mathbf{F}$, provided that we lower-bound $|T(x_1^n)|$, as defined by (13), in terms of $P_{\mathbf{F}}(x_1^n)$, i.e., the "maximum likelihood" of x_1^n w.r.t. the model family $\mathbf{F}$. The reference families in this section contain machines with limited resources, that implement schemes whose ith assignment, $i \geq 1$, depends on the past data x_1^{i-1} only through a state variable z_{i-1}, which is determined by an FSM.

Specifically, an FSM F is defined by a state space S of finite cardinality k, with the transitions between states being determined by a "next-state" function f that maps $S \times A$ into S. When a sequence $x_1\ x_2 \cdots$ is fed into F, the state variable evolves recursively according to

$$z_i \triangleq f(z_{i-1}, x_i), \qquad i \geq 1 \quad (14)$$

where z_0 is a given initial state. An FSM 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, although later on we dispense with this assumption. Markovian machines are a special case where the state is formed by sliding a finite window on the recent past of the data. A machine F is completely characterized by the quadruple (S, k, f, z_0). We also assume that the graph determined by f is strongly connected, but the results can be extended to any FSM. If $\mathbf{F}$ is the family of schemes defined by an FSM F, then by (1) the probability assigned to x_1^n by a scheme $\mathcal{F} \in \mathbf{F}$ has the product form

$$P_{\mathcal{F}}(x_1^n) = \prod_{i=1}^{n} p(x_i|z_{i-1}) \quad (15)$$

where $p(a|z)$, $a \in A$, $z \in S$, is a vector of conditional probabilities that represents the free parameters of the machine, and $z_0^n = z_0\ z_1 \cdots z_n$ denotes the sequence of states, as defined by (14). Now, defining for every $z \in S$ and every $a \in A$ the count

$$\mu_j(za) \triangleq \sum_{i=1}^{j} \delta(z_{i-1}, z; x_i, a), \quad (16a)$$

where

$$\delta(z_{i-1}, z; x_i, a) \triangleq \begin{cases} 1 & \text{if } z_{i-1} = z \text{ and } x_i = a \\ 0 & \text{otherwise,} \end{cases} \quad (16b)$$

then (15) takes the form

$$P_{\mathcal{F}}(x_1^n) = \prod_{a \in A,\ z \in S} p(a|z)^{\mu_n(za)}. \quad (17)$$

Thus, by (13) and (17), the equivalence classes defining the partition to be used in conjunction with $\mathbf{F}$ and Lemma 1, are given by sequences having the same *FS-type* w.r.t. F, i.e., sequences having the same counts $\mu_n(za)$ for every $z \in S$ and every $a \in A$. Furthermore, by the FS property (17), any scheme in $\mathbf{F}$ must use a fixed strategy each time a symbol $a \in A$ is received at state $z \in S$. It follows by Gibb's inequality that, for a given input sequence x_1^n, the maximum probability $P_{\mathbf{F}}(x_1^n)$ is attained when each $p(a|z)$ agrees the "empirical" conditional probability $\hat{P}_n(a|z)$ relative to x_1^n, defined by

$$\hat{P}_n(a|z) \triangleq \begin{cases} 0 & \text{if } \mu_n(z) \triangleq \sum_{a \in A} \mu_n(za) = 0 \\ \frac{\mu_n(za)}{\mu_n(z)} & \text{otherwise,} \end{cases} \quad (18)$$

and derived from the joint empirical measure over $S \times A$

$$\hat{P}_n(za) \triangleq \frac{\mu_n(za)}{n}. \quad (19)$$

Thus, the best scheme in $\mathbf{F}$ yields a total probability

$$P_{\mathbf{F}}(x_1^n) = 2^{-n\hat{H}(x_1^n|F)} \quad (20)$$

where $\hat{H}(x_1^n|F)$ is the conditional entropy w.r.t. F of the empirical measure, namely

$$\hat{H}(x_1^n|F) \triangleq -\sum_{a \in A} \sum_{z \in S} \hat{P}_n(za) \log \hat{P}_n(a|z) \quad (21)$$

where hereafter the logarithms are taken to the base 2 and $0 \log 0 \triangleq 0$.

Next, we use Corollary 1 to derive a bound on the performance of any scheme $\mathcal{M}$ w.r.t. the family of schemes defined by an FSM F, for "most" sequences, in the sense defined by (9). By (20), we need to lower-bound the difference

$$-n^{-1} \log P_{\mathcal{M}}(x_1^n) - \hat{H}(x_1^n|F). \quad (22)$$

This is done in the following theorem, which is the main result of this section.

Theorem 1: Let $\mathcal{M}$ be an arbitrary probability assignment scheme. For any FSM $F = (S, k, f, z_0)$ supported by a strongly connected graph, fix $\varepsilon > 0$ and let $B_{\mathcal{M}}(\varepsilon|F)$ denote the set of n-sequences for which $\mathcal{M}$ assigns a probability $P_{\mathcal{M}}(\cdot)$ such that

$$-\frac{1}{n} \log P_{\mathcal{M}}(x_1^n) < \hat{H}(x_1^n|F) + \left[\frac{k(\alpha - 1)}{2} - \varepsilon\right] \frac{\log n}{n}. \quad (23)$$

Let $N_{\mathcal{M}}(\varepsilon|F)$ denote the number of FS-types T w.r.t. F for which

$$\frac{|B_{\mathcal{M}}(\varepsilon|F) \cap T|}{|T|} > n^{-\varepsilon/3}, \quad (24)$$

and let N denote the total number of FS-types. Then,

$$\lim_{n\to\infty} \frac{N_{\mathcal{M}}(\varepsilon|F)}{N} = 0. \tag{25}$$

Theorem 1 tells us that the log-likelihood assigned by any scheme to "*most*" sequences of "*most*" FS-types w.r.t. F, is asymptotically far away from the maximum by a quantity at least as large as $0.5k(\alpha - 1)\log n$. This, of course, does not imply that the bound holds for all but a vanishing fraction of sequences. However, the rationale in measuring the relative sizes of the exception sets type by type is that it amounts to considering regions where the competing FSM behaves alike, and only the behavior of $\mathcal{M}$ may vary from one sequence to another.

As observed from (7), the use of Corollary 1 in the proof of Theorem 1 requires auxiliary lower bounds on the number N of different FS-types and on the size $|T|$ of an FS-type. The former bound is stated in Lemma 2 below, and its proof, which was given to us by N. Alon, is not reproduced here. The lower bound on $|T|$, stated in Lemma 3, is derived in Appendix A and holds for all but a vanishing fraction of types, as stated in Lemma 4. An alternative bound that holds for every FS-type is given in [3], but it is not tight enough for our purposes.

Lemma 2: Let $F = (S, k, f, z_0)$ be an FSM supported by a strongly connected graph. Then,

$$N > Cn^{k(\alpha-1)} \tag{26}$$

where C is a constant that depends only on F.

Lemma 3: Given $\varepsilon > 0$, let T be an FS-type relative to an FSM F supported by a strongly connected graph, such that for every $z \in S$ and every $a \in A$,

$$\mu_n(za) > \mu_n(z)\delta_{\varepsilon,F}(n) \tag{27}$$

where $\delta_{\varepsilon,F}(n)$ is a vanishing function that depends only on ε and on F. Then, for all sufficiently large n we have

$$\log|T| > n\hat{H}(T|F) - \left[\frac{k(\alpha-1)}{2} + \varepsilon\right]\log n \tag{28}$$

where $\hat{H}(T|F)$ denotes the empirical conditional entropy $\hat{H}(x_1^n|F)$, which depends on x_1^n only through its type $T(x_1^n) = T$.

The FS-types not covered by Lemma 3 represent a vanishing fraction of the total number of FS-types, as stated in Lemma 4 below.

Lemma 4: Given an FSM F supported by a strongly connected graph, together with a vanishing function $\delta(n)$, let $N(\delta)$ denote the number of FS-types T relative to F, such that $\mu_n(za) \le \mu_n(z)\delta(n)$ for every $z \in S$ and every $a \in A$. Then,

$$\lim_{n\to\infty} \frac{N(\delta)}{N} = 0. \tag{29}$$

The proof of Lemma 4 is given in Appendix B. Lemmas 1–4 provide the tools to prove Theorem 1.

Proof of Theorem 1: Let $\delta_{\varepsilon/6,F}(n)$ denote the vanishing function defined in Lemma 3, and let T_δ denote the set of FS-types that satisfy the corresponding condition of Lemma 3. First, consider a sequence $x_1^n \in B_{\mathcal{M}}(\varepsilon|F)$ such that $T(x_1^n) \in T_\delta$. By Lemma 3, and using the simplified notation $K \triangleq k(\alpha - 1)$, we have

$$\begin{aligned}
-\log P_{\mathcal{M}}(x_1^n) &< n\hat{H}(x_1^n|F) + \left(\frac{K}{2} - \varepsilon\right)\log n \\
&< \log|T(x_1^n)| + \left(K - \frac{5\varepsilon}{6}\right)\log n \\
&= \log|T(x_1^n)| + \left(1 - \frac{\log N - \log n^{K-5\varepsilon/6}}{\log N}\right) \\
&\quad \cdot \log N.
\end{aligned} \tag{30}$$

By Lemma 2, for all sufficiently large n we have

$$N > n^{K-\varepsilon/6}. \tag{31}$$

Thus, (30) yields

$$-\log P_{\mathcal{M}}(x_1^n) < \log|T(x_1^n)| + \left(1 - \frac{\log n^{2\varepsilon/3}}{\log N}\right)\log N. \tag{32}$$

By Corollary 1, the fraction of FS-types that belong to T_δ and such that

$$\frac{|B_{\mathcal{M}}(\varepsilon|F) \cap T|}{|T|} > N^{-\frac{\log n^{2\varepsilon/3}}{2\log N}} = n^{-\varepsilon/3} \tag{33}$$

never exceeds $n^{-\varepsilon/3}$. Now, by Lemma 4, only a vanishing fraction of FS-types does not belong to T_δ. Hence, the total fraction $N^{-1}N_{\mathcal{M}}(\varepsilon|F)$ of types satisfying (24) vanishes. Q.E.D.

As discussed in Section III, we consider the achievability of the bound of Theorem 1 relative to the entire class of FS schemes, rather than a specific machine F. To this end, the following corollary is needed.

Corollary 2: Let $\boldsymbol{J}$ denote an arbitrary set of FSM's $F = (S, k, f, z_0)$, each supported by a strongly connected graph. Fix $\varepsilon > 0$, and let $B_{\mathcal{M}}(\varepsilon)$ denote the set of n-sequences for which

$$\begin{aligned}
&-\frac{1}{n}\log P_{\mathcal{M}}(x_1^n) \\
&\qquad < \min_{F\in\boldsymbol{J}}\left\{\hat{H}(x_1^n|F) + \left[\frac{k(\alpha-1)}{2} - \varepsilon\right]\frac{\log n}{n}\right\}.
\end{aligned} \tag{34}$$

Then,

$$B_{\mathcal{M}}(\varepsilon) = \bigcap_{F\in\boldsymbol{J}} B_{\mathcal{M}}(\varepsilon|F). \tag{35}$$

The above corollary states that the "exceptional" set $B_{\mathcal{M}}(\varepsilon)$ of the sequences for which $\mathcal{M}$ successfully competes [as defined by (23)] with the best FS scheme, based on any machine $F \in \boldsymbol{J}$, is the intersection of the exceptional sets for all F and, *a fortiori*, it is "small."

Comparison with previously reported results: So far we have been concerned with optimal probability assignments, and this allows the consideration of general sequential decision problems as noiseless coding, gambling, and prediction. In the coding case, each probability measure $P_{\mathcal{M}}(x_{i+1}|x_1^i)$, $0 \le i < n$, can be used with an arithmetic coder [24] to perform a noiseless code $L_{\mathcal{M}}$ with total code length

$$L_{\mathcal{M}}(x_1^n) = -\sum_{i=0}^{n-1} \log P_{\mathcal{M}}(x_{i+1}|x_1^i) = -\log P_{\mathcal{M}}(x_1^n). \quad (36)$$

Thus, one can define the scheme $\mathcal{M}$ as a noiseless sequential encoder competing with FS encoders, as was done in [33], with the *additional constraint* that the ideal code lengths $L_{\mathcal{F}}(z, a)$, $a \in A$, $z \in S$, assigned by the FS encoder at each state, satisfy the Generalized Kraft inequality $\sum_{a \in A} 2^{-L_{\mathcal{F}}(z,a)} \le 1$ for every z. Now, proceeding as in [27, Theorem 1], one can easily show that for any scheme $\mathcal{M}$, any machine F, and every n, we have

$$\max_{x_1^n \in A^n} [n^{-1} L_{\mathcal{M}}(x_1^n) - \hat{H}(x_1^n|F)] \ge n^{-1} \log \left[\sum_{x_1^n \in A^n} 2^{-n\hat{H}(x_1^n|F)} \right]. \quad (37)$$

It can be further shown that an asymptotically equivalent lower bound is

$$\max_{x_1^n \in A^n} [n^{-1} L_{\mathcal{M}}(x_1^n) - \hat{H}(x_1^n|F)] \ge k(\alpha - 1)\frac{\log n}{2n} - O(n^{-1}). \quad (38)$$

Thus, Theorem 1 strengthens the pointwise minimax bound (37), due to Shtar'kov [27, Theorem 1], in a way similar to that in which Rissanen's bound on the average code length [22] strengthens the earlier minimax bounds [6]–[8], [13], in the probabilistic case. In this process, we obtain an asymptotic result, instead of a bound that holds for any value of n, as in [27]. Again, a similar phenomenon occurs when comparing the lower bounds of [22] and [7].

Moreover, Theorem 1 has a completely deterministic meaning, unlike the setting of [27], where the bound involves an (implicit) average criterion, along with the maximum taken over A^n, which is suited to individual sequences. In [27], x_1^n is emitted by an FS source (supported by an FSM F) with probability $P_{\boldsymbol{\theta}}(x_1^n|F)$, where $\boldsymbol{\theta}$ denotes a vector of conditional symbol probabilities. Although (22) is there defined as a "maximum own redundancy" for single messages [with a length function satisfying (36)], the definition relies on the interpretation of

$$\frac{1}{n}[L_{\mathcal{M}}(x_1^n) + \log P_{\boldsymbol{\theta}}(x_1^n|F)] \quad (39)$$

as a redundancy. Now, the use of (39) assumes, implicitly, that having a code length close to $-\log P_{\boldsymbol{\theta}}(x_1^n|F)$ is a desirable goal, which of course is the case if we need to minimize also the *average* code length w.r.t. $P_{\boldsymbol{\theta}}(\cdot|F)$. In this respect, we notice that some authors refer to (39), in a probabilistic setting, as a "pointwise redundancy" (see, e.g., [5], [17, Definition 1]), unlike our definitions, where this term is reserved to (22).

Finally, we comment that although we assumed the graph of the considered FSM's to have exactly α outgoing edges per state, the results can be easily extended to any strongly connected graph with E edges. In this case the model cost [i.e., the second-order terms in the right-hand sides of (23) and (34)] would be $(E - k)(\log n)/2n$. For example, $E = k^2$ corresponds to a first-order Markov chain, while $E = \alpha k$ is the case considered so far. This allows different alphabet sizes per state, which is essentially different from letting α be the maximum size and taking some transition probabilities to be zero.

III. Achievability and the "Plug-in" Approach

Next, we discuss the achievability of the bounds given by Theorem 1 and Corollary 1. First, assume that $F = (S, k, f, z_0)$ is a fixed, given FSM. Following Theorem 1 and considering second-order asymptotics, we define a scheme $\mathcal{M}$ as *universal* w.r.t. F, if for all sufficiently large n

$$\max_{x_1^n \in A^n} [-n^{-1} \log P_{\mathcal{M}}(x_1^n) - \hat{H}(x_1^n|F)] \le k(\alpha - 1)\frac{\log n}{2n} + O(n^{-1}). \quad (40)$$

A scheme satisfying (40) achieves uniformly (up to an ε) the bound (23) for a given F and every sequence. Since $\mathcal{M}$ is sequential by definition, the probabilities it assigns must satisfy the condition (3) on the marginals. Moreover, the considered schemes are *strongly sequential*, i.e., the probability assigned to x_{i+1} depends neither on future outcomes nor on the length n of the entire sequence x_1^n to be processed. Noiseless codes derived as in (36), with a probability assignment satisfying (3), have been termed *regular* [20]. A universal probability assignment mechanism (for a specific F) and its sequential code, have been studied in [13], [27], where it is shown that

$$P'(x_1^n|F) = \prod_{i=1}^{n} \theta'_{i-1}(x_i|z_{i-1}, F), \quad z_i \triangleq f(z_{i-1}, x_i), \quad 1 \le i \le n, \quad (41)$$

where

$$\theta'_i(a|z, F) \triangleq \frac{\mu_i(za) + 1/2}{\mu_i(z) + \alpha/2}, \quad (42)$$

satisfies (40). The assignment (41)–(42) is similar to Laplace's rule of succession, except for a different bias.

Here, however, we consider a stronger definition of universality, that corresponds to Corollary 2. Let Ψ denote the set of all FSM's (of any number of states), and let $\boldsymbol{J}$ denote a finite (unspecified) subset of Ψ. Following Corollary 2, a scheme $\mathcal{M}$ is defined as universal in the class of FS schemes, if for *every* $\boldsymbol{J}$ and all sufficiently large n (depending on $\boldsymbol{J}$), it satisfies

$$\max_{x_1^n \in A^n} \left\{ -\frac{1}{n} \log P_{\mathcal{M}}(x_1^n) - \min_{F \in \boldsymbol{J}} \cdot \left[\hat{H}(x_1^n|F) + k(\alpha - 1)\frac{\log n}{2n} \right] \right\} \le O(n^{-1}). \quad (43)$$

Thus, we are interested in schemes that are *simultaneously universal* w.r.t. *every* FSM $F \in \boldsymbol{J}$, with an unspecified number of states, where $\boldsymbol{J}$ is finite but unknown. Note that although the class of FS models is not nested (as opposed to the discussion in Section I), each k-state FSM is nested in all its refinements with $k' > k$ states, and thus the same arguments apply to the minimum in (43). Universality relative to the FS class, can be expressed in terms of universality w.r.t. each F, as stated in Lemma 5 below.

Lemma 5: A scheme $\mathcal{M}$ is universal in the FS class, if and only if it is universal w.r.t. every $F \in \Psi$.

Proof: First, assume that $\mathcal{M}$ is universal w.r.t. every $F \in \Psi$. Thus, (40) holds for every F, with the $O(n^{-1})$ term being a function $\varepsilon_F(n)$ that depends only on F. Next, consider a finite subset $\boldsymbol{J}$ of Ψ. Since $\boldsymbol{J}$ is finite, there exists an $O(n^{-1})$ function that uniformly upper-bounds $\varepsilon_F(n)$ over $\boldsymbol{J}$. Hence, (40) (with $F \in \boldsymbol{J}$) clearly implies (43), and since $\boldsymbol{J}$ is arbitrary, $\mathcal{M}$ is universal in the FS class.

Conversely, if $\mathcal{M}$ is universal in the FS class, (43) holds for every finite subset $\boldsymbol{J}$ of Ψ. In particular, it holds when $\boldsymbol{J}$ is a single machine F. Consequently, (40) holds for every $F \in \Psi$. Q.E.D.

A code derived from such a universal scheme $\mathcal{M}$ asymptotically achieves, for every sequence and in a predictive manner, the MDL given by a two-part code in the class of FS models, without any prior knowledge of the model structure. Codes whose universality applies not only to a given F but to every F simultaneously, have been termed "twice-universal" [25]. However, the universality of the codes considered in [25] has been established merely in a probabilistic sense.

To the best of our knowledge, the problem of achieving the MDL both pointwise and sequentially, has not been explicitly treated before. The semi-predictive code of [20, Theorem 2] is not sequential, for a prescan step is needed to find the optimal machine F. In [27, Theorem 3], where the questions of strong sequentiality and pointwise optimality are also addressed, F is assumed as known, while we require universality w.r.t. the entire class. Although this requirement [see (43)] is also stronger than that of [27, Corollary 4], the mixing approach employed there would work in our framework if an upper bound on k was known. However, this approach fails when the set $\boldsymbol{J}$ is unknown. Also, if $\boldsymbol{J}$ grows with n, as suggested in [27], the marginality condition (3), needed for strong sequentiality, might be violated. Finally, the scheme proposed in [29] is sequential, but achieves the MDL in the probabilistic sense only, with the data being a sample of some FS source.

Due to their success in solving related problems, the sequential schemes of [27] and [29] are plausible for universal probability assignment in the sense defined by (43). Observe that although the conditional measure (42), used for universality in [27] for a given F, results from a mixture of measures over the parameter space, it can also be viewed as "plugging" the "estimated parameter" $\theta'_i(x_{i+1}|z_i, F)$ at time $i+1$. In the same spirit, a "plug-in" approach is also used for double universality in a probabilistic setting [29], where at each time instant i a model structure $\hat{F}_i$, which recursively generates a state sequence $\hat{z}_i$, is estimated, and then the empirical probability $\theta'_i(x_{i+1}|\hat{z}_i, \hat{F}_i)$ induced by the estimated model is assigned to the next symbol x_{i+1}. The total probability assigned to x_1^n by this "plug-in" scheme is $P^{PI}(x_1^n) = \prod_{i=0}^{n-1} \boldsymbol{\theta}'_i(x_{i+1}|\hat{z}_i, \hat{F}_i)$. To assess its performance in a deterministic setting, we must compare

$$-\frac{1}{n}\log P^{PI}(x_1^n) = -\frac{1}{n}\sum_{i=1}^{n}\log \theta'_i(x_{i+1}|\hat{z}_i, \hat{F}_i) \triangleq l^{PI}(x_1^n) \tag{44}$$

to the per-symbol MDL

$$\min_{F\in\boldsymbol{J}}\left\{\hat{H}(x_1^n|F) + k(\alpha-1)\frac{\log n}{2n}\right\} \tag{45}$$

where k is the number of states in F and $\boldsymbol{J}$ is any finite set of FSM's. As discussed in Section II, (45) is essentially the optimal code length for "most" sequences.

A natural model structure estimator to be used sequentially in this plug-in approach is the asymptotic MDL estimator w.r.t. the data x_1^i observed so far, i.e.,

$$\hat{F}_i = \arg\min_{F\in\Psi}\left\{\hat{H}(x_1^i|F) + k(\alpha-1)\frac{\log i}{2i}\right\}. \tag{46}$$

The resulting per-symbol code length $l^{PI}(x_1^n)$, defined in [21] as the *predictive stochastic complexity* of x_1^n w.r.t. the appropriate class of models, was conjectured there to be asymptotically equivalent to the nonpredictive stochastic complexity, i.e., to the MDL. This conjecture has been confirmed [29] both in expectation and with probability one using a slight modification of (46). However, it can be shown that for any sequential model structure estimator of the form

$$\hat{F}_i = \arg\min_{F\in\Psi}\{\hat{H}(x_1^i|F) + \nu(i)\} \tag{47}$$

where $\nu(i)$ is a vanishing positive "penalty term," the conjecture fails to hold uniformly for every sequence. In other words, there is a counterexample sequence whose code length in the plug-in approach is larger than (45) by at least an $O(n^{-1}\log n)$ term. Note that both (46) and the estimator [29] are special cases of (47). Estimators having the asymptotic form (47) are reasonable since, by (20), $\hat{H}(x_1^i|F)$ is the minimum of $-i^{-1}\log P_{\mathcal{F}}(x_1^i)$ over all probability distributions $P_{\mathcal{F}}(\cdot)$ that are based on F and, furthermore, higher order models in the class that are refinements of lower order models, uniformly yield a smaller empirical entropy. Thus, a penalty term $\nu(i)$ is needed. In Appendix C we sketch the idea underlying the construction of the counterexample sequence. More details can be found in [12].

IV. The "Mixture" Approach

Rather than trying to estimate the best machine F that fits the data, we construct a universal scheme $\mathcal{M}$ satisfying (43) for every finite set $\boldsymbol{J}$ of FSM's and all sufficiently large n, by computing a "mixture" of schemes satisfying (40), one for each specific F. This idea has already been applied in [26] for prediction in a probabilistic setting. The difficulty is that $\boldsymbol{J}$ is unknown and the number of model families is countably infinite, so a naive mixture involves an infinite summation and hence is inapplicable. Fortunately, it turns out that by defining the components of the mixture appropriately, the contributions

of all machines with sufficiently many states can be made identical. Since these machines assign the same probability, this enables the computation of the mixture. This approach is applied also in Section V to other situations of interest.

Specifically, for a given F, consider the probability assignment defined by (41) and (42). Following [20], we order the set $\Psi = \{F_1, F_2, \cdots\}$ of all FSM's, such that a machine with fewer states precedes another with more states, and the ordering of machines with the same number of states is arbitrary. Let $\{\gamma(j)\}_{j\geq 1}$ denote a positive summable sequence, $\Gamma_\infty = \sum_{j=1}^{\infty}\gamma(j)$, and $\Gamma_i = \sum_{j=1}^{i}\gamma(j)$. Define the probability measure

$$P'(x_1^n) \triangleq \Gamma_\infty^{-1}\sum_{j=1}^{\infty}\gamma(j)P'(x_1^n|F_j). \tag{48}$$

Note that $\Gamma_\infty^{-1}\gamma(j)$ serves as a prior on $F_j \in \Psi$. This slowly decaying weighting is needed to cope with the infiniteness of Ψ. By (41) and (42),

$$\sum_{a\in A}P'(x_1^i a) = P'(x_1^i) \tag{49}$$

for every $i \geq 0$, with x_1^0 denoting the null string λ, for which $P'(\lambda|F) \triangleq 1$. Hence,

$$P_{\mathcal{M}'}(x_{i+1}|x_1^i) \triangleq \frac{P'(x_1^{i+1})}{P'(x_1^i)} \tag{50}$$

is a well-defined probability assignment for every $i \geq 0$. Since the probability assigned to each symbol depends on the past string only, this assignment can be implemented by a sequential scheme $\mathcal{M}'$. Moreover, for every $F_j \in \Psi$ with k_j states we have

$$P'(x_1^n) \geq \frac{\gamma(j)}{\Gamma_\infty}P'(x_1^n|F_j), \tag{51}$$

or, equivalently,

$$-\log P'(x_1^n) \leq -\log P'(x_1^n|F_j) + \log\frac{\Gamma_\infty}{\gamma(j)}. \tag{52}$$

Now, by [13], [27], for every $x_1^n \in A^n$

$$-\frac{1}{n}\log P'(x_1^n|F_j) \leq \hat{H}(x_1^n|F_j) + k_j(\alpha-1)\frac{\log n}{2n} + O(n^{-1}) \tag{53}$$

which together with (52) implies

$$-\frac{1}{n}\log P'(x_1^n) \leq \hat{H}(x_1^n|F_j) + k_j(\alpha-1)\frac{\log n}{2n} + O(n^{-1}), \tag{54}$$

where the last term in the right-hand side of (52) has been included in the $O(n^{-1})$ term. Consequently, $\mathcal{M}'$ is universal w.r.t. every $F_j \in \Psi$. By Lemma 5, it is also universal in the class of FSM's. Of course, (48) cannot be computed since it involves an infinite summation. In addition, the real constant Γ_∞ is needed if we want $P'(\cdot)$ to be a probability measure. Finally, observe that a naive approach of mixing a number of terms that grows with n, might violate (3) and hence must be ruled out. We next propose a method that overcomes these obstacles.

Define for each $F = (S, k, f, z_0) \in \Psi$ an auxiliary distribution

$$P(x_1^n|F) \triangleq \prod_{i=1}^{n}\theta_{i-1}(x_i|z_{i-1}, F), \quad z_i = f(z_{i-1}, x_i),\ 1 \leq i \leq n \tag{55}$$

where

$$\theta_i(a|z, F) \triangleq \begin{cases}\alpha^{-1} & \text{if } i < k\\ \theta_i'(a|z) & \text{otherwise,}\end{cases} \tag{56}$$

and $P(\lambda|F) \triangleq 1$. The auxiliary parameter vector $\theta_i(a|z, F)$ differs from $\theta_i'(a|z)$ only in a number of symbols that equals the model order and, hence, is independent of n. Thus, the deviation is asymptotically inconsequential. This idea was already employed by Ryabko [26, eq. (9)] in the Markovian case. However, the goal in [26] is the definition of a Markovian process in a probabilistic framework, without specifying an initial state, which requires arbitrary probability assignments until there are enough symbols for determining a state. Now, for every $n \geq 0$, let $F_{j(n)}$ denote the last machine in Ψ with n states. Given a constant $\Gamma \geq \Gamma_\infty$, define

$$P(x_1^n) \triangleq \Gamma^{-1}\sum_{j=1}^{j(n)}\gamma(j)P(x_1^n|F_j) + \left(1 - \frac{\Gamma_{j(n)}}{\Gamma}\right)\alpha^{-n}, \tag{57}$$

which is clearly a probability measure for each $n \geq 0$. The particular assignment (56) guarantees a finite number (growing with n) of machines in the summation of (57), while the contribution of the others, that assign a uniform probability to all the sequences, is gathered in the rightmost term of (57). On the other hand, the assignment (56) for $i < k$ is unimportant in [26], where the issue of the infiniteness of the summation is not considered, thus maintaining the infinite summation of [26, eq. (11)]. We interpret (57) as a mixture of probability measures with a prior $\Gamma^{-1}\gamma(j)$ on $F_j \in \Psi$, $j \leq j(n)$, where the machines with $j > j(n)$ share a common weight $1-\Gamma^{-1}\Gamma_{j(n)}$. The use of Γ is aimed to avoid a real constant Γ_∞. Note that the prior depends on n whenever $\Gamma \neq \Gamma_\infty$, but in a way that guarantees the marginality condition (3), as claimed in Theorem 2, where a new universal machine $\mathcal{M}$ is defined.

Theorem 2: Let

$$P_{\mathcal{M}}(x_{i+1}|x_1^i) \triangleq \frac{P(x_1^{i+1})}{P(x_1^i)}. \tag{58}$$

Then $P_{\mathcal{M}}(\cdot|\cdot)$ is a well-defined sequential probability assignment for every $i \geq 0$, and the corresponding scheme $\mathcal{M}$ is universal in the class of FSM's.

Proof: First, we show that $P_{\mathcal{M}}(x_{i+1}|x_1^i)$ is a well-defined conditional probability. Consider the sum

$$\sum_{a\in A}P(x_1^n a) = \Gamma^{-1}\sum_{j=1}^{j(n+1)}\left[\gamma(j)\sum_{a\in A}P(x_1^n a|F_j)\right] + \left(1 - \frac{\Gamma_{j(n+1)}}{\Gamma}\right)\alpha\cdot\alpha^{-(n+1)}. \tag{59}$$

By (55), for every $n \geq 0$ and $j \geq 1$ we have

$$\sum_{a \in A} P(x_1^n a | F_j) = P(x_1^n | F_j). \tag{60}$$

In addition, by (55) and (56), if $j > j(n)$ then $P(x_1^n|F_j) = \alpha^{-n}$ for every $x_1^n \in A^n$. Hence, (59) takes the form

$$\begin{aligned}\sum_{a \in A} P(x_1^n a) &= \Gamma^{-1} \sum_{j=1}^{j(n)} \gamma(j) P(x_1^n|F_j) \\ &\quad + \frac{\Gamma_{j(n+1)} - \Gamma_{j(n)}}{\Gamma} \alpha^{-n} \\ &\quad + \left(1 - \frac{\Gamma_{j(n+1)}}{\Gamma}\right) \alpha^{-n} \\ &= \Gamma^{-1} \sum_{j=1}^{j(n)} \gamma(j) P(x_1^n|F_j) \\ &\quad + \left(1 - \frac{\Gamma_{j(n)}}{\Gamma}\right) \alpha^{-n} = P(x_1^n). \end{aligned} \tag{61}$$

Consequently, $P_{\mathcal{M}}(x_{i+1}|x_1^i)$ is indeed a probability measure for every $i \geq 0$, allocated by a scheme $\mathcal{M}$.

Next, we prove that $\mathcal{M}$ is universal in the class of FS schemes. By Lemma 5, it suffices to prove that it is universal w.r.t. every given $F_l = (S, k, f, z_0) = F \in \Psi$. For all sufficiently large n, we have $k < j(n)$. In addition, $\Gamma \geq \Gamma_\infty > \Gamma_{j(n)}$ for every $n \geq 0$. Hence, denoting $\gamma = \gamma(l)$, for all sufficiently large n and every $x_1^n \in A^n$ we have, by (55), (56), and (57),

$$\begin{aligned} P(x_1^n) &> \frac{\gamma}{\Gamma} P(x_1^n|F) = \frac{\gamma}{\Gamma} \alpha^{-k} \prod_{i=k}^{n-1} \theta_i'(x_{i+1}|z_i,\ F) \\ &> \frac{\gamma}{\Gamma} \alpha^{-k} \prod_{i=0}^{n-1} \theta_i'(x_{i+1}|z_i,\ F) = \frac{\gamma}{\Gamma} \alpha^{-k} P'(x_1^n|F) \end{aligned} \tag{62}$$

where the last equality follows from (41). Consequently, by (53) and (62),

$$\begin{aligned} -\frac{1}{n} \log P(x_1^n) &< \frac{1}{n} \log \frac{\Gamma}{\gamma} + \frac{k}{n} \log \alpha - \frac{1}{n} \log P'(x_1^n|F) \\ &\leq \hat{H}(x_1^n|F) + k(\alpha - 1) \frac{\log n}{2n} + O(n^{-1}). \end{aligned} \tag{63}$$

Therefore, $\mathcal{M}$ is universal w.r.t. F. Since F is arbitrary, the proof is complete. Q.E.D.

Note that a similar approach can be used with more restricted model families. Examples are Markov and finite-memory (FSMX) models, i.e., models where each state is determined by a bounded number of past symbols [20], [30]. If an upper-bound on the length of the states is known (so that $\boldsymbol{J}$ in (43) is a *given* finite subset of Ψ), the number of terms in the mixture can be made finite. In this case, an elegant recursive algorithm that computes the mixture without explicit enumeration of the machines has been proposed recently [32].

As we noted for the bound in Section II, we can reach optimality in a stronger sense, by considering in the mixture machines supported by strongly connected graphs with an arbitrary number E of edges. A machine with less than α outgoing edges at some state, assigns probability zero to some sequences, for which it does not contribute to the mixture. With this approach we achieve, whenever possible, a smaller model cost within each family, namely $(E - k)(\log n)/2n$.

V. Application to Other Models

The idea of creating a universal measure $P(\cdot)$ by using a countable mixture of models $\{F_j\}_{j \geq 1}$ which degenerates to a finite mixture, is applicable to other situations of interest. Consider the following two examples.

a) *"Piecewise stationary" data:* Suppose that the data is expected to have a "piecewise stationary" structure, namely, it can be divided into segments having different characteristics, but in each segment the structure is relatively "simple" in the sense that it can be compressed efficiently by a simple machine, say, a single-state machine ($k = 1$). An extreme example is $x_1^{28} = 0000000000011111111111111111$. In such cases, the natural approach would be to partition the data into two or more segments and to handle them differently. For instance, in the case of two segments, we wish to find a universal sequential scheme $\mathcal{M}$ that nearly attains $m\hat{H}(x_1^m|F) + (n - m)\hat{H}(x_{m+1}^n|F)$, where m is the length of the first segment in the best partition and F is a given k-state model whose assignments are "adapted" to the current segment as defined in (20). More precisely, a reasonable measure of the performance of a universal scheme w.r.t. x_1^n and to F, would be

$$\begin{aligned} \Delta(x_1^n, \mathcal{M}\|F) \triangleq n^{-1} \Big[&-\log P_{\mathcal{M}}(x_1^n) - \min_{1 \leq m \leq n} \\ &(m\hat{H}(x_1^m|F) + (n - m)\hat{H}(x_{m+1}^n|F)) \Big]. \end{aligned} \tag{64}$$

When data flows serially, the universal scheme $\mathcal{M}$ knows *a priori* neither the best parameters associated with each segment, nor the best boundary point m. Following the idea of (57), consider a prior $\Gamma_\infty^{-1}\gamma(\cdot)$ on $m \geq 1$. Note that even though the relevant subset $\boldsymbol{J}$ of Ψ in (57) was finite, we needed a nonuniform prior in order to cope with the infinite number of models, for $\boldsymbol{J}$ was unknown. Similarly, here m can take only $n + 1$ values but, since n is unknown (strongly sequential scheme) this prior is proposed in order to define a universal measure independent of the final value $i = n$, which is constructed by the mixture

$$\begin{aligned} P(x_1^i) = \Gamma^{-1} &\sum_{j=1}^{i-1} \gamma(j) P(x_1^j|F) P(x_{j+1}^i|F) \\ &+ (1 - \Gamma^{-1}\Gamma_{i-1}) P(x_1^i|F) \end{aligned} \tag{65}$$

where $\Gamma \geq \Gamma_\infty$. This measure clearly satisfies the marginality condition (3), thus yielding $P_{\mathcal{M}}(x_{i+1}|x_1^i)$, defined as in (58). This can be interpreted as a mixture of measures with a prior on m given by $\Gamma^{-1}\gamma(m)$. Thus, with probability $1 - \Gamma_{i-1}\Gamma^{-1}$, m might be as large as i, which means that no transition occurs in the first i symbols. It is easy to see, using the same technique as in the proof of Theorem 2, that the resulting redundancy is

upper-bounded for every x_1^n as

$$\Delta(x_1^n, \mathcal{M}\|F) \le 2\frac{k(\alpha-1)}{2}\cdot\frac{\log n}{n} + \frac{\log n}{n} + O(n^{-1}) = [k(\alpha-1)+1]\frac{\log n}{n} + O(n^{-1}), \quad (66)$$

where each segment contributes an $0.5[k(\alpha-1)](\log n)/n$ term and an extra $(\log n)/n$ term is due to the unknown boundary point m. It has been shown in [15], where this problem has been studied in a probabilistic setting (i.e., coding for piecewise stationary information sources), that this is essentially the minimum achievable *expected* redundancy. *A fortiori*, it is a lower bound on the *minimax* redundancy $\max_{x_1^n\in A^n}\Delta(x_1^n, \mathcal{M}\|F)$ corresponding to the setting of [27]. Finally, if we also desire a doubly-universal scheme (i.e., F is not specified), another mixture, this time on F as proposed in (57), is needed.

b) *Block coding:* Suppose we want to design a strongly sequential, universal code that competes with the family of l-length block encoders, where l is unspecified, i.e., the family of uniquely decodable schemes that map input blocks of (unknown) length l to variable length codewords. These are single-state encoders defined on the super-alphabet of l-tuples. The comparison basis is similar to the one used with the FS family. Note that any such encoder can be simulated by an FS scheme with the same number of free parameters [24, Theorem 1] and, hence, the universal code of Section IV can successfully compete with this family, but there might be situations where the "natural" model for the data is an l-extension, and hence computing the mixture (57) would be unnecessarily costly.

Proceeding as in Section II, one can readily show, using Kraft's and Gibb's inequalities, that for a fixed l the best code in the family assigns to x_1^n a per-symbol code length that equals the normalized lth-order empirical entropy $\hat{H}_l(x_1^n)$, defined by the relative frequency of nonoverlapping l-tuples (we assume that l divides n). Denoting by $L_C(x_1^n)$ the length assigned to x_1^n by any code C, we can proceed as in Section II to show that for any given $\varepsilon > 0$, any code C, any block length l, and "most" sequences x_1^n we have, for all sufficiently large n,

$$n^{-1}L_C(x_1^n) - \hat{H}_l(x_1^n) \ge (\alpha^l - 1 - \varepsilon)\frac{\log n}{2n} \quad (67)$$

where the term "most" is defined as in Theorem 1, with types being defined w.r.t. the lth extension of A. Thus, as in Lemma 5, a code C is universal w.r.t. the family of block codes, if for every block length l and all sufficiently large multiples n of l, it satisfies

$$\max_{x_1^n\in A^n}[n^{-1}L_C(x_1^n) - \hat{H}_l(x_1^n)] \le (\alpha^l - 1)\frac{\log n}{2n} + O(n^{-1}). \quad (68)$$

Arguing as in Sections IV and V a), a strongly sequential, universal code for the family of block models is obtained as follows. For each block length l and every multiple i of l, define the probability measure

$$P(x_1^i|l) \triangleq \prod_{j=0}^{(i/l)-1}\eta_{jl}(x_{jl+1}^{(j+1)l}|l) \quad (69)$$

where $\eta_{jl}(y|l)$ is defined, for every l-tuple y that occurred $\mu_{jl}(y)$ times at nonoverlapping phases $ml+1$, $0 \le m < j$, in x_1^{jl}, as

$$\eta_{jl}(y|l) \triangleq \frac{\mu_{jl}(y) + 1/2}{j + \alpha^l/2}. \quad (70)$$

For values of i that are not multiples of l, define $r = \lfloor i/l \rfloor$, and

$$P(x_1^i|l) \triangleq P(x_1^{lr}|l)\sum_{y\in A^{l(r+1)-i}}\eta_{lr}(x_{lr+1}^i y|l). \quad (71)$$

It can be readily seen that, for each l, $P(\cdot|l)$ satisfies the marginality condition (3). Now, for every $i \ge 1$, define

$$P(x_1^i) \triangleq \Gamma^{-1}\sum_{l=1}^{i-1}\gamma(l)P(x_1^i|l) + (1 - \Gamma^{-1}\Gamma_{i-1})\alpha^{-i} \quad (72)$$

where $\gamma(\cdot)$, Γ_i, and Γ, are defined as in Section IV. This probability measure also satisfies (3). Thus, the code length $-\log P(x_1^i)$ defines a strongly sequential, regular code. The code is universal w.r.t. the family of block models in the sense that for all sufficiently large multiples n of l, the upper bound (68) holds. Note that although the universal code is defined for *every* sequence length, its universality w.r.t. a block length l is well-defined only for lengths that are a multiple of l. As in the previous cases, this can be interpreted as a mixture of measures with a prior $\Gamma^{-1}\gamma(l)$ on l. With probability $1 - \Gamma^{-1}\Gamma_{i-1}$, l might be as large as i, with the resulting probability α^{-i} for the sequence, as follows from (71).

Finally, we point out that these ideas can be easily generalized to the case where not only the size l of the best alphabet extension is unknown, but also the choice of an optimal phase is allowed. In this case, for each l, the probability $P(x_1^i|l)$ would involve an additional mixture over all possible phases.

APPENDIX A
A LOWER BOUND ON THE SIZE OF AN FS-TYPE

Our lower bound on $|T|$ is based on Whittle's formula [31] for the size of an FS-type. To state this formula we need some further notation. For a type T, let Φ^T denote a $k \times k$ matrix whose rows and columns are labeled by the states in S, and such that for any pair (s, z) of states, the entry ϕ_{sz}^T is the number of transitions from s to z in T. Note that at most $k\alpha$ entries in Φ^T are nonzero, and that the row-sum corresponding to a state z is $\mu_n(z)$, while the sum of all the entries is n. Moreover, the difference between the row-sum and the column-sum for z is $\delta_{zz_0} - \delta_{zz_n}$, where z_n denotes the final state, which is uniquely determined by Φ^T. Now, divide every nonzero row of Φ^T by its row-sum, and substract the resulting matrix from the $k \times k$ identity matrix I_k, thus obtaining a matrix denoted $\overline{\Phi}^T$. Finally, denote the (s, z)-cofactor of $\overline{\Phi}^T$

by $\overline{\Phi}^T(s, z)$, $s, z \in S$. Whittle's formula states that

$$|T| = \overline{\Phi}^T(z_n, z_0) \cdot \prod_{z \in S} \frac{\mu_n(z)!}{\prod_{a \in A} \mu_n(za)!}. \quad \text{(A.1)}$$

An asymptotic lower bound on (A.1) for every FS-type T is given in [3], but it is not tight enough for our purposes. Instead, we use Lemma 3, that provides a tighter bound that holds for most types.

In our way to the proof of Lemma 3, we notice that if all the rows of Φ^T are nonzero, then $\overline{\Phi}^T$ takes the form $I_k - \Phi$, where Φ is a stochastic matrix. The following lemma applies to this case.

Lemma A.1: Let Φ denote a $k \times k$ stochastic matrix such that each column has an off-diagonal nonzero entry. Then, all the cofactors of $I_k - \Phi$ are nonzero.

We notice that, by Whittle's formula, the lemma holds whenever Φ corresponds to an actual FS-type, for otherwise the type would be empty. However, our bounding technique in the proof of Lemma 3 requires that this hold for any stochastic matrix satisfying the conditions of Lemma A.1.

Proof of Lemma A.1: Let $\overline{\Phi} \triangleq I_k - \Phi$. First, note that $\det \overline{\Phi} = 0$ (1 is an eigenvalue for every stochastic matrix). Next, suppose conversely that the (i, j)-cofactor $\overline{\Phi}(i, j)$ of $\overline{\Phi}$ is also zero. Then, expanding $\det \overline{\Phi}$ by its ith row, we conclude that it is independent of the (i, j)th entry $\overline{\phi}_{ij}$ of $\overline{\Phi}$. Thus, we can change the value of $\overline{\phi}_{ij}$ without affecting neither the determinant, nor the cofactors of the entries in the ith row. Replace it by any arbitrary value. Furthermore, both the determinant and the (i, i)-cofactor remain unchanged if we also replace the ith column by the sum of all the columns. Since the sum of the entries in each row of $\overline{\Phi}$ is zero, except for the ith row, where $\overline{\phi}_{ij}$ has been replaced by a different value and hence the new row-sum is some $\sigma \neq 0$, we can expand the unchanged $\det \overline{\Phi}$ by the (new) ith column, thus obtaining $\sigma \cdot \overline{\Phi}(i, i)$. It follows that $\overline{\Phi}(i, i)$ must also be zero. (In fact, the same can be shown for all the cofactors of the ith row.)

Now, let Φ^i denote the matrix obtained from Φ by deleting the ith row and the ith column. By definition,

$$0 = \overline{\Phi}(i, i) = \det \overline{\Phi}^i \quad \text{(A.2)}$$

where $\overline{\Phi}^i \triangleq I_{k-1} - \Phi^i$. Since Φ^i is a substochastic matrix, the entries $\overline{\phi}^i_{lj}$, $0 < l, j < k$ of $\overline{\Phi}^i$ satisfy

$$|\overline{\phi}^i_{ll}| \geq \sum_{\substack{1 \leq j \leq k-1 \\ j \neq l}} |\overline{\phi}^i_{lj}| \quad \text{(A.3)}$$

for every l (note that (A.3) is satisfied with equality for every l only if Φ^i is also a stochastic matrix). Since by our assumptions some off-diagonal entry in the removed column is positive, Φ^i is not stochastic and, hence, strict inequality holds in (A.3) for at least one row l. Now, an extension by Hadamard of a theorem by Lévy [2, p. 69], states that this is a sufficient condition for a determinant to be nonzero. Hence, we have a contradiction. Q.E.D.

Proof of Lemma 3: First, we lower-bound the cofactor in Whittle's formula. As an auxiliary step, we consider FS-types for which

$$\mu_n(za) \geq \mu_n(z)\delta \neq 0 \quad \text{(A.4)}$$

for every $z \in S$, every $a \in A$, and some constant $\delta > 0$. For these types, $\overline{\Phi}^T$ takes the form $I_k - \Phi$, for some stochastic matrix Φ. Furthermore, each entry of Φ corresponding to an edge of F is at least δ, which, in particular, by the strong connectivity of F, ensures that Φ satisfies the condition of Lemma A.1. Now, consider each cofactor of the entries of $I_k - \Phi$ as a continuous function of the nonzero entries of Φ, which by (A.4) belong to a compact subset Δ of the space $(0, 1)^{k\alpha}$. By Lemma A.1, the sign of these k^2 functions is constant over Δ and, by Whittle's formula, must be positive. By the compactness of Δ, each function attains a positive minimum, yielding a unique minimum value $g_F(\delta)$, independent of n, over all the functions. Thus, the function $g_F(\delta)$, which uniformly lower-bounds all the cofactors over all the matrices $\overline{\Phi}^T$, possesses the following properties: it is continuous, nondecreasing, positive for every $\delta > 0$, and $\lim_{\delta \to 0} g_F(\delta) = 0$, since with $\Phi = I_k$ all the cofactors are clearly zero. It follows that $g_F(\delta)$ is strictly increasing for sufficiently small values of δ, for otherwise it would be zero in a neighborhood of 0. Consequently, it has an inverse $g_F^{-1}(x)$ which tends to 0 when x tends to 0.

Next, given $\varepsilon > 0$, let $\delta_{\varepsilon, F}(n) \triangleq g^{-1}(n^{-\varepsilon/2})$. Clearly,

$$\lim_{n \to \infty} \delta_{\varepsilon, F}(n) = 0. \quad \text{(A.5)}$$

By the definition of $g(\cdot)$, for every FS-type T satisfying (27) for every $z \in S$ and every $a \in A$, we have

$$\overline{\Phi}^T(z_n, z_0) \geq g(g^{-1}(n^{-\varepsilon/2})) = n^{-\varepsilon/2}. \quad \text{(A.6)}$$

As for the other terms in Whittle's formula we have, by Stirling's inequalities,

$$\begin{aligned}
&\log \prod_{z \in S} \frac{\mu_n(z)!}{\prod_{a \in A} \mu_n(za)!} \\
&\quad > n\hat{H}(x_1^n|F) + \frac{1}{2}\sum_{z \in S} \log \frac{2\pi\mu_n(z)}{\prod_{a \in A} 2\pi[\mu_n(za) + 1]} \\
&\quad \geq n\hat{H}(x_1^n|F) + \frac{1}{2}\sum_{z \in S} \log \frac{2\pi\mu_n(z)}{[2\pi\mu_n(z)]^\alpha} \\
&\quad = n\hat{H}(x_1^n|F) - \frac{(\alpha - 1)}{2}\sum_{z \in S} \log 2\pi\mu_n(z) \\
&\quad \geq n\hat{H}(x_1^n|F) - \frac{k(\alpha - 1)}{2} \log 2\pi n \\
&\quad > n\hat{H}(x_1^n|F) - \frac{k(\alpha - 1) + \varepsilon}{2} \log n \quad \text{(A.7)}
\end{aligned}$$

where the last inequality holds for sufficiently large n. By (A.1) and (A.5)–(A.7), the proof is complete. Q.E.D.

APPENDIX B
PROOF OF LEMMA 4

By Lemma 2, and since $\delta(\cdot)$ is a vanishing function of n, (29) will follow if we can upper-bound $N(\delta)$ as

$$N(\delta) \leq R\delta(n)(n+1)^{(\alpha-1)k} \tag{A.8}$$

for some constant R. Since $\mu_n(z) \leq n$ for every $z \in S$, it suffices to show that the number $N'(\delta)$ of types such that $\mu_n(za) \leq n\delta(n)$ for every $z \in S$ and every $a \in A$, is upper-bounded by the right-hand side of (A.8). Now, given a type, each edge of the graph of F (which we also denote by F), has an associated transition count, namely the number of transitions corresponding to this edge for that type. Let $N_e(\delta)$ denote the number of types with given final state z_n and such that a fixed edge e has an associated transition count not larger than $n\delta(n)$. Clearly, using a union bound where we let z_n range over S and e range over the set of edges, we obtain

$$N'(\delta) \leq k^2 \alpha N_e(\delta). \tag{A.9}$$

Thus, a sufficient condition for the desired upper bound on $N'(\delta)$, is

$$N_e(\delta) \leq \delta(n)(n+1)^{(\alpha-1)/k} \tag{A.10}$$

for any edge e of F and any $z_n \in S$. By a simple counting argument, (A.10) will, in turn, follow, if we show that given n and the final state of a type, we have $k(\alpha-1)$ degrees of freedom in the choice of the entries of the corresponding matrix Φ^T, and that any fixed entry (corresponding to the edge e) can be considered as one of the free parameters, so that a factor $n\delta(n)$ appears in the upper bound. In other words, we need to prove that there exists a subset E of k edges in F, with $e \notin E$, whose associated counts (for any type) are uniquely determined by the remaining $k(\alpha-1)$ counts, that correspond to the set $\overline{E}$ of the other edges of F. Next, we demonstrate one such set E or, equivalently, a partial graph G of F with k vertices (the same as F) and k edges, which are exactly the members of E.

In our way to G, we first delete e from F (since, by definition, e is not in G). Since F is strongly connected, the resulting graph is (at least) connected. Hence, it possesses a partial graph which is a tree [1, p. 153]. Add to this tree an additional edge from F, different from e, thus obtaining a partial graph of F, with k vertices and k edges. This is our graph G, so that it remains to prove that it has the desired property, namely that the transition counts corresponding to its edges are uniquely determined by the $k(\alpha-1)$ counts corresponding to the edges in $\overline{E}$, for any type. Since G is connected and has as many vertices as edges, it is either a single cycle, or has a vertex with only one incident edge e' (such a vertex is termed *pendant*, [1, p. 152]). In the latter case, since the difference between the row-sum and the column-sum of Φ^T for any state z is $\delta_{zz_0} - \delta_{zz_n}$, we have a linear equation that uniquely determines the count corresponding to e' from the counts corresponding to the edges in $\overline{E}$ incident at the pendant vertex. Remove e', whose count has been determined, and the corresponding pendant vertex from G, thus obtaining a new connected graph with $k-1$ edges and $k-1$ vertices. We can recursively continue with this process, until we end up with a cycle. Now, consider any edge in the final cycle, outgoing from state $z \in S$ and labeled with a symbol $a \in A$. Had the corresponding count $\mu_n(za)$ been determined, one could have determined the remaining counts in the cycle by linear operations, as described for the pendant vertices. Thus, there is a linear equation relating $\mu_n(za)$ with the counts that are still to be determined. Furthermore, the sum of all the counts associated with the edges in F is n. Hence, there is a linear equation that either determines $\mu_n(za)$ (in case the counts are compatible), or has no solution. Assuming that the counts correspond to an actual type, they must be compatible, which proves our claim. The proof is complete. Q.E.D.

APPENDIX C
THE FAILURE OF THE PLUG-IN APPROACH

The counterexample sequences are constructed in a simple case where only zero- and first-order Markov models are considered. The former is a single state machine, that assigns to each symbol some constant probability distribution, while for the latter the state at time i is x_i, and so it can assign one of α probability distributions to x_{i+1}. Let $\hat{H}_m(x_1^i)$ denote the empirical conditional entropy of x_1^i w.r.t. model order m. Thus, the estimator (47) takes the form

$$\begin{aligned} &\text{Estimate model order 0 if } \hat{H}_0(x_1^i) - \hat{H}_1(x_1^i) < \nu(i), \\ &\text{Estimate model order 1 if } \hat{H}_0(x_1^i) - \hat{H}_1(x_1^i) \geq \nu(i). \end{aligned} \tag{A.11}$$

We further consider a binary alphabet, for which the MDL uses $\nu(i) = 0.5i^{-1}\log i$. Now, if

$$\lim_{i\to\infty} \frac{2i\nu(i)}{\log i} \neq 1, \tag{A.12}$$

then either

$$\nu(i) - \frac{\log i}{2i} > C\frac{\log i}{2i} \tag{A.13}$$

for some positive C, in case $\nu(i)$ dominates $0.5i^{-1}\log i$, or

$$\frac{\log i}{2i} - \nu(i) > c\frac{\log i}{2i} \tag{A.14}$$

for some $0 < c < 1$, whenever $\nu(i)$ is dominated by $0.5i^{-1}\log i$.

For both cases where (A.12) holds, the counterexample sequences constructed in [12] satisfy $\hat{H}_0(x_1^i) = 1$ and $\hat{H}_1(x_1^i) = 1 - \delta(i)$ up to an $O(i^{-1})$ term, where $\delta(i)$ lies between $\nu(i)$ and $(\log i)/2i$ and, for large i, is away from both by at least an $O(i^{-1}\log i)$ term. The resulting code length is

$$l^{PI}(x_1^n) \geq \min_{m=0,1}\left\{\hat{H}_m(x_1^n) + 2^m\frac{\log n}{2n}\right\} + K_1\frac{\log n}{2n} \tag{A.15}$$

for large enough n, where K_1 is a positive constant. More precisely, in case (A.13) holds, the MDL of the entire sequence is attained by model order 1, and its value is $\hat{H}_1(x_1^n) + (2\log n)/2n$, while the plug-in scheme that uses $\nu(\cdot)$ as a penalty term always estimates an order 0, thus leading to a per-symbol code length

$$l^{PI}(x_1^n) = \hat{H}_0(x_1^n) + \frac{\log n}{2n} = \hat{H}_1(x_1^n) + \delta(n) + \frac{\log n}{2n} \tag{A.16}$$

(within an $O(n^{-1})$ term). In the case where (A.14) holds, the MDL of the entire sequence is attained by model order 0 and its value is $\hat{H}_0(x_1^n) + (\log n)/2n$, while a plug-in scheme with penalty term $\nu(\cdot)$ always estimates an order 1, with a resulting per-symbol code length

$$l^{PI}(x_1^n) = \hat{H}_1(x_1^n) + \frac{2\log n}{2n} = \hat{H}_0(x_1^n) - \delta(n) + \frac{\log n}{n}, \tag{A.17}$$

again, within an $O(n^{-1})$ term.

Finally, when (A.12) does not hold, (namely, for the asymptotic MDL estimator), the counterexample sequence constructed in [12] satisfies $\hat{H}_0(x_1^i) = 1$ and $\hat{H}_1(x_1^i) = 1 - \nu(i)$ up to an $O(i^{-1})$ term. Furthermore, for this sequence $\hat{H}_0(x_1^i) - \hat{H}_1(x_1^i)$ *fluctuates* around $\nu(i)$ so that about half of the time $\hat{H}_0(x_1^i) \geq \hat{H}_1(x_1^i) + \nu(i)$, and about half of the time $\hat{H}_0(x_1^i) \leq \hat{H}_1(x_1^i) + \nu(i)$. In addition, in the first case, i.e., whenever order 1 is estimated, the next symbol is such that the probability assigned by this model is smaller than the one assigned by a zero-order one, and vice versa when order 0 is chosen. Now, given the entire sequence x_1^n, the MDL is attained by an order m which is either 0 or 1. Since the sequential probability assignment (42) associated with order m satisfies (40), the MDL of x_1^n is the corresponding code length. On the other hand, as stated above, both models are estimated alternately by the plug-in scheme. It can be shown that whenever the estimated order differs from m, there is an extra code length of $O(\sqrt{(\ln i)/i})$ (see [12]). This overhead is contributed alternately, about half the time along the sequence, and hence the per-symbol excess code length above the MDL is about

$$\frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^{n} \sqrt{\frac{\ln i}{i}} \geq K_2 \sqrt{\frac{\log n}{n}}, \tag{A.18}$$

for some positive constant K_2. It follows that, for large enough n, the corresponding code length satisfies

$$l^{PI}(x_1^n) \geq \min_{m=0,1} \left\{ \hat{H}_m(x_1^n) + 2^m \frac{\log n}{2n} \right\} + K_2 \sqrt{\frac{\log n}{2n}}. \tag{A.19}$$

For further experiments regarding the sequential plug-in approach for coding, the reader is referred to [16].

In summary, the plug-in approach that assigns an asymptotically optimal code length in the probabilistic setting, does not attain the MDL lower bound (45) for each sequence in the deterministic setting, for any fixed vanishing penalty term.

ACKNOWLEDGMENT

The authors wish to thank N. Alon for providing the proof of Lemma 2.

REFERENCES

[1] C. Berge, *The Theory of Graphs and its Applications*. London, England: Methuen, 1962.
[2] E. Bodewig, *Matrix Calculus*, 2nd ed. Amsterdam, The Netherlands: North-Holland, 1959.
[3] L. B. Boza, "Asymptotically optimal tests for finite Markov chains," *Ann. Math. Statist.*, vol. 42, pp. 1992–2007, 1971.
[4] T. M. Cover, "On the competitive optimality of Huffman codes," *IEEE Trans. Inform. Theory*, vol. 37, pp. 172–174, Jan. 1991.
[5] I. Csiszár, Lecture Notes, Stanford University, Mar. 1993.
[6] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[7] ——, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 211–215, Mar. 1983.
[8] L. D. Davisson, R. J. McEliece, M. B. Pursely, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 269–279, May 1981.
[9] A. P. Dawid, "Present position and potential developments: Some personal views, statistical theory, the prequential approach," *J. Roy. Stat Soc., Ser. A*, vol. 147, part 2, pp. 278–292, 1984.
[10] M. Feder, "Gambling using a finite-state machine," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1459–1465, Sept. 1991.
[11] ——, "A note on the competitive optimality of the Huffman code," *IEEE Trans. Inform. Theory*, vol. 38, pp. 436–439, Mar. 1992.
[12] ——, "On the plug-in approach for sequential coding of individual sequences." Dep. EE–Syst., Tel Aviv Univ., Tech. Rep. EE-S-93-09, Mar. 1993.
[13] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.
[14] G. G. Langdon, Jr., "A note on the Lempel-Ziv model for compressing individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 284–287, Mar. 1983.
[15] N. Merhav, "On the minimum description length principle for sources with piecewise constant parameters," *IEEE Trans. Inform. Theory* vol. 39, pp. 1962–1967, Nov. 1993.
[16] R. Oz, "On sequential order estimation for universal coding of individual sequences." M.Sc. thesis, Dep. of EE–Syst., Tel-Aviv Univ., 1992 (in Hebrew).
[17] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel-Ziv algorithm," *IEEE Trans. Inform. Theory*, vol. 38, pp. 66–72, Jan. 1992.
[18] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.
[19] ——, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, pp. 416–431, 1983.
[20] ——, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 526–532, July 1986.
[21] ——, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, Sept. 1986.
[22] ——, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.
[23] ——, *Stochastic Complexity in Statistical Inquiry*. New Jersey: World Scientific, 1989.
[24] J. Rissanen and G. G. Langdon, Jr., "Universal modeling and coding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 12–23, Jan. 1981.
[25] B. Y. Ryabko, "Twice-universal coding," *Problems of Inform. Trans.*, vol. 20, pp. 173–177, July-Sept. 1984.
[26] ——, "Prediction of random sequences and universal coding," *Problems of Inform. Trans.*, vol. 24, pp. 87–96, Apr.-June 1988.
[27] Y. M. Shtar'kov, "Universal sequential coding of single messages," *Problems of Inform. Trans.*, vol. 23, pp. 175–186, July-Sept. 1987.
[28] R. J. Solomonoff, "A formal theory of inductive inference," *Inform. Contr.*, vol. 7, part I: pp. 1–22, part II: pp. 224–254, 1964.
[29] M. J. Weinberger and M. Feder, "Predictive stochastic complexity and model estimation for finite-state processes," to appear in *J. Statist. Planning and Inference*, vol. 39, pp. 353–372, May 1994.
[30] M. J. Weinberger, A. Lempel, and J. Ziv, "A sequential algorithm for the universal coding of finite-memory sources," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1002–1014, May 1992.
[31] P. Whittle, "Some distributions and moment formulae for the Markov chain," *J. Roy. Stat. Soc., Ser. B*, vol. 17, pp. 235–242, 1955.
[32] F. M. J. Willems, Y. M. Shtar'kov, and T. J. Tjalkens, "Context tree weighting A sequential universal source coding procedure for FSMX sources," in *Proc. 1993 IEEE Int. Symp. Inform. Theory*.
[33] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.