Strongly Consistent Code-Based Identification
and Order Estimation for Constrained
Finite-State Model Classes

John C. Kieffer

Abstract—Observations are made of data generated by a
stationary ergodic finite-alphabet information source according
to an unknown statistical model. Two modeling problems, the
identification problem and the order estimation problem, are
considered. In the identification problem, one wishes to decide
from the observed data whether the source model belongs to
a given model class. In the order estimation problem, one wishes
to decide from the observed data to which of infinitely many
given model classes the source model belongs. It is required that
the given model class in the identification problem and that
each given model class in the order estimation problem be a
constrained finite-state model class, which is a type of model
class that includes many model classes of information-theoretic
interest. Strongly consistent decision rules are exhibited in both
the identification problem and the order estimation problem.
The decision rules are code-based in that a model class is shown
based upon how well a certain code for that class encodes the
observed data. The code used for a model class is based upon
the maximum likelihood code for that class, and asymptotic code
performance is gauged by means of a key property of divergence-
rate distance.

Index Terms—Finite-state sources, order estimation, maximum
likelihood codes, divergence-rate distance.

I. INTRODUCTION

Consider a finite-alphabet information source which
generates a random sequence of symbols \(X_1, X_2, X_3, \ldots\).
We assume that this random sequence is stationary and er-

dotic. Suppose that the distribution \(\mu\) of the source sequence
\((X_i)\) is not known. By observing a large number of terms of
the source sequence, we may hope to model \(\mu\) as a member of
a certain type of model class, with a high degree of confidence.

In this paper, we consider two kinds of modeling problems
that arise in this context, identification problems and order
estimation problems. In an identification problem, one wishes
to learn whether or not \(\mu\) lies in a given model class; for
example, one may wish to know whether \(\mu\) is i.i.d. In an
order estimation problem, one is given a sequence of model
classes whose union contains \(\mu\) and wishes to determine which
of the classes contains \(\mu\); for example, one may know that \(\mu\) is
Markovian of finite unknown order and may wish to determine
the order. The following two subsections detail what shall
be our approach to identification problems and order estima-
tion problems.

A. Identification Problems

We consider again the source whose output \((X_i)\) has un-
known distribution \(\mu\). Let \(M\) be a model class consisting of
probability measures on the source alphabet sequence space.
The identification problem for the model class \(M\) would
involve the determination of a binary-valued decision rule
\(Y_n\) based on the first \(n\) source samples \((X_1, X_2, \ldots, X_n)\) in which
the decision that \(\mu \in M\) is made if \(Y_n = 1\) and the decision
that \(\mu \notin M\) is made if \(Y_n = 0\). We desire a strongly consistent
decision rule, which means that with probability one, the
correct decision is made for \(n\) sufficiently large. We also
require an information-theoretic solution to the identification
problem in that the decision as to whether \(\mu\) belongs to
\(M\) is based upon seeing which of two noiseless binary
codes encodes \((X_1, X_2, \ldots, X_n)\) using the smaller number of
bits. We call this approach to the identification problem code-

Based on the first

Based on the second

Based on the third

Based on the fourth

Based on the fifth

Based on the sixth

Based on the seventh

Based on the eighth

Based on the ninth

Based on the tenth

Based on the eleventh

Based on the twelfth

Based on the thirteenth

Based on the fourteenth

Based on the fifteenth

Based on the sixteenth

Based on the seventeenth

Based on the eighteenth

Based on the nineteenth

Based on the twentieth

Based on the twenty-first

Based on the twenty-second

Based on the twenty-third

Based on the twenty-fourth

Based on the twenty-fifth

Based on the twenty-sixth

Based on the twenty-seventh

Based on the twenty-eighth

Based on the twenty-ninth

Based on the thirty-first

Based on the thirty-second

Based on the thirty-third

Based on the thirty-fourth

Based on the thirty-fifth

Based on the thirty-sixth

Based on the thirty-seventh

Based on the thirty-eighth

Based on the thirty-ninth

Based on the forty-first

Based on the forty-second

Based on the forty-third

Based on the forty-fourth

Based on the forty-fifth

Based on the forty-sixth

Based on the forty-seventh

Based on the forty-eighth

Based on the forty-ninth

Based on the fifty-first

Based on the fifty-second

Based on the fifty-first

Based on the fifty-second

Based on the fifty-third

Based on the fifty-fourth

Based on the fifty-fifth

Based on the fifty-sixth

Based on the fifty-seventh

Based on the fifty-eighth

Based on the fifty-ninth

Based on the sixty-first

Based on the sixty-second

Based on the sixty-first

Based on the sixty-second

Based on the sixty-third

Based on the sixty-fourth

Based on the sixty-fifth

Based on the sixty-sixth

Based on the sixty-seventh

Based on the sixty-eighth

Based on the sixty-ninth

Based on the seventy-first

Based on the seventy-second

Based on the seventy-first

Based on the seventy-second

Based on the seventy-third

Based on the seventy-fourth

Based on the seventy-fifth

Based on the seventy-sixth

Based on the seventy-seventh

Based on the seventy-eighth

Based on the seventy-ninth

Based on the eighty-first

Based on the eighty-second

Based on the eighty-first

Based on the eighty-second

Based on the eighty-third

Based on the eighty-fourth

Based on the eighty-fifth

Based on the eighty-sixth

Based on the eighty-seventh

Based on the eighty-eighth

Based on the eighty-ninth

Based on the ninety-first

Based on the ninety-second

Based on the ninety-first

Based on the ninety-second

Based on the ninety-third

Based on the ninety-fourth

Based on the ninety-fifth

Based on the ninety-sixth

Based on the ninety-seventh

Based on the ninety-eighth

Based on the ninety-ninth

Based on the one-hundred-first

Based on the one-hundred-second

Based on the one-hundred-first

Based on the one-hundred-second

Based on the one-hundred-third

Based on the one-hundred-fourth

Based on the one-hundred-fifth

Based on the one-hundred-sixth

Based on the one-hundred-seventh

Based on the one-hundred-eighth

Based on the one-hundred-ninth

Based on the one-hundred-twentieth

Based on the one-hundred-twenty-first

Based on the one-hundred-twenty-second

Based on the one-hundred-twenty-third

Based on the one-hundred-twenty-fourth

Based on the one-hundred-twenty-fifth

Based on the one-hundred-twenty-sixth

Based on the one-hundred-twenty-seventh

Based on the one-hundred-twenty-eighth

Based on the one-hundred-twenty-ninth

Based on the one-hundred-thirty-first

Based on the one-hundred-thirty-second

Based on the one-hundred-thirty-first

Based on the one-hundred-thirty-second

Based on the one-hundred-thirty-third

Based on the one-hundred-thirty-fourth

Based on the one-hundred-thirty-fifth

Based on the one-hundred-thirty-sixth

Based on the one-hundred-thirty-seventh

Based on the one-hundred-thirty-eighth

Based on the one-hundred-thirty-ninth

Based on the one-hundred-thirty-twentieth

Based on the one-hundred-thirty-twenty-first

Based on the one-hundred-thirty-twenty-second

Based on the one-hundred-thirty-twenty-third

Based on the one-hundred-thirty-twenty-fourth

Based on the one-hundred-thirty-twenty-fifth

Based on the one-hundred-thirty-twenty-sixth

Based on the one-hundred-thirty-twenty-seventh

Based on the one-hundred-thirty-twenty-eighth

Based on the one-hundred-thirty-twenty-ninth

Based on the one-hundred-thirty-thirtieth

Based on the one-hundred-thirty-first

Based on the one-hundred-thirty-second

Based on the one-hundred-thirty-first

Based on the one-hundred-thirty-second

Based on the one-hundred-thirty-third

Based on the one-hundred-thirty-fourth

Based on the one-hundred-thirty-fifth

Based on the one-hundred-thirty-sixth

Based on the one-hundred-thirty-seventh

Based on the one-hundred-thirty-eighth

Based on the one-hundred-thirty-ninth
the identification problem for constrained finite-state model classes. We achieve this result in Theorem 1 of Section II.

B. Order Estimation Problems

Consider again the source whose output \((X_i)\) has unknown distribution \(\mu\). Let \(M^1, M^2, M^3, \ldots\) be an increasing sequence of model classes, each consisting of measures on the source alphabet sequence space. Suppose that one knows that \(\mu\) is a member of one of the classes \(\{M^i\}\), but does not know which \(i\) for which \(\mu \in M^i\). In the order estimation problem for the classes \(\{M^i\}\), one wishes to estimate the order of the source relative to the classes \(\{M^i\}\) for large \(n\). We desire that our order estimates be strongly consistent as \(n \to \infty\), and we require that the estimates be code-based in that the order estimate based on the sample \((X_1, X_2, \ldots, X_n)\) is \(i\), if and only if the code \(\phi_i\) in some sequence of codes \(\{\phi_1, \phi_2, \ldots\}\) assigns to \((X_1, X_2, \ldots, X_n)\) the codeword of shortest length.

One widely used code-based approach to order estimation problems has been popularized by J. Rissanen [15]–[20]. In this approach, the codes \(\{\phi_i\}\) are selected according to the minimum description length (MDL) criterion. In the code-based approach employed in this paper, we select the codes \(\{\phi_i\}\) in a different way, so as to ensure strong consistency of the resulting sequence of order estimators. In Section II, we discuss the difference in structure between the codes we employ and the codes employed by Rissanen.

We investigate the order estimation problem for classes \(\{M^i\}\) in which all the \(M^i\) are constrained finite-state model classes on the same finite source alphabet. We exhibit a strongly consistent sequence of order estimators for this problem in Theorem 2 of Section II. Special cases of this problem had earlier been considered. Tong [24] derived a sequence of order estimators using Akaike’s information criterion [1] for the order estimation problem in which \(M^i\) is the class of \(i\)-th order Markov models \((i = 1, 2, \ldots)\). However, Tong did not show strong consistency. Merhav et al. [14] considered the same classes \(\{M^i\}\) as Tong, but arrived at a sequence of order estimators different from that of Tong. The Merhav et al. order estimators employ the codeword lengths in the Lempel–Ziv data compression algorithm. Ziv and Merhav [27] extended the approach in [14] to the case in which \(M^i\) is the class of all finite-state source models with \(i\) states \((i = 1, 2, \ldots)\).

The order estimators found in the papers [14], [27] are not consistent; however, Liu and Narayan [13] showed how these order estimators could be slightly modified to achieve strong consistency. Liu and Narayan [13] also provided their own strongly consistent order estimators in the case in which \(M^i\) is the class of all \(i\)-state hidden Markov models \((i = 1, 2, \ldots)\).

There is an extensive literature on order estimation problems, and it is not our purpose here to survey all of the order estimation results that have been obtained. In the previous paragraph, we mentioned those results that were relevant to the result on order estimation presented in this paper. In Section II, we cite some other results on order estimation. The reader may find numerous other publications on order estimation cited in the bibliography of the recent monograph [20].

The rest of the paper is organized as follows. In Section II, we give the necessary background so that our results on identification and order estimation can be formally stated. In Sections III and IV, we give a treatment of the concept of divergence-rate distance and of the concept of the rate profile of a code (respectively), which are key concepts in our investigation of identification problems and order estimation problems. In Section V, which concludes the paper, we give the proofs of our identification and order estimation results.

II. STATEMENT OF PRINCIPAL RESULTS

We present several subsections which give necessary background, finishing with a subsection in which we state our results on identification problems and order estimation problems.

A. Generic Notations

We point out notational conventions that shall be in force throughout the paper. The notation \(\mathbb{N}\) denotes the set of positive integers. If \(j \in \mathbb{N}\), then \(I_j\) denotes the set \(\{1, 2, \ldots, j\}\). Let \(S\) be a finite set. Then \(|S|\) denotes the cardinality of \(S\), \(S^n (n \in \mathbb{N})\) denotes the set of all \(n\)-tuples that can be formed from the elements of \(S\), and \(S^*\) denotes the set of all strings of finite length that can be formed from the elements of \(S\) (i.e., \(S^* = \bigcup_n S^n\)).

All logarithms shall be to base two. Also, the notation \(\exp_{2q}(Q)\) denotes \(2^Q\).

If \(x\) is a real number, then \([x]\) shall denote the smallest positive integer \(\geq x\). Finally, the notation \(\arg \min_i g(i)\), where \(g = g(i)\) is a positive-integral valued function of a positive-integral valued variable \(i\), denotes that we take the smallest \(i\) at which the minimum of \(q\) is achieved.

B. Finite Alphabet Sequence Space

We fix throughout the paper a finite symbol set \(A\) of size \(|A| \geq 2\), which will serve as the alphabet for all the information sources that we consider. We let \(A^\infty\) denote the sequence space consisting of all infinite sequences \((x_1, x_2, \ldots)\) whose terms belong to \(A\). For each \(n \in \mathbb{N}\) and each \(x = (x_1, x_2, \ldots, x_n) \in A^n\), we let \([x]\) denote the finite-dimensional cylinder set consisting of all sequences in \(A^n\) whose first \(n\) terms are \(x_1, \ldots, x_n\), respectively. We then make \(A^\infty\) into the measurable space whose measurable sets are those sets which belong to the sigma-field spanned by the finite-dimensional cylinder sets \(\{[x] : x \in A^n, n \in \mathbb{N}\}\).

C. Measures on the Sequence Space

Let \(M(A)\) denote the set of all probability measures on the sequence space \(A^\infty\). If \(\mu \in M(A)\) and \(n \in \mathbb{N}\), we let \(\mu_n\) denote the probability measure on \(A^n\) in which \(\mu_n(x) = \mu([x]), x \in A^n\). The measures \(\{\mu_n\}\) can be described from the point of view of random sequences. Suppose \(X = (X_1, X_2, \ldots)\) is a \(A^\infty\)-valued random sequence. For each \(n \in \mathbb{N}\), let \(X^n\) denote the random vector \((X_1, X_2, \ldots, X_n)\).
consisting of the first \( n \) entries of \( X \). Then, if \( \mu \in M(A) \) is the
distribution of \( X \), the measure \( \mu_n \) is the distribution of \( X^n \).

If \( M \) is a subset of \( M(A) \) and \( n \in \mathbb{N} \), we let \( M_n \) denote the set \( \{ \lambda_n : \lambda \in M \} \) of probability measures on \( A^n \).

A sequence \( \{ \lambda_n \} \) of measures in \( M(A) \) is said to converge weakly to a measure \( \lambda \) in \( M(A) \) if \( \lambda_n(E) \) converges to \( \lambda(E) \) for every finite-dimensional cylinder subset \( E \) of \( A^n \).

We endow \( M(A) \) with the unique metrizable topology in
which convergence of sequences of measures is precisely weak
convergence. Since we take \( M(A) \) to have a topology, we
shall need the concept of the

**Information Sources**

For the purposes of this paper, an *information source* (or simply source) is a pair \( [X, \mu] \) in which \( X \) is a \( A^\infty \)-valued random sequence \( (X_1, X_2, X_3, \ldots) \) (called the source output), and \( \mu \in M(A) \) is the distribution of the source output \( X \).

We say that a source \( [X, \mu] \) is stationary if \( \mu \) is shift-invariant, and we say that the source \( [X, \mu] \) is ergodic if \( \mu \in M_e \).

In the solution to the identification problem that we present later (see Theorem 1), we shall need the concept of the *entropy-rate* of a source. If \( n \in \mathbb{N} \) and \( \lambda \) is a probability measure on \( A^n \), the entropy of \( \lambda \) is the number \( H(\lambda) \) defined by

\[
H(\lambda) = \sum_{x \in A^n} -\lambda(x) \log \lambda(x).
\]

If \( [X, \mu] \) is a stationary source, then its entropy-rate is the number \( H(\mu) \) defined by

\[
H(\mu) = \lim_{n \to \infty} n^{-1} H(\mu_n).
\]

(The preceding limit is known to exist [5, Theorem 4.2.1].)

**E. Constrained Finite-State Model Classes**

We now give a formal definition of the concept of constrained
finite-state model class. If \( j \in \mathbb{N} \), a *j*-state transition
function is a vector \( p = [p(u, x|u') : u, u' \in I_j, x \in A] \) in which

\[
p(u, x|u') \geq 0, \quad \forall u, u', x
\]

and

\[
\sum_{u', f} p(u, x|u') = 1, \quad \forall u'.
\]

We say that a set \( C \) of \( j \)-state transition functions is *closed* if, whenever \( \{ p_t : t \in \mathbb{N} \} \) is a sequence of \( j \)-state transition functions from \( C \) converging coordinate-wise to a \( j \)-state transition function \( p \), then \( p \in C \). (Saying that a set \( C \) of \( j \)-state transitions is closed is equivalent to saying that \( C \) is a compact subset of a Euclidean space, since every \( j \)-state transition function can be thought of as a point in a \( j^3 |A| \)-dimensional Euclidean space.) If \( q \) is a probability measure on \( I_j \times A \) and \( p \) is a \( j \)-state transition function, we define \( \mu^{q, p} \) to be the measure in \( M(A) \) such that, for every \( n \geq 2 \) and every \( (x_1, x_2, \ldots, x_n) \in A^n \),

\[
\mu^{q, p}([(x_1, x_2, \ldots, x_n)]) = \sum_{u_1, u_2, \ldots, u_n} q(u_1, x_1) p(u_2, x_2|u_1) p(u_3, x_3|u_2) \cdots p(u_n, x_n|u_{n-1}).
\]

For some \( j \in \mathbb{N} \), suppose \( C \) is a closed set of \( j \)-state transition functions. We let \( M(C) \) denote the model class \{ \( \mu^{q, p} : q \) a prob. dist. on \( I_j \times A, p \in C \) \}. The model classes \{ \( M(C) : C \) a closed set of \( j \)-state transition functions for some \( j \in \mathbb{N} \) \} shall be called *constrained finite-state model classes*.

**F. Codes**

For the purposes of this paper, we define a *code* \( \phi \) to be a mapping from \( A^* \) to the set of binary strings \( \{0, 1\}^* \) such that whenever \( a_1 \) and \( a_2 \) are two different strings in \( A^* \) of the same length, then \( \phi(a_1) \) is not a prefix of \( \phi(a_2) \). If \( b \) is a binary string, we let \( L[b] \) denote the length of the string \( b \).

We state two fundamental facts about codes [5, ch. 5] that we shall need.

**Fact 1**: Let \( \phi \) be a code. Then the function \( \sigma : A^* \to \mathbb{N} \) defined by

\[
\sigma(x) = L[\phi(x)], \quad x \in A^*
\]

satisfies

\[
\sum_{x \in A^n} \exp_2(-\sigma(x)) \leq 1, \quad x \in A^n, \quad n \in \mathbb{N}.
\]

**Fact 2**: If \( \sigma : A^* \to \mathbb{N} \) is a function such that (2.2) holds, then there is a code \( \phi \) such that (2.1) holds.

**G. Maximum Likelihood Codes**

The concept of maximum likelihood code was introduced by Shtrakov, who used this concept in the context of universal source coding [22], [23]. We shall use maximum likelihood codes in this paper in the context of the code-based identification problem for constrained finite-state model classes. In this subsection, we define the concept of maximum likelihood code.

Let \( M \) be any model class. We let \( \tau_M \) be the function defined on \( A^* \) by

\[
\tau_M(x) = \sup_{\nu \in M} \nu_n(x), \quad x \in A^n, \quad n \in \mathbb{N}.
\]
Furthermore, we let \( \tau_M(n) \) be the sequence of numbers defined by
\[
\tau_M(n) = \sum_{x \in A^n} \tau_M(x), \quad n \in \mathbb{N}.
\]
Using Fact 2, there must exist a code \( \phi_M \) such that for each \( n \in \mathbb{N} \) and each \( x \in A^n \),
\[
L[\phi_M(x)] = 1 + \lceil -\log(\tau_M(x)/\tau_2(n)) \rceil, \quad \tau_M(x) > 0
\]
\[
= 1 + \lceil 2n \log |A| \rceil, \quad \text{otherwise}.
\]
Any such code \( \phi_M \) shall be called a **maximum likelihood code** for the model class \( M \).

**H. Code-Based Identification**

Let \( M \) be an arbitrary model class. In our code-based approach to the identification problem for the model class \( M \), we seek a pair of codes \( \{ \phi, \psi \} \) such that
\[
\lim_{n \to \infty} \{ L[\psi(X^n)] - L[\phi(X^n)] \} = \infty \quad \text{a.s.},
\]
\( \forall \) ergodic source \( [X, \mu], \mu \in M \). (2.3)

\[
\lim_{n \to \infty} \{ L[\psi(X^n)] - L[\phi(X^n)] \} = -\infty \quad \text{a.s.},
\]
\( \forall \) ergodic source \( [X, \mu], \mu \notin M \). (2.4)

Given \( \phi, \psi \) satisfying (2.3)–(2.4), one easily obtains a decision rule that can be used in the identification problem for \( M \). This decision rule operates as follows: Having observed the first \( n \) source samples \( X_1, X_2, \ldots, X_n \), decide that the source is modeled by a measure in \( M \), if and only if \( L[\psi(X_1, \ldots, X_n)] \geq L[\phi(X_1, \ldots, X_n)] \). Then, from (2.3)–(2.4) it follows that, for any ergodic source \( [X, \mu] \)
\[
\Pr [\text{for } n \text{ suff. large, correct decision is made based on samples } X_1, \ldots, X_n] = 1. \quad (2.5)
\]
Statement (2.5) tells us that our decision rule is strongly consistent, as desired.

The first of the two main results of this paper (Theorem 1) tells us how to obtain codes satisfying (2.3)–(2.4) in the identification problem for a constrained finite-state model class. To achieve this result, one must employ a code \( \psi \) which satisfies the property
\[
\lim_{n \to \infty} n^{-1} L[\hat{\psi}(X^n)] = H(\mu) \quad \text{a.s.},
\]
\( \forall \) ergodic source \( [X, \mu] \). (2.6)
It is known that codes \( \hat{\psi} \) satisfying (2.6) exist [11]. Perhaps the best known code \( \hat{\psi} \) for which (2.6) holds is the Lempel–Ziv code [26].

We are now ready to state Theorem 1. (The proof is given in Section V.)

**Theorem 1:** Let \( M \) be a constrained finite-model class. Select a positive real number \( \beta \) (whose existence is guaranteed by Lemma A.2 of the Appendix) such that the sequence \( \{ \tau_M(n)n^{-\beta} : n \in \mathbb{N} \} \) is bounded. Let \( \alpha \) be a real number such that \( \alpha > \beta + 1 \). Fix a code \( \bar{\psi} \) for which (2.6) holds. Let \( \psi \) be any code for which
\[
L[\psi(x)] = L[\bar{\psi}(x)] + \lceil \alpha \log n \rceil, \quad x \in A^n, \quad n \in \mathbb{N}.
\]
Let \( \phi \) be the code \( \phi = \phi_M \). Then the pair of codes \( \{ \phi, \psi \} \) satisfies (2.3)–(2.4).

**Discussion:** The statistic \( L[\psi(X^n)] - L[\phi(X^n)] \) arising from the codes \( \{ \phi, \psi \} \) given in Theorem 1 is the statistic that one examines for large \( n \) in order to make one’s decision in the identification problem. If the code \( \psi \) is taken to be the Lempel–Ziv code, then it is easy to sequentially evaluate \( L[\psi(X^n)] \) in terms of the number of phrases in the Lempel–Ziv parsing of the sample \( X^n \). However, the evaluation of \( L[\phi(X^n)] \) can take a long time since it may be difficult to compute the maximum likelihood function \( \tau_M(X^n) \). To get around this difficulty, one can seek a computationally efficient sequential algorithm that will approximate \( \tau_M(X^n) \) sufficiently well for large \( n \) so that the resulting approximation to the statistic \( L[\psi(X^n)] - L[\phi(X^n)] \) will yield a strongly consistent decision rule in the identification problem for the constrained finite-state model class \( M \). It is an open problem to find such an algorithm.

**I. Code-Based Order Estimation**

Let \( M^1, M^2, \ldots \) be an increasing sequence of model classes. We define the order of a source \( [X, \mu] \) (relative to the classes \( \{ M^i \} \) to be the number \( \text{ord}(\mu) \) which is equal to the smallest \( i \) such that \( \mu \in M^i \). In our code-based approach to the order estimation problem for the model classes \( \{ M^i \} \), we seek a sequence of codes \( \phi_1, \phi_2, \ldots \) such that
\[
\arg \min_i L[\phi_i(X^n)] \to \text{ord}(\mu) \quad \text{a.s.},
\]
\( \forall \) ergodic source \( [X, \mu], \mu \in \bigcup_i M^i \). (2.7)

In Section V, we shall prove the following result on order estimation.

**Theorem 2:** Let \( \{ M^i : i \in \mathbb{N} \} \) be a sequence of constrained finite-state model classes in which \( M^1 \subset M^2 \subset M^3 \subset \cdots \). Choose positive integers \( \{ j_i : i \in \mathbb{N} \} \) such that
\[
j_{i+1} \geq 2(j_i + 1), \quad i \in \mathbb{N};
\]
\[
\tau_M(n) \leq n^{j_i}, \quad n \geq 2, \quad i \in \mathbb{N}. \quad (2.8)
\]
For each \( i \in \mathbb{N} \), let \( \phi_i \) be a code in which
\[
L[\psi_i(x)] = L[\phi_M(x)] + \lceil j_i \log n \rceil, \quad x \in A^n.
\]
Then the sequence of codes \( \{ \phi_i \} \) satisfies (2.7).

**Discussion:** In code-based approaches to the order estimation problem for a sequence of model classes \( \{ M^i \} \), one proposes a sequence of codes \( \{ \phi_i \} \) and then uses the statistic \( \arg \min_i L[\psi_i(X^n)] \) as the estimator of the unknown source order based on the source sample \( X^n \). For the codes \( \{ \phi_i \} \) that we selected in Theorem 2, the codeword length \( L[\phi_i(X^n)] \) takes the approximate form
\[
L[\psi_i(X^n)] \approx -\log \tau_M(X^n) + C_i \log n, \quad (2.10)
\]
for large $n$, where $\{C_i\}$ is a certain nondecreasing sequence of positive constants that is determined from the model classes $\{M_i\}$. Let $\{Y_n(X^n)\}$ be the sequence of statistics in which

$$Y_n(X^n) = \arg\min_n \{ - \log \tau_{M_i}(X^n) + C_l \log n \}, \quad \forall n.$$ 

The approximation of each $L[\phi_i(X^n)]$ by (2.10) is good enough for large $n$ so that $\{Y_n(X^n)\}$ can be shown to be strongly consistent for the order estimation problem of Theorem 2. In Rissanen’s code-based approach to order estimation for the classes $\{M_i\}$, the MDL criterion is imposed to select the code $\phi_i$ for the class $M_i$ whose expected redundancy grows at the smallest possible rate for “almost all” sources whose outputs are modeled by members of $M_i$ (see [20, Theorem 3.1] for a precise statement). As a consequence, if one writes the codeword length $L[\phi_i(X^n)]$ for the code $\phi_i$ as selected by the MDL criterion in the approximate form

$$L[\phi_i(X^n)] \approx - \log \tau_{M_i}(X^n) + D_i \log n,$$ 

for a sequence of positive constants $\{D_i\}$, it will follow that $D_i \leq C_l$ for all $i$. Making use of the approximation (2.11), one obtains MDL order estimators $\{Z_n(X^n)\}$ of the form

$$Z_n(X^n) = \arg\min_n \{ - \log \tau_{M_i}(X^n) + D_i \log n \}, \quad \forall n.$$ 

(It should be remarked that Schwarz [21], independently of Rissanen, arrived at precisely the same order estimators $Z_n(X^n)$ on the basis of purely Bayesian considerations.) As the estimator $Z_n(X^n)$ employs a penalty function $D_i \log n$ no larger than the penalty function $C_l \log n$ employed by the estimator $Y_n(X^n)$, it would be of interest to determine for which sequences $\{M_i\}$ of constrained finite-state model classes it is true that the MDL order estimators $\{Z_n(X^n)\}$ are consistent order estimators for every ergodic source $[X, \mu]$ in which $\mu$ is a member of one of the $\{M_i\}$.

We cite some results of this type that are known for other types of model classes $\{M_i\}$. Suppose we allow the underlying source alphabet to be infinite. We also allow only finitely many different $M_i$ and each $M_i$ must be of the form $M_i = \{ \mu : \theta \in R_i \}$ where $R_i$ is the closure of a connected open subset of a finite-dimensional Euclidean space (whose dimension depends on $i$), and the measure $\mu_\theta$ is i.i.d. and depends smoothly on the parameter $\theta$. From a result of Barron [4, Theorem 4.3] one can deduce that under certain additional conditions, there exists a prior distribution on $\cup_i M_i$ such that the MDL estimators $\{Z_n(X^n)\}$ are consistent estimators of the order of the source $[X, \mu]$ for almost every $\mu$ relative to this prior. Under the additional requirement that each $M_i$ be a family of exponential type, Haughton [9] has shown that $\{Z_n(X^n)\}$ is a consistent sequence of order estimators for every source $[X, \mu]$ in which $\mu \in \cup_i M_i$. If one makes the further requirement that each $M_i$ consist of Gaussian distributions, one can in the context of some applications that the MDL estimators $\{Z_n(X^n)\}$ are strongly consistent for every source $[X, \mu]$ in which $\mu \in \cup_i M_i$; for example, Zhao et al. [25] have a result of this type for the problem of the detection of the number of signals in the presence of Gaussian white noise.

### III. Investigations Into Divergence-Rate Distance

Our solution to the identification problem will involve a notion of distance from a measure to a model class called divergence-rate distance. It is the purpose of this section to define this distance and to prove a key property of this distance in relation to constrained finite-state model classes.

#### A. Kullback-Leibler and Divergence-Rate Distances

If $\lambda_1$ and $\lambda_2$ are two probability measures on $A^n$ for a given $n$, then $D(\lambda_1||\lambda_2)$, the Kullback-Leibler distance (also called Kullback-Leibler divergence) between $\lambda_1$ and $\lambda_2$ is defined by

$$D(\lambda_1||\lambda_2) = \sum_{x \in A^n} \lambda_1(x) \log \frac{\lambda_1(x)}{\lambda_2(x)}$$

with the usual convention that a term in the above sum is taken to be zero whenever $\lambda_1(x) = 0$ and is taken to be $\infty$ whenever $\lambda_1(x) > 0$ and $\lambda_2(x) = 0$. The following properties are well known [7, ch. 5]:

1. $0 \leq D(\lambda_1||\lambda_2) \leq \infty$.
2. $D(\lambda_1||\lambda_2) = 0$, if $\lambda_1 = \lambda_2$.
3. $D(\lambda_1||\lambda_2) \geq D(\lambda_1||\lambda_2)$, $x \in A^n$.

If $\mu$ is a probability measure on $A^n$ and $\Lambda$ is a family of probability measures on $A^n$, then $D(\mu||\Lambda)$ denotes the quantity

$$\inf_{\nu \in \Lambda} D(\mu||\nu).$$

Let $\mu \in M(\Lambda)$ and let $\Lambda$ be a nonempty subset of $M(\Lambda)$. We define the divergence-rate distance $D(\mu||\Lambda)$ from $\mu$ to $\Lambda$ by

$$D(\mu||\Lambda) = \lim_{n \to \infty} n^{-1} D(\mu||\Lambda_n),$$

provided that the sequence $\{n^{-1} D(\mu||\Lambda_n)\}$ has a limit.

From properties 1–2 of Kullback-Leibler distance, we obtain the following properties of divergence-rate distance:

1. $0 \leq D(\mu||\Lambda) \leq \infty$.
2. $D(\mu||\Lambda) = 0$, if $\mu \in \Lambda$.

We present a result which shows that $D(\mu||\Lambda)$ exists whenever $\mu$ is shift-invariant and $\Lambda$ is the set of mixture models arising from a constrained finite-state model class.

**Proposition 1.** Let $M$ be a constrained finite-state model class. Let $[X, \mu]$ be a stationary source. Then the sequences $\{n^{-1} D(\mu||\Lambda_n) : n \in N\}$ and $\{n^{-1} E[\log \tau_{M_i}(X^n)] : n \in N\}$ both have a limit and the following formula is valid:

$$D(\mu||\Lambda) = -H(\mu) - \lim_{n \to \infty} n^{-1} E[\log \tau_{M_i}(X^n)].$$

**Proof:** We first show that the sequence $\{n^{-1} E[\log \tau_{M_i}(X^n)]\}$ possesses a limit. Suppose $E[\log \tau_{M_i}(X^n)] = -\infty$ for some $m \in N$. By Lemma A.5 of the Appendix,

$$\log \tau_{M_i}(X^n) \leq \log \tau_{M_i}(X^n)$$

everywhere, $n > m$. (3.2) Taking the expected value of both sides of (3.2), we see that $E[\log \tau_{M_i}(X^n)] = -\infty$, $n > m$. Hence, all but finitely many terms of the sequence $\{n^{-1} E[\log \tau_{M_i}(X^n)]\}$ are $-\infty$ and so this sequence possesses a limit (namely, $-\infty$).

Now suppose $E[\log \tau_{M_i}(X^n)]$ is finite for every $n$. Using Lemma A.5 of the Appendix again, we see the
sequence \( \{ E[\log \tau_M(X^n)] \} \) is subadditive and so the sequence \( \{ n^{-1} E[\log \tau_M(X^n)] \} \) possesses a limit.

We now relate the two sequences \( \{ n^{-1} D(\mu_v || \tilde{M}_n) \} \) and \( \{ n^{-1} E[\log \tau_M(X^n)] \} \). Note that if \( \nu \in \tilde{M} \) then,

\[
D(\mu_v || \tilde{M}_n) = E[\log (\nu(x^n) / \mu_v(x^n))] \\
\geq E[\log (\nu(x^n) / \tau_M(x^n))],
\]

and so

\[
D(\mu_v || \tilde{M}_n) \geq -H(\mu_v) - E[\log \tau_M(X^n)], \quad \forall \ n. \tag{3.3}
\]

From Lemma A.1 of the Appendix, there exists for each \( n \) a measure \( \nu^n \in \tilde{M}_n \) such that

\[
\tau_M(X^n) \leq C_n \nu^n(X^n) \quad \text{everywhere},
\]

where \( \{ C_n \} \) is a sequence of positive constants for which \( n^{-1} \log C_n \to 0 \) as \( n \to \infty \). This gives us the bound

\[
D(\mu_v || \tilde{M}_n) \leq E[\log (\nu^n(x^n) / \mu_v(x^n))] \\
\leq \log C_n - H(\mu_v) - E[\log \tau_M(X^n)], \quad \forall \ n. \tag{3.4}
\]

Combining (3.3) and (3.4) we see that the sequence \( \{ n^{-1} D(\mu_v || \tilde{M}_n) \} \) possesses a limit because the sequence \( \{ n^{-1} E[\log \tau_M(X^n)] \} \) does. Therefore, the distance \( D(\mu || \tilde{M}) \) is defined and, appealing to (3.3) and (3.4) again, we see that formula (3.1) is valid.

**B. A Property of Divergence-Rate Distance**

We now present our key result on divergence-rate distance. We shall exploit this result later to obtain our main results Theorems 1 and 2.

**Proposition 2:** Let \( M \) be a constrained finite-state model class, and let \( \mu \in M(A) \) be shift-invariant. Suppose \( D(\mu || \tilde{M}) = 0 \). Then, \( \mu \in \tilde{M} \); furthermore, \( \mu \in M \) if \( \mu \) is ergodic.

The proof of the Proposition will be accomplished by means of two lemmas. First, we develop some notation concerning conditional distributions. If \( \mu \in M(A) \) and \( x \in A^* \), then \( \mu(\cdot | x) \) denotes the probability measure in \( M(A) \) such that

\[
\mu(\cdot | x) = \mu, \quad \text{if} \ \mu([x]) = 0; \\
\mu(\cdot | x) = \mu([x,y]) / \mu([x]), \quad y \in A^*, \quad \text{otherwise}.
\]

(In other words, if \( [x, \mu] \) is a source and \( x \in A^n \), then \( \mu(\cdot | x) \) is the conditional distribution of \( (x_{n+1}, x_{n+2}, \cdots) \) given \( (x_1, x_2, \cdots, x_n, y) = x \) if \( \mu \in M(A) \) is shift-invariant, then \( \mu(\cdot | x) : x \in A^\infty \) denotes a family of measures in \( M(A) \) such that 1) \( \mu(E|x) \) is a \( A^\infty \)-measurable function of \( x \) for each \( A^\infty \)-measurable set \( E \), and 2) if \( (X_i : i = 0, \pm 1, \pm 2, \cdots) \) is a bilateral process such that the distribution of \( (X_1, X_{i+1}, \cdots) \) is \( \mu \) \( \forall \ i \), then \( \mu(\cdot | x) \) serves as the conditional probability distribution for the process \( (X_1, X_2, \cdots) \) given \( (X_0, X_{-1}, X_{-2}, \cdots) = x \).

**Lemma 1:** Let \( \Lambda \) be a nonempty subset of \( M(A) \) such that \( \nu(\cdot | x) \in \Lambda \) for all \( \nu \in \Lambda \) and \( x \in A^* \). Let \( \mu \in M(A) \) be shift-invariant and let \( (X_i : i = 0, \pm 1, \pm 2, \cdots) \) be a process such that the distribution of \( (X_0, X_{i+1}, \cdots) \) is \( \mu \) for all \( i \). Suppose \( D(\mu || \Lambda) \) is defined. Then,

\[
D(\mu || \Lambda) \geq \sup_j j^{-1} E[D(\mu(\cdot | X_0, X_{-1}, X_{-2}, \cdots) || \Lambda_j)]. \tag{3.5}
\]

**Proof:** Fix \( j \in \mathbb{N} \). Then, for any \( n \geq 2 \) and any \( \nu \in \Lambda \), we may write

\[
D(\mu_v || \nu_{n_j}) = D(\mu_v || \nu_j) + \sum_{i=2}^{n} E[D(\mu_j(\cdot | X_{i}^{(i-1)}) || \nu_j(\cdot, X_{i}^{(i-1)}))] \tag{3.6}
\]

Taking the infimum over \( \nu \) then yields

\[
D(\mu_v || \Lambda_{n_j}) \geq \sum_{i=2}^{n} E[D(\mu_j(\cdot | X_{i}^{(i-1)}) || \Lambda_j)]. \tag{3.7}
\]

Let \( \epsilon_n \) be the quantity defined by

\[
\epsilon_n \triangleq n^{-1} \sum_{i=2}^{n} \left[ H(X_i | X_0, X_{-1}, X_{-2}, \cdots) \right] - H(X_i | X_0, X_{-1}, \cdots, X_{i-1}, \cdots) \}
\]

where the notation \( H(U|V) \) is used to designate the conditional entropy of the random variable \( U \) given the random variable \( V \). The right side of (3.6) equals

\[
(n-1)E[D(\mu_j(\cdot | X_0, X_{-1}, X_{-2}, \cdots) || \Lambda_j)] + \epsilon_n \tag{3.8}
\]

The sequence \( \epsilon_n \) tends to zero with \( n \) for any bilateral process \( (X_i) \) whose entries come from a finite set. Hence, we have the following bound for each \( j \):

\[
\lim_{n \to \infty} j^{-1} D(\mu_v || \Lambda_{n_j}) \geq \sup_j j^{-1} E[D(\mu(\cdot | X_0, X_{-1}, X_{-2}, \cdots) || \Lambda_j)]. \tag{3.9}
\]

The left side of (3.7) is \( \leq D(\mu || \Lambda) \) and so (3.5) follows.

**Lemma 2:** Let \( \Lambda \) be a closed convex non-empty subset of \( M(A) \) such that \( \nu(\cdot | x) \in \Lambda \) for all \( \nu \in \Lambda \) and \( x \in A^* \). Let \( \mu \in M(A) \) be shift-invariant. Suppose \( D(\mu || \Lambda) \) is defined. If \( D(\mu || \Lambda) = 0 \), then \( \mu \in \Lambda \).

**Proof:** Let \( (X_i) \) be a stationary process such that the distribution of \( (X_i, X_{i+1}, \cdots) \) is \( \mu \) for all \( i \). Suppose \( D(\mu || \Lambda) = 0 \). From Lemma 1,

\[
D(\mu(\cdot | X_0, X_{-1}, X_{-2}, \cdots) || \Lambda_j) = 0 \quad \text{a.s.,} \quad \forall \ j,
\]

from which it follows that

\[
\mu(\cdot | X_0, X_{-1}, \cdots) \in \Lambda, \quad \text{a.s.,} \tag{3.10}
\]

using P.3) and the fact that \( \Lambda \) is a closed subset of \( M(A) \). We have

\[
E[\mu(S | X_0, X_{-1}, \cdots)] = \mu(S), \quad S \text{ a meas. subset of } A^\infty. \tag{3.11}
\]
Let \( \Phi \) be the distribution of the random measure \( \mu(\cdot | X_0, X_1, X_2, \ldots) \) we see from (3.8) that \( \Phi \) is concentrated on \( \Lambda \). This fact allows us to re-write (3.9) as

\[
\mu(S) = \int_{\Lambda} \nu(S) \, d\Phi(\nu), \quad \forall S.
\]

Since \( \Lambda \) is closed and convex, we can deduce that \( \mu \in A \) from the preceding relation.

**Proof of Proposition 2:** Let \( M \) be a constrained finite-state model class, and suppose \( D(\mu | M) = 0 \), where \( \mu \) is shift-invariant. Lemma A.4 of the Appendix tells us that \( \nu(\cdot | x) \in M \) for any \( \nu \in M \) and \( x \in \mathbb{A}^\infty \). The model class \( M \) is clearly convex. To see that \( \mu \) is a closed subset of \( M(A) \), we note that \( M \) is a closed subset of \( M(A) \) and therefore is compact since \( M(A) \) is compact. Applying the Banach–Alaoglu Theorem [2, Theorem 3.5.16], the set of probability measures on \( M(A) \) is compact and a closed subset of \( M \).

Appealing to Lemma 2 (with \( \Lambda = M \)), we have \( \mu \in M \). Furthermore, if \( \mu \) is ergodic we may apply Lemma A.7 of the Appendix to conclude that \( \mu \in M \).

**IV. CONCEPT OF RATE PROFILE**

A code will encode different sources at different rates. Informally speaking, the rate profile of a code is a function which gives these different rates. In this section, we formally define the concept of rate profile and examine the rate profile of maximum likelihood codes. The determination of the rate profile of a maximum likelihood code proves to be crucial in our attack upon the identification problem for constrained finite-state model classes.

**A. Rate Profile of a General Code**

The rate \( R(X^n) \) (bits per source symbol) at which a code \( \phi \) encodes the first \( n \) symbols \( X^n \) generated by a source is the ratio \( \log \mu(\phi(X^n))/n \). We are interested in the asymptotic behavior of the rate \( R(X^n) \) as \( n \to \infty \). To this end, we say that a code \( \phi \) is stable if, for any ergodic source \( [X, \mu] \), the sequence \( \{R(X^n)|\phi|: n \in \mathbb{N}\} \) converges almost surely to a constant \( R(\mu|\phi) \in [0, \infty] \). If a code is stable, its rate profile is defined to be the function \( \mu \to R(\mu|\phi) \) on \( M_e \).

**B. Rate Profile of Maximum Likelihood Codes**

In this subsection, we shall prove the following result which gives the rate profile of the maximum likelihood codes that are associated with constrained finite-state model classes.

**Proposition 3:** Let \( M \) be a constrained finite-state model class. Then the maximum likelihood code \( \phi_M \) is stable, and for each \( \mu \in M_e \), the following is true:

\[
R(\mu|\phi_M) = H(\mu) + D(\mu||M), \quad \text{if} \quad D(\mu||M) < \infty, \quad \forall \ n = 2 \log |A|, \quad \text{otherwise.} \tag{4.1}
\]

**Proof:** Fix \( \mu \in M_e \). Fix a source whose output \( X \) has distribution \( \mu \). Then, by Lemma A.6 of the Appendix, one of the following two conditions holds:

(C.1) \( \Pr[\tau_M(X^n) > 0 \ | \ n] = 1 \).

(C.2) \( \Pr[\tau_M(X^n) > 0, \ n \text{ suff. large}] = 1 \).

Condition C.1 is equivalent to the condition \( D(\mu_n || M) < \infty \) for all \( n \). Under condition C.1, the sequence \( \{R(\mu_n||M): n \in \mathbb{N}\} \) has the same almost sure limiting behavior as the sequence \( \{-n^{-1} \log \tau_M(X^n): n \in \mathbb{N}\} \). But, in view of Lemma A.5 of the Appendix and the subadditive ergodic theorem [12], this latter sequence is almost surely convergent to the extended real number \( \lim_{n \to \infty} -n^{-1} \log \tau_M(X^n) \) under condition C.1. Under condition C.2, the sequence \( \{R(\mu_n||M)\} \) is almost surely convergent to \( 2 \log |A| \). Hence, the code \( \phi_M \) is stable and

\[
R(\mu|\phi_M) = \lim_{n \to \infty} -n^{-1} \log \tau_M(X^n), \quad \text{if} \quad D(\mu_n||M) < \infty, \quad \forall \ n = 2 \log |A|, \quad \text{otherwise.}
\]

Applying formula (3.1) to the preceding, we see that formula (4.1) is valid.

**Corollary:** Let \( M \) be a constrained finite-state model class. Let \( \mu \in M_e \). Then, \( R(\mu|\phi_M) = H(\mu) \) if \( \mu \in M \) and \( R(\mu|\phi_M) > H(\mu) \) if \( \mu \in M \).

**Proof:** Fix \( \mu \in M_e \). We first suppose that \( \mu \in M \). We see that the condition \( D(\mu_n||M) < \infty \) for all \( n \) holds. (Indeed, \( D(\mu_n||M) = 0 \) for all \( n \).) Proposition 3 then gives us \( R(\mu|\phi_M) = H(\mu) + D(\mu||M) \). But \( D(\mu||M) = 0 \) since \( \mu \in M \), and so we conclude that \( R(\mu|\phi_M) = H(\mu) \).

We now suppose that \( \mu \not\in M \). If the condition \( D(\mu_n||M) < \infty \) for all \( n \) does not hold, then \( R(\mu|\phi_M) = 2 \log |A| \) by Proposition 3. Since \( 2 \log |A| > \log |A| \geq H(\mu) \), we can conclude that \( R(\mu|\phi_M) > H(\mu) \). On the other hand, assume that the condition \( D(\mu_n||M) < \infty \) for all \( n \) does not hold. The Proposition 3 tells us that \( R(\mu|\phi_M) = H(\mu) + D(\mu||M) \). As \( \mu \not\in M \), we must have \( D(\mu||M) > 0 \) in view of Proposition 2. We are again led to the conclusion that \( R(\mu|\phi_M) > H(\mu) \).

**V. PROOFS OF PRINCIPAL RESULTS**

In this final section, we prove Theorems 1 and 2.

**Proof of Theorem 1:** Let \( M \) be a constrained finite-state model class and let \( \beta, \alpha, \phi \) be as specified in the statement of Theorem 1. Fix an ergodic source \([X, \mu]\).

**Case 1:** \( \mu \in M \). Fix an arbitrary real constant \( Q \). For each \( n \in \mathbb{N} \), let \( E_n \) be the set

\[
E_n \triangleq \{ x \in \mathbb{X}^n : L[\phi(x)] \leq Q \}.
\]

Then condition (2.3) will hold if we can show that the sequence \( \{\mu_n(E_n): n \in \mathbb{N}\} \) is summable. Since \( \phi \) is the maximum likelihood code for \( M \), we have

\[
\mu_n(x) \leq \tau_M(x) \leq 4T_M(n) \exp(-L[\phi(x)]), \quad x \in \mathbb{X}^n, \quad n \in \mathbb{N}.
\]
From this, it follows that
\[ \mu_n(x) \leq 4\tau_M(n)n^{-\alpha}2^Q \exp\left(-L[\phi_{M}(x)]\right), \quad x \in E_n. \]
Summing then yields (in view of Fact 1)
\[ \mu_n(E_n) \leq 4\tau_M(n)n^{-\alpha}2^Q, \quad n \in \mathbb{N}. \]
By choice of \( \alpha \), the sequence \( \{\tau_M(n)n^{-\alpha}\} \) is summable and therefore the sequence \( \{\mu_n(E_n)\} \) is summable.

Case 2: \( \mu \notin M \). From statement (2.6), we see that the sequence \( \{n^{-1}L[\phi(X^n)] : n \in \mathbb{N}\} \) converges almost surely to \( H(\mu) \). From the fact that maximum likelihood codes are stable and the Corollary to Proposition 3, we see that the sequence \( \{n^{-1}L[\phi(X^n)] : n \in \mathbb{N}\} \) converges almost surely to the constant \( R(\mu, \phi) \) which is greater than \( H(\mu) \). Hence, statement (2.4) holds.

**Proof of Theorem 2:** Let \( \{M_i : i \in \mathbb{N}\} \) be an increasing sequence of constrained finite-state model classes. Let \( \{\phi_i : i \in \mathbb{N}\} \) be the sequence of positive integers and let \( \{\phi_i : i \in \mathbb{N}\} \) be the sequence of codes defined in the statement of Theorem 2. Fix an ergodic source \([X, \mu]\) in which \( \mu \) lies in the union of the classes \( \{M_i\} \). The proof will be complete if we can prove
\[ \lim_{n \to \infty} \arg \min_{\phi_i} L[\phi_i(X^n)] \leq r \text{ a.s.,} \quad \mu \in M^r \] (5.1)
and
\[ \lim_{n \to \infty} \arg \min_{\phi_i} L[\phi_i(X^n)] \geq r + 1 \text{ a.s.,} \quad \mu \notin M^r. \] (5.2)

**Proof of (5.1):** Fix \( r > r \) such that \( \mu \in M^r \). Then the sequence \( \{n^{-1}L[\phi_i(X^n)] : n \in \mathbb{N}\} \) is almost surely convergent to \( H(\mu) < \infty \) and so we may choose a positive real number \( B \) such that
\[ \Pr\{L[\phi_i(X^n)] < nB, n \text{ suff. large}\} = 1. \] (5.3)
Let \( \{i_n : n \in \mathbb{N}\} \) be the sequence of positive integers in which
\[ i_n \doteq \lceil \log n B \rceil, \quad n \in \mathbb{N}. \]
It is easy to show that
\[ \exp\{i_n - 1\} \log n \geq nB, \quad n \geq 4. \] (5.4)
The sequence \( \{j_i : i \in \mathbb{N}\} \) satisfies
\[ j_i \geq 2^{i-1}, \quad i \in \mathbb{N}. \] (5.5)
We deduce from (5.4)–(5.5) that
\[ \min\{j_i \log n : i \geq i_n\} \geq nB, \quad n \geq 4, \]
and so, coupling this with (5.3), we have
\[ \Pr\{\min_{i \leq i_n} L[\phi_i(X^n)] > L[\phi_i(X^n)], n \text{ suff. large}\} = 1. \] (5.6)
For each pair of positive integers \( n, i \), define \( E_{n, i} \) to be the set
\[ E_{n, i} \doteq \{x \in A^n : L[\phi_i(x)] \geq L[\phi_i(X^n)]\}. \]
As in the proof of Theorem 1, we have
\[ \mu_n(x) \leq 4\tau_M(n)n^{-\alpha}2^Q \exp\{L[\phi_M(x)]\}, \quad x \in A^n, \quad n \in \mathbb{N}. \]
Then, applying (2.9), we obtain
\[ \mu_n(x) \leq 4n^n \exp\{L[\phi_M(x)]\}, \quad x \in A^n, \quad n \geq 2. \] (5.7)
It can be checked that
\[ -L[\phi_M(x)] \leq 1 - L[\phi_M(x)] + (j_r - j_i) \log n, \quad x \in E_{n, i}, \quad \forall n, i. \] (5.8)
It is easily seen that \( 2j_r - j_i \geq -2 \) for \( i > r \); consequently, using (5.7) and (5.8) in combination with Fact 1 we see that
\[ \mu_n(E_{n, i}) \leq 8n^{-2}, \quad \forall n, i. \]
We can now deduce that
\[ \mu_n(\cup\{E_{n, i} : r < i \leq i_n\}) \leq 8(\log n B)n^{-2}, \quad \forall n. \] (5.9)
The sequence on the right side of (5.9) is summable, and so we conclude that
\[ \Pr\{\min_{i \leq i_n} L[\phi_i(X^n)] > L[\phi_i(X^n)], n \text{ suff. large}\} = 1. \]
This fact, in combination with (5.6), gives us (5.1).

**Proof of (5.2):** Fix \( r > r \) such that \( \mu \notin M^r \). We have \( \mu \notin M^r \) for all \( i < r \) and from the fact that maximum likelihood codes are stable and the Corollary to Proposition 3, there exists \( \alpha > 0 \) such that
\[ \Pr\{\min_{i \leq r} L[\phi_i(X^n)] > n(H(\mu) + \alpha), n \text{ suff. large}\} = 1. \] (5.10)
Pick \( s > r \) such that \( \mu \in M^s \). We have
\[ \Pr\{L[\phi_i(X^n)] < n(H(\mu) + \alpha), n \text{ suff. large}\} = 1. \] (5.11)
Putting (5.10) and (5.11) together, we see that (5.2) must hold.

**APPENDIX**

**Lemma A1:** Let \( j \in \mathbb{N} \) and let \( C \) be a closed set of \( j \)-state transition functions. Let \( M \) be the model class \( M = M(C) \). Then there exists for each \( n \geq 2 \) a measure \( \sigma^n \in M_n \) such that
\[ \tau_M(x) \leq j|A|^n A_s^n(x), \quad x \in A^n, \quad n \geq 2. \] (A.1)

**Proof:** Fix \( n, \sigma \geq 2 \). We introduce some notation that will make the proof go smoothly. The notation \( u^n \) shall represent an \( n \)-dimensional vector variable \( u^n = (u_1, \ldots, u_n) \) whose components \( u_i \) come from \( I_j \). The notation \( z^n \) shall represent an \( n \)-dimensional vector variable \( z^n = (x_1, \ldots, x_n) \) whose components \( x_i \) come from \( A \). For each \( n \in \mathbb{N} \), let \( S_n \) be the
set of all pairs \((u^n, x^n)\) in which the variables \(u^n\) and \(x^n\) are allowed to take on all possible values.

For each \(j\)-state transition function \(p\), let \(W_p\) be the function defined on \(S_n\) in which
\[
W_p(u^n, x^n) = p(u_2, x_2|u_1) p(u_3, x_3|u_2) \cdots p(u_n, x_n|u_{n-1}) .
\]
The functions \(\{W_p\}\) are important for the following reason: If \(q\) is a probability measure on \(I_j \times A\) and \(p \in C\), then,
\[
\mu^q_p(x') = \sum_{u^n, x^n} q(u_1, x_1) W_p(u^n, x^n),
\]

(A.2)

Our proof uses the concept of type pioneered by Csizsar and Körner [6, ch. 2]. Two pairs \((u^n, x^n)\) and \((\tilde{u}^n, \tilde{x}^n)\) are members of \(S_n\) of the same type if for any \(a, b, c\) in which \(a, b \in I_j\) and \(c \in A\), the two sets \(\{i : (u_i, u_{i+1}, x_i) = (a, b, c)\}\) \(\{i : (\tilde{u}_i, \tilde{u}_{i+1}, \tilde{x}_i) = (a, b, c)\}\) have the same cardinality. The basic fact about types that we need here is: If \((u^n, x^n)\) and \((\tilde{u}^n, \tilde{x}^n)\) are members of \(S_n\) of the same type, then \(W_p(u^n, x^n) = W_p(\tilde{u}^n, \tilde{x}^n)\) for every \(j\)-state transition function \(p\).

For each \(s \in S_n\) let \(T_n(s)\) be the set of all members of \(S_n\) that have the same type as \(s\), and let \(T_n\) be a subset of \(S_n\) such that the sets \(T_n(s), s \in T_n\), form a partition of \(S_n\). For each \(s \in T_n\) let \(p(s)\) be a member of \(C\) for which \(W_p(s) = \max_{p \in C} W_p(s)\). (For fixed \(s\), the mapping \(p \to W_p(s)\) is a continuous function on the compact set \(C\), so there exists such an element \(p(s)\) of \(C\).)

From (A.2), we have
\[
\tau_M(x') \leq \sum_{(u^n, x^n): x^n = x'} \max_{p \in C} W_p(u^n, x^n),
\]

(A.3)
The sum on the right can be re-expressed as the double sum
\[
\sum_{s \in T_n} \sum_{(u^n, x^n) \in T_n(s): x^n = x'} \max_{p \in C} W_p(s).
\]
This double sum is equal to
\[
\sum_{s \in T_n} \sum_{(u^n, x^n) \in T_n(s): x^n = x'} W_p(s) (u^n, x^n).
\]
We can upper bound this quantity to get the following upper bound for the left side of (A.3):
\[
\tau_M(x') \leq \sum_{s \in T_n} \sum_{(u^n, x^n) \in T_n(s): x^n = x'} W_p(s) (u^n, x^n),
\]

(A.4)
Using \(q(u_1, x_1) = j^{-1}|A|^2\) in (A.2) we see that each of the interior summations on the right in (A.4) yields an element of \(M_n\) when divided by \(j|A|^2\). Formalizing this observation, let \(g\) be the probability measure equidistributed over \(I_j \times A\). Let \(\sigma^n \in M_n\) be the distribution that one obtains by averaging up the \(n\)th order marginals of the measures \(\{\mu^q_p(s) : s \in T_n\}\). Then rewriting the right side of (A.4) we see that
\[
\tau_M(x') \leq j|A| |T_n| \sigma^n(x'), \quad x' \in A^n .
\]

(A.5)
The cardinality \(|T_n|\) may be interpreted as the number of different types among the elements of \(S_n\). The number of types is \(\leq n^{|I_j|}|A|^2\) [6, lemma 2.2]. This bound, applied to (A.5), yields the desired conclusion (A.1).

The following lemma is an immediate consequence of Lemma A.1.

\textbf{Lemma A.2:} Let \(M\) be a constrained finite-state model class. Then there is a polynomial \(p\) such that \(\tau_M(n) \leq p(n)\) for all \(n\).

\textbf{Proof:} Let \(j \in \mathbb{N}\) be a \(j\)-state transition function, and let \(q\) be a probability measure on \(I_j \times A\). Let \(\{(S_i, X_i) : i \in \mathbb{N}\}\) be the homogeneous Markov chain with state space \(I_j \times A\) and with transition probabilities \(p((S_{i+1}, X_{i+1}) = (s, x)|(S_i, X_i) = (s', x')) = p(s, x|s')\) for all \(s', s, x, x'\), such that the distribution of \((S_1, X_1)\) is \(q\). Let \(m \in \mathbb{N}\). If \(x \in A^m\), let \(q^x\) be the conditional distribution of \((S_{m+1}, X_{m+1})\) given \(X^m = x\). With probability one, if \(X^m = x\) then \(\mu^q_p(x) = \mu^q(x^m)\).

\textbf{Proof:} Fix \(x \in A^m\). The distribution of \(\{X^i : i \geq 1\}\) is \(\mu^q_p\) and so \(\mu^q_p(x)\) is the conditional distribution of \(\{X^i : i > m\}\) given \(X^m = x\). To complete the proof, we arrive at this same conditional distribution in another way. Note that for any \(m' > m\) the three random objects \(X^m, \{(S^i, X^i) : m < i < m'\}\) and \(\{(S^i, X^i) : i > m'\}\) form a Markov chain, and so, conditioned on the event \(X^m = x\), the random sequence \(\{(S^i, X^i) : i > m\}\) is a homogeneous Markov chain with the same transition probabilities as the chain \(\{(S^i, X^i) : i \geq 1\}\). Thus, the conditional distribution of \(\{X^i : i > m\}\) given \(X^m = x\) is also equal to \(\mu^q(x^m)\).

\textbf{Lemma A.4:} Let \(M\) be a constrained finite-state model class. If \(\lambda(x)\) in \(M\) then \(\lambda(x) \leq \bar{M}\), for every \(x \in A^n\).

\textbf{Proof:} Let \(j \in \mathbb{N}\) and let \(C\) be a closed set of \(j\)-state transition functions such that \(M = C\). Let \(D_j\) be the set of all pairs \((q, p)\) in which \(q\) is a probability measure on \(I_j \times A\) and \(p \in C\). The set \(D_j\) can be viewed, in an obvious way, as a compact subset of a finite-dimensional Euclidean space, and we endow \(D_j\) with the measurable space structure it inherits from the underlying Euclidean space. Let \(\lambda \in M\). There is a probability measure \(\sigma\) on \(D_j\) such that
\[
\lambda = \int \mu^q_p d\sigma(q, p).
\]
Fix \(x \in A^*\) such that \(\lambda(x) > 0\). We argue that \(\lambda(x) x \in M\). Let \(\sigma^*\) be the probability measure on \(D_j\) specified in differential form as
\[
d\sigma^*(q, p) = (\mu^q_p(x)/\lambda(x)) d\sigma(q, p).
\]
From (A.6) and (A.7), we have
\[
\lambda(x) = \int \mu^q_p(x) d\sigma^*(q, p).
\]
Using the result and the notation of Lemma A.3, we may rewrite the preceding equation as
\[
\lambda(x) = \int \mu^q p(x) d\sigma^*(q, p).
\]
From this equation, it is clear that \(\lambda(x) \in M\).
Lemma A.5: Let $M$ be a constrained finite-state model class. Then

$$\tau_M((x,y)) \leq \tau_M(x) \tau_M(y), \quad x,y \in A^*.$$  \hspace{1cm} (A.8)

Proof: Let $\lambda \in M$. We have $\lambda([x,y]) = \lambda([x]) \lambda([y|x])$. Since $\lambda([x]) \in M$ in view of Lemma A.3, we then have $\lambda([x,y]) \leq \lambda([x]) \tau_M(y)$. Statement (A.8) then follows by taking the supremum over $\lambda$. \hfill $\square$

Lemma A.6: Let $M$ be a constrained finite-state model class. Let $(X_1, X_2, \cdots)$ be the output generated by an ergodic source. Then, one of the following two conditions holds:

C.A.1) $\Pr\{\tau_M(X^n) > 0 \forall n\} = 1$.

C.A.2) $\Pr\{\tau_M(X^n) = 0, \ n \text{ suff. large}\} = 1$.

Proof: By Lemma A.5, the event $\{\tau_M(X^n(1), X^n(2), \cdots, X^n(n)) > 0, \ n \text{ suff. large}\}$ implies the shifted event $\{\tau_M(X^n(2), X^n(3), \cdots, X^n(n+1)) > 0, \ n \text{ suff. large}\}$. From the fact that we are dealing with an ergodic source, we can then assert that $\Pr\{\tau_M(X^n) > 0, \ n \text{ suff. large}\} = 0$ or 1. Hence, the proof is complete once we show that the events C.A.1) and C.A.2) are complementary. To see this, simply note that the events $\{\tau_M(X^n) > 0, \ \text{inf. many } n\}$ and $\{\tau_M(X^n) > 0, \ \text{n suff. large}\}$ are complementary, and by Lemma A.5, the events $\{\tau_M(X^n) > 0, \ \text{inf. many } n\}$ and $\{\tau_M(X^n) > 0, \ \text{inf. many } n\}$ coincide. \hfill $\square$

Lemma A.7: Let $M$ be a constrained finite-state model class. Let $\mu \in M_G$ and $\mu \in \hat{M}$. Then $\mu \in M$.

Proof: Fix $\mu$ as hypothesized. Fix $j \in \mathbb{N}$ and a closed set $\mathcal{C}$ of $j$-state transition functions such that $M = M(\mathcal{C})$. We may write

$$\mu = \int_M v \, d\sigma(v),$$

for some probability measure $\sigma$ on $M$. Shifting, averaging, and taking a limit yields

$$\mu = \int_M \bar{v} \, d\sigma(v),$$  \hspace{1cm} (A.9)

where $\bar{v}$ denotes the stationary mean of the a.m.s. measure $v$ [8]. From (A.9) and ergodicity, we can say that

$$\mu = \bar{v}, \quad \text{for } \sigma \text{ almost every } v \in M.$$  \hspace{1cm} (A.10)

Let $v \in M$. Then we may write $v = \mu_\sigma^p$ for some $p \in C$. It follows from Lemma A.3 that each shift of $v$ is of the form $\mu_\sigma^q$. Therefore, $\bar{v}$ is also of the form $\mu_\sigma^q$ for some $q'$. We have shown that $\bar{v} \in M$ whenever $v \in M$. This fact, coupled with (A.10), tells us that $\mu \in M$. \hfill $\square$

ACKNOWLEDGMENT

The author thanks A. Barron, I. Csiszar, and the anonymous reviewers for helpful suggestions in the preparation of this paper.

REFERENCES


